UNIVERSIDAD DE GRANADA

Programa de Biomedicina

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

GENÉTICA DE ENFERMEDADES COMPLEJAS

UNIDAD DE BIOINFORMÁTICA

Análisis integrado de datos ómicos en enfermedades autoinmunes

Memoria de Tesis Doctoral

Daniel Toro Domínguez

Granada 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Daniel Toro Domínguez
ISBN: 978-84-1306-349-2
URI: http://hdl.handle.net/10481/57740

El trabajo desarrollado durante esta tesis doctoral ha sido financiado con fondos del proyecto europeo PRECISESADS (*Molecular Reclassification to Find Clinically Useful Biomarkers for Systemic Autoimmune Diseases*).

La investigación incluida en esta tesis doctoral fue desarrollada en su totalidad en el Centro Pfizer - Universidad de Granada - Junta de Andalucía de Genómica e Investigación Oncológica (GENYO), dentro del grupo de Genética de enfermedades complejas, liderado por la doctora Marta E. Alarcón Riquelme y dentro de la Unidad de bioinformática, dirigida por el doctor Pedro Carmona Sáez.

AGRADECIMIENTOS

Gracias por creer en mí, por darme aquella primera oportunidad de demostrar que podía hacerlo 5 años atrás, por enseñarme todo cuanto sé y por enseñarme a saber aprender. Gracias por el pragmatismo, por ayudarme a buscar ser mejor, por tu paciencia y por hacerlo fácil. Gracias por hacer que trabajar sea como una tarde de domingo, por ayudarme a mantener el ritmo en esta carrera de fondo con las mismas ganas que el primer día. Gracias por los cafés, por las risas y por los momentos serios, por ser un inquebrantable apoyo y por hacer que cada año sea "el año". Por ser un amigo. Esta tesis hubiera sido imposible sin tu ayuda. Sin más, gracias Pedro.

Gracias por enseñarme a amar la ciencia, por alimentar mi curiosidad hasta hacerla insaciable, por enseñarme a avanzar, a competir con uno mismo, a no conformarse. Gracias por hacerme un hueco, por incitarme a pensar, por las largas charlas donde surgen mil hipótesis, por hablar claro, por darme fuerzas y por ser ejemplo. Gracias por tu incondicional y contagiosa pasión por lo que haces. Gracias por hacer de la investigación, un sueño. Gracias Marta.

Gracias a mis directores de tesis. Esta tesis es vuestra.

A mis compañeros, gracias por aguantarme, por hacer de la oficina un hogar, por las risas y las medias de tomate. Por aconsejarme y ayudarme. Por amar vuestro trabajo y hacer amarlo al resto. Por ocupar una gran parte de mí. Es una suerte y un honor trabajar con vosotros. Gracias Jordi, Raúl, Adrián, Alba, Juan Antonio y Pilar.

Gracias por el incansable e incondicional apoyo, ayuda y cariño, por enseñarme unos valores férreos, por haber siempre creído en mí, sin dudar un solo instante, por enseñarme a que rendirse nunca es una opción y a darlo todo en cada intento. Por enseñarme a ser yo. Por abanderar la humildad, el respeto y el trabajo. Gracias papá, gracias mamá.

Gracias a todos aquellos que directa o indirectamente han contribuido a hacer esto posible, a los colaboradores y a todo el grupo de Genética e enfermedades complejas, por hacerme sentir uno más de la familia.

ABREVIATURAS Y GLOSARIO

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

Anticuerpo: Los anticuerpos son glucoproteínas que actúan como receptores de los linfocitos B y son empleados por el sistema inmunitario para identificar y neutralizar elementos patógenos o externos.

Antígeno: Sustancia con la capacidad de ser reconocida por el sistema inmunitario y provocar una respuesta en éste.

AR o RA: Artritis reumatoide (*rheumatoid arthritis*).

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

Auto-anticuerpo: Anticuerpo que reconoce y actúa directamente sobre antígenos del propio individuo. También conocido como anticuerpo autoreactivo.

CDK: Quinasa dependiente de ciclina (*cyclin-dependent kinase*).

Clustering: Conjunto de técnicas y algoritmos destinados a realizar análisis de agrupamiento, en los cuales los datos de un conjunto se fragmentan y forman núcleos más pequeños y similares entre ellos.

CMAP: *Connectivity Map*.

Dataset: Conjunto de datos de un tipo normalmente pertenecientes a un mismo estudio.

Endógeno: Referente a un producto o molécula originada dentro del propio organismo.

Exógeno: Referente a un producto o molécula originada fuera del propio organismo.

Firma genética: Conjunto de genes alterados en una condición respecto a una referencia.

Flare: Aumento medible en la actividad del lupus en uno o más sistemas orgánicos que involucran signos y síntomas clínicos nuevos o peores y / o mediciones de laboratorio.

Fold change: Valor que mide la diferencia de una magnitud entre dos condiciones. En estudios de transcriptómica, el *fold change* hace referencia a la diferencia de medias de expresión génica entre dos condiciones de estudio.

GSEA: Análisis de enriquecimiento por conjunto de genes (*Gene Set Enrichment Análisis*)

IFN I: Interferón tipo I.

IGF: Factor de crecimiento Insulínico (*insulin growth factor*).

ImaGEO: *Integrative Meta-Analysis of GEO Data*.

In sílico: Término que significa hecho por computadora o vía simulación computacional.

INMEX: *INtegrative MEta-analysis of eXpression data*.

LES o SLE: Lupus eritematoso sistémico o generalizado (*systemic lupus erythematosus*).

NCBI GEO: National Center for Biotechnology Information Gene Expression Omnibus.

Patógeno: Agente infeccioso que pueden provocar daño a su huésped.

PCR: Reacción en cadena de la polimerasa (*polymerase chain reaction*).

Perturbagen: Este término hace referencia a cualquier compuesto o experimento que causa cambios en la expresión genética de un sistema.

RNA-Seq: Técnica de secuenciación de ARN de nueva generación (*RNA Sequencing*).

SjS: Síndrome de Sjögren (Sjögren' syndrome).

SLEDAI: *Systemic Lupus Erythematosus Disease Activity Index*.

Sonda o cebador: Secuencia corta de ácido nucleico donde hibrida un determinado gen y que sirve como punto de partida para la replicación.

RESUMEN

En las últimas décadas, el desarrollo de las técnicas de generación masiva de datos, o técnicas ómicas, ha supuesto un punto de inflexión en el avance del conocimiento científico, cambiando el paradigma científico y el modo de abordar la generación de hipótesis. Estas técnicas permiten analizar grandes conjuntos de datos biológicos y generar conclusiones a partir de ellos, lo que ha llevado a categorizar la investigación biomédica como una ciencia de datos.

Una de estas técnicas es la transcriptómica, con la cual se puede medir los eventos de expresión génica de todos los genes, o la cantidad de veces que cada gen se expresa en una o un conjunto de muestras biológicas. La expresión génica es un reflejo directo de qué mecanismos moleculares están activados o inhibidos en un determinado contexto, y, por ende, prácticamente cualquier estado biológico posee patrones de expresión génica diferenciadores. Esto es realmente útil para el estudio de enfermedades, donde se puede analizar la expresión diferencial de los genes con respecto a individuos sanos y, de esa manera, profundizar en el conocimiento de los mecanismos moleculares patogénicos.

Junto con el potencial de las técnicas de generación masiva de datos, su abaratamiento ha causado un crecimiento exponencial en su uso y el almacenamiento de los datos en repositorios públicos, los cuales están disponibles para que cualquier investigador pueda analizarlos. Esto ha creado la necesidad de desarrollar nuevas técnicas que hagan posible tanto el manejo y procesado como la explotación e interpretación de los datos. Las ómicas por tanto van

obligatoriamente unidas al concepto de bioinformática, el cual se define como la aplicación de tecnologías computacionales y la estadística a la gestión y análisis de datos biológicos.

En esta tesis doctoral se han aprovechado las ventajas que nos ofrecen la transcriptómica y la bioinformática a la hora de generar nuevo conocimiento en enfermedades autoinmunes, una serie de patologías consideradas raras, por su baja incidencia en la población, y complejas, por su carácter multigénico y multifactorial. Además, se ha aprovechado la disponibilidad de usar datos almacenados en bases de datos públicas, lo que permite trabajar con grandes conjuntos de datos y generar resultados más robustos y sistemáticos, resolviendo el problema de que, generalmente, y debido a la baja prevalencia de estas enfermedades en la población, se suele trabajar con cohortes pequeñas de pacientes.

Los trabajos realizados durante el desarrollo de la tesis se contextualizan dentro de dos marcos principales. Primero, consideramos que dentro de las enfermedades autoinmunes existen ciertos patrones moleculares comunes. Con el primer trabajo recogido en la tesis, buscamos identificar los patrones genéticos homogéneos entre 3 enfermedades autoinmunes diferentes, el lupus eritematoso sistémico, la artritis reumatoide y el síndrome de Sjögren, mediante el uso de técnicas de meta-análisis basado en expresión génica. Como resultados, obtuvimos una serie de mecanismos moleculares de los cuales había sido descrita su relevancia en alguna de las enfermedades de manera individual, como la ruta de señalización de interferón tipo I o la apoptosis celular, y otras rutas biológicas aún no relacionadas con la autoinmunidad o no ampliamente estudiadas, como la desregulación en los procesos de translación proteica. En el segundo trabajo nos centramos sólo en lupus y continuamos con la búsqueda de los patrones homogéneos entre diferentes pacientes, con el objetivo de identificar nuevos fármacos candidatos para tratar esos patrones conservados dentro de la enfermedad mediante el uso de técnicas de reutilización de fármacos *in sílico*. Con este trabajo, propusimos una serie de nuevos posibles tratamientos con capacidad de revertir la firma genética patológica, entre los cuales había algunos en fases iniciales de estudio clínico, como los inhibidores de la ruta de la fosfoinositol quinasa.

Por otra parte, aunque observamos que hay mecanismos genéticos comunes dentro de lupus y entre diferentes patologías autoinmunes, desde el punto de vista clínico se observa una enorme heterogeneidad sintomatológica, serológica y de progresión de las enfermedades, y que se ve

reflejada además en una respuesta diferencial a fármacos entre los diferentes pacientes. Este hecho, da pie a hipotetizar sobre la existencia de patrones de expresión génica diferentes entre los distintos pacientes. El segundo marco conceptual de esta tesis se enfoca en el estudio de la heterogeneidad dentro de una misma enfermedad, con el objetivo de estratificar o agrupar los pacientes en subgrupos más homogéneos genética y molecularmente, reduciendo así la heterogeneidad global. El tercer artículo englobado en la tesis doctoral es un trabajo pionero desarrollado en lupus, pero aplicable en cualquier otra enfermedad compleja, donde desarrollamos un enfoque de estratificación de pacientes usando datos longitudinales de expresión, y en el que obtuvimos y replicamos 3 subgrupos homogéneos de progresión de la enfermedad dentro del lupus. Estos grupos fueron formados de acuerdo a cómo varia la expresión genética cuando la actividad de la enfermedad aumenta. La caracterización clínica nos mostró que en dos de los grupos el porcentaje de neutrófilos aumenta cuando la actividad de la enfermedad aumenta, mientras que en el tercer grupo es el porcentaje de linfocitos el que aumenta con la actividad de la enfermedad. Este hallazgo representa dos mecanismos de progresión de la enfermedad totalmente distintos. Además, estos grupos obtenidos mostraban diferencias significativas en cuanto a sintomatología clínica. Por ejemplo, entre otras, los grupos relacionados con neutrófilos mostraban un mayor número de casos de desarrollo de nefritis proliferativa, mientras que el grupo dirigido con linfocitos se asociaba a una mayor comorbidad con otras enfermedades autoinmunes, como el síndrome de Sjögren. Esta estratificación puede tener gran relevancia clínica, pudiendo usarse para el estudio de respuesta a fármacos dentro de cada grupo específicamente.

Por tanto, en esta tesis se han definido patrones de expresión génica comunes entre 3 enfermedades autoinmunes diferentes, se ha realizado un análisis de búsqueda de nuevos fármacos candidatos para el lupus basándonos en los mecanismos genéticos compartidos entre los diferentes pacientes, y se han definido 3 subgrupos de pacientes de lupus, caracterizados por mecanismos diferentes de progresión de la enfermedad.

ÍNDICE

1. INTRODUCCIÓN

1.1. CONCEPTOS BÁSICOS DEL SISTEMA INMUNE

El término **sistema inmunológico** engloba una red compleja de órganos, tipos celulares y procesos biológicos que constituyen la defensa natural del cuerpo contra agentes externos dañinos o patógenos, entre los que podemos encontrar virus, bacterias y otros parásitos, los cuales son identificados y diferenciados del tejido propio del organismo para su posterior eliminación.

Los mecanismos de defensas conforman un **sistema escalado** en función de la especificidad de la respuesta inmune ante el patógeno (1). La primera línea de defensa natural son las barreras de superficie que evitan la entrada de agentes externos al organismo, donde encontramos barreras físicas como la piel, mecánicas como el estornudo y la tos, que promueven expulsar agentes no deseados del tracto respiratorio, las lágrimas, el cerumen o la orina. También encontramos defensas químicas como las B-defensinas, péptidos antimicrobianos segregados por las células más en contacto con el exterior en piel y tracto respiratorio o ciertas enzimas segregadas en saliva, además de la creación de entornos ácidos para evitar la proliferación patogénica en estómago y en secreciones vaginales. Por último, un eslabón fundamental en la primera línea de defensa la constituye la flora endógena bacteriana, actuando en simbiosis con el organismo como una barrera biológica activa al competir por los nutrientes y el espacio,

impidiendo de este modo la proliferación de posibles poblaciones de microorganismos perjudiciales para el hospedador (2).

Si las barreras de superficie son superadas, el **sistema inmune innato** conforma el siguiente nivel de defensa. El sistema inmune innato proporciona una respuesta inmediata pero poco específica que consiste principalmente en el reclutamiento de macrófagos, mastocitos y células dendríticas al lugar de la infección en un proceso denominado inflamación. Estas células reconocen moléculas ampliamente compartidas entre un gran abanico de patógenos. Tras la primera respuesta, se produce vasodilatación en la zona y se liberan moléculas como la histamina o la serotonina que atraen a otros fagocitos, especialmente a neutrófilos, y se inicia en las células dendríticas la ruta de señalización mediada por interferón de tipo I (IFN I), que causa una retroalimentación positiva de la inflamación (3).

La liberación de citoquinas mayormente por parte de neutrófilos y los procesos de presentación de antígeno mediados por las células dendríticas son el germen de la última barrera defensiva, el **sistema inmune adaptativo**, orquestado en gran medida por los linfocitos T y los linfocitos B. La presentación de antígeno es comúnmente llevada a cabo entre una célula dendrítica y los receptores de histocompatibilidad de tipo I de los linfocitos T para agentes endógenos y los de tipo 2 para agentes exógenos. Este proceso inicia la activación de las células linfoides, las cuales median la respuesta inmune por diferentes mecanismos. La respuesta adaptativa mediada por linfocitos T es conocida como respuesta celular. Los linfocitos T citotóxicos median la eliminación de células dañadas o infectadas por reconocimiento de antígenos superficiales. Cuando este tipo de linfocito reconoce esos antígenos, libera citotoxinas que forman poros en la membrana plasmática de la célula diana, permitiendo la entrada en ella de iones, agua y toxinas, alterando su homeostasis y provocando apoptosis celular. Este proceso es clave para evitar la propagación de infecciones y la replicación de los virus, eliminando a las células hospedadoras. Los linfocitos T "helper" o colaboradores carecen de actividad citotóxica, pero son fundamentales en el control, balanceo y dirección de la respuesta inmune mediante la liberación de citoquinas y la expresión de receptores de superficie que promueven la activación de otras células inmunes. Por otro lado, la respuesta mediada por los linfocitos B es denominada respuesta humoral. Estas células identifican los patógenos mediante la unión de un anticuerpo de superficie celular a un determinado antígeno específico, proceso por el cual se activan los

linfocitos B, los cuales comienzan a dividirse y cuya descendencia segrega millones de copias del anticuerpo que reconoce al antígeno. Estos anticuerpos se unen a los patógenos portadores de esos antígenos, dejándolos marcados para su destrucción o para ser ingeridos por los fagocitos.

La característica principal de este sistema de defensa es la **memoria inmune** (3), es decir, una vez un determinado antígeno haya sido presentado y combatido por el sistema inmune a lo largo de la vida del organismo, éste almacenará células B con anticuerpos específicos contra ese patógeno, denominadas células plasmáticas. De este modo, la respuesta inmune es más rápida y eficiente ante una segunda infección del mismo patógeno, puesto que se tienen preparados los mecanismos de reconocimiento y eliminación para esa determinada amenaza.

Los órganos involucrados en el sistema inmunológico se denominan **órganos linfoides** y afectan al crecimiento, desarrollo y liberación de linfocitos. Los vasos sanguíneos y los vasos linfáticos son partes importantes de los órganos linfoides, debido a que median el transporte de los linfocitos hacia y desde diferentes partes del cuerpo. Cada órgano linfoide desempeña una función en la producción y la activación de los linfocitos y entre los que encontramos la médula ósea, el timo, los ganglios linfáticos, el bazo o el tejido linfoide asociado a las mucosas.

1.2. AUTOINMUNIDAD Y ENFERMEDADES AUTOINMUNES

Una actividad incorrecta del sistema inmune puede derivar en una serie de patologías entre las cuales encontramos la hipersensibilidad, o respuesta exacerbada del sistema inmune ante un agente patogénico o no patogénico, la inmunodeficiencia, o incapacidad del sistema inmune de hacer frente a patógenos debido a una actividad pobre o una cantidad baja de alguno de los componentes del sistema inmunológico, o la **autoinmunidad**, en la cual el sistema inmune falla en el reconocimiento adecuado y la distinción entre los tejidos propios del organismo y los agentes externos, generando una respuesta defensiva contra el propio cuerpo y dañando por tanto al mismo (4). Existen mecanismos naturales de control de la autoinmunidad, como la edición de los receptores de los linfocitos autoreactivos o la eliminación de éstos, procesos cuyo

funcionamiento parece estar alterado en eventos desencadenantes de autoinmunidad. Esta eliminación puede ocurrir a nivel central, durante la formación de los linfocitos T y B en el timo y la médula ósea, respectivamente, o periférico, donde sistemas de respuesta negativa llamada anergia, evitan la respuesta crónica a lo propio. Los linfocitos expresan receptores que inhiben las respuestas inmunológicas excesivas o detienen la respuesta iniciada evitando así su propagación y cronicidad. Estos mecanismos se engloban dentro de lo que se conoce como **tolerancia inmunológica**, que dicta los límites de qué debe ser atacado y que no dentro del sistema inmune.

Cuando un proceso autoinmune se vuelve crónico, nos encontramos ante una **enfermedad autoinmune**. Éstas engloban a una serie de más de 80 patologías consideradas raras o poco frecuentes de forma individual pero que en conjunto afectan en torno al 5% de la población (5), la cual debe someterse de por vida al consumo de medicamentos con fuertes efectos secundarios para paliar sus síntomas a menudo graves. Es por eso por lo cual se requiere profundizar más tanto en nuevas terapias más eficientes y menos agresivas como en la compleja etiología aún desconocida de las enfermedades autoinmunes. A día de hoy, son muchos los estudios y los conocimientos referentes a mecanismos moleculares implicados en las diferentes enfermedades y, aunque aún insuficientes como para reflejarse en una clínica eficaz, han sido descritas tanto asociaciones a nivel genético como implicaciones a nivel ambiental.

Como objeto de estudio de esta tesis encontramos las siguientes enfermedades autoinmunes:

- *Lupus Sistémico Eritematoso* (6).

El **lupus eritematoso sistémico** (LES o SLE) es una de las enfermedades autoinmunes con más prevalencia, siendo enormemente más frecuente en mujeres que en hombres y caracterizada por la presencia de anticuerpos autoreactivos. Es una enfermedad altamente heterogénea en cuanto a manifestaciones moleculares, alteraciones serológicas y afecciones clínicas, pudiendo causar desde erupciones cutáneas, daño a órganos como riñones, hígado, corazón o sistema nervioso, hasta fallos muliorgánicos y problemas inflamatorios a nivel sistémico. Además, la enfermedad presenta un curso clínico impredecible, alternando entre periodos de baja actividad de la enfermedad o remisiones, con afecciones leves o inexistentes, y periodos de alta actividad de

la enfermedad conocidos como brotes o *flares*, momentos en los que se producen la mayoría de los daños.

- *Artritis Reumatoide* (7).

La **artritis reumatoide** (AR o RA) es una enfermedad sistémica caracterizada por una inflamación continuada de las articulaciones de forma simétrica, produciendo su destrucción progresiva y generando distintos grados de deformidad, aunque al igual que el SLE, las manifestaciones clínicas pueden ser amplias, afectando a la capacidad funcional de diferentes órganos. Esta enfermedad se asocia con la presencia de autoanticuerpos contra péptidos cíclicos citrulinados.

- *Síndrome de Sjögren* (8).

El **síndrome Sjögren** (SjS) es una enfermedad que afecta también a las articulaciones y puede afectar a diferentes órganos de manera sistémica, aunque los principales objetivos atacados son las glándulas exocrinas, provocando la sequedad e incapacidad funcional de las mismas por infiltración linfocitaria. De igual modo, esta enfermedad está caracterizada por la presencia de autoanticuerpos.

Aunque estas patologías están caracterizadas por fenotipos y manifestaciones clínicas específicas, a menudo pacientes catalogados bajo una enfermedad concreta desarrollan fenotipos de otras, por lo que parece que debe existir una serie de mecanismos moleculares comunes entre ellas. A este hecho se suma la enorme heterogeneidad sintomatológica, serológica y de progresión dentro de una misma enfermedad, que se ve reflejada además en una respuesta diferencial a fármacos entre los diferentes pacientes, pudiéndose dar casos de pacientes de distintas enfermedades que compartan más características patogénicas que en comparación con los demás pacientes de su propia enfermedad. Estos conceptos son objeto de estudio dentro del proyecto europeo **PRECISESADS**, del cual forman parte un total de 28 instituciones de 12 países (http://www.precisesads.eu/about-the-project/), el cual ha financiado y soportado ideológicamente esta tesis, y el cual parte de las siguientes hipótesis

- La clasificación clínica de las enfermedades autoinmunes no refleja patrones moleculares únicos y diferenciadores para cada una de las enfermedades.

- Todas las enfermedades autoinmunes comparten ciertos mecanismos patogénicos.
- Dentro de cada enfermedad hay mucha heterogeneidad fenotípica, lo que hace suponer que refleja diferencias moleculares.
- Unos pacientes, y otros no, con distintas enfermedades autoinmunes, pueden compartir misma sintomatología clínica.

Bajo estas premisas, el proyecto busca la reclasificación de las enfermedades autoinmunes, olvidando la etiqueta clínica que define a cada paciente bajo el marco de una enfermedad concreta y buscando establecer nuevos subgrupos de pacientes más homogéneos molecular y fenotípicamente, usando para ello datos de múltiples ómicas. Con esto, no sólo se conseguiría una reagrupación funcional y molecular de pacientes, sino que se abre las puertas hacia una medicina más personalizada dentro de la autoinmunidad, al estudiar las posibles terapias no dentro de una enfermedad altamente heterogénea, sino dentro de subgrupos guiados por los mismos mecanismos patogénicos. Durante la tesis, se han desarrollado y publicado una serie de trabajos que se sustentan y a la vez apoyan las hipótesis mencionadas anteriormente.

1.3. TRANSCRIPTÓMICA Y BIOINFORMÁTICA

El avance del conocimiento científico y el desarrollo tecnológico son dos conceptos estrechamente ligados (9). El desarrollo de nuevas tecnologías de análisis genómico y el abaratamiento en los costes de éstas han permitido a la comunidad investigadora no sólo profundizar en los estudios, sino que han cambiado el paradigma científico y el modo de abordar la generación de hipótesis. En especial, las técnicas de generación masiva de datos, conocidas como "**ómicas**", han supuesto un punto de inflexión en las ciencias biomédicas, ya que permiten extraer grandes cantidades de información y son extremadamente útiles para caracterizar a nivel global sistemas biológicos de muestras experimentales. Entre las ómicas más comúnmente utilizadas encontramos la genómica, que busca la secuenciación del DNA de una muestra de estudio determinada con posibles múltiples aplicaciones, desde la búsqueda de mutaciones hasta la caracterización filogenética. La metabolómica es la cuantificación de metabolitos o productos moleculares de procesos metabólicos mediante técnicas como la espectrometría de

masas o la espectrometría de resonancia magnética nuclear. La proteómica identifica y cuantifica los niveles de proteínas en una muestra de manera similar al caso anterior. Por otro lado, la metilación es un proceso por el cual se añaden pequeñas moléculas o grupos metilo a zonas concretas del DNA, actuando como regulación de la transcripción genética. Este proceso se encuentra fuertemente influenciado por el entorno ambiental y nos puede dar pistas de cómo este entorno influye a nivel molecular en el organismo. La epigenómica es el estudio cuantitativo de los niveles de metilación por gen y es usado comúnmente en estudios comparativos para determinar qué procesos están siendo diferentemente regulados entre diferentes condiciones. Por último, hablaremos sobre la transcriptómica como fuente principal de los datos usados durante el desarrollo de esta tesis doctoral. La **transcriptómica** engloba una amplia gama de técnicas que permiten cuantificar el nivel de expresión de todos los genes transcritos a RNA en una muestra, pudiendo medir millones de moléculas de RNA al mismo tiempo.

Inicialmente, los ***micro-arrays*** o chips de DNA (10) eran las técnicas más usadas en este campo, que consistían en una superficie sólida de vidrio sobre la que se fijaban sondas de DNA o secuencias cortas de DNA de genes conocidos. El RNA de la muestra de estudio es amplificado mediante la reacción en cadena de la polimerasa (PCR) y marcado con fluorocromos para posteriormente difundirlo sobre la superficie del *microarray*, donde hibrida y se fija a las sondas. Cuantas más copias haya de un determinado gen, más hibridará en las sondas complementarias específicas para ese gen, y, por tanto, el nivel de expresión será proporcional a la cantidad de fluorescencia medida en el punto del *microarray* correspondiente al gen.

Las técnicas de secuenciación de RNA de nueva generación o **RNA-Seq** (11), han desbancado recientemente en uso a los *microarrays* debido a una serie de ventajas como que no es necesario conocer los genes previamente a la cuantificación, pudiendo medir la expresión de genes desconocidos, además de que los niveles de expresión se miden de forma directa, contando el número de copias de cada gen y no extrapolando a partir de los niveles de fluorescencia. El proceso comienza con la conversión de todas las moléculas de RNA a DNA complementario mediante PCR reversa. El DNA complementario es amplificado mediante PCR para formar lo que se denomina librería de DNA, la cual puede ser secuenciada. El proceso de secuenciación puede variar dependiendo de la tecnología usada, aunque el objetivo de todas ellas es la

identificación del orden correcto de nucleótidos de una secuencia de DNA. Una de las más utilizadas es la secuenciación desarrollada por la empresa Illumina, en este caso, dos secuencias de unión son incorporadas a cada una de las secuencias de DNA en la generación de las librerías, luego se vierten dentro de una celda de flujo que contiene millones de nanopozos con 2 tipos de cebadores fijados, o secuencias complementarias a los sitios de unión insertados en las librerías. En cada nanopozo entra una sola hebra de DNA y se fija a uno de las sondas y comienza un proceso de amplificación en puente, donde las nuevas hebras creadas se unen por cada extremo a cada uno de los cebadores. Este proceso hace que se amplifique una secuencia de DNA de manera localizada dentro de cada nanopozo, lo que se denomina amplificación clonal. Por último, junto con DNA polimerasa, se añade una solución que contiene un único tipo de oligonucleótido marcado con un fluoróforo. Cada oligonucleótido posee un fluoróforo de distinto color a modo de etiqueta, por lo que es posible identificar si el oligonucleótido ha sido incorporado y a qué secuencia mediante la medición del color. Este proceso es repetido alternando entre cada uno de los oligonucleótidos y procesos de lavado. De este modo es identificado el orden de los nucleótidos de una secuencia de DNA. Finalmente, los conjuntos de secuencias obtenidas son alineadas sobre un genoma de referencia y es sobre éste en el que se realiza el conteo del número de copias de secuencias que alinean en cada gen del genoma de referencia.

Como incorporación más reciente a esta familia de técnicas encontramos la **secuenciación de RNA de célula única** (*scRNA-seq*) (12), la cual permite medir la transcriptómica de cientos o miles de células, pero de manera individual, célula a célula, lo que resulta realmente útil, por ejemplo, en el estudio de cómo ciertos microambientes influyen en la expresión genética.

Cuando hablamos de datos ómicos, hablamos de millones de mediciones las cuales hay que procesar y analizar para darles un sentido. Esto ha creado la necesidad de desarrollar nuevas técnicas que hagan posible tanto el manejo y procesado como la explotación e interpretación de los datos. Las ómicas por tanto van obligatoriamente unidas al concepto de **bioinformática**, el cual se define como la aplicación de tecnologías computacionales y la estadística a la gestión y análisis de datos biológicos. Esta relativamente nueva ciencia no sólo ha sido clave en la consolidación de las técnicas de generación masiva de datos, sino que, al ser excesivamente

heterogénea y amplia, sus aplicaciones son prácticamente incontables en multitud de campos. A lo largo de la tesis veremos algunas de estas técnicas junto con su aplicación práctica.

En los últimos años se ha podido apreciar un auge que continua en aumento en la cantidad de este tipo estudios de generación masiva de datos y esto, como mencionaba al inicio de este apartado, junto con el desarrollo paralelo de la bioinformática, ha cambiado en gran medida el modo de hacer ciencia. Generalmente, un estudio científico empieza con una hipótesis la cual debe ser probada, y para lo que se requiere un nivel suficientemente profundo en el campo como para poder formular los diferentes conceptos en base a dicha hipótesis. El potencial de estas nuevas técnicas ofrece un cambio en el orden lógico del método científico, ya que no es necesario partir de ninguna hipótesis, sino que las hipótesis son generadas en base a los datos o los resultados (Figura 1A). Es decir, estas nuevas tecnologías nos permiten extraer de los resultados nuevos conocimientos ni siquiera tenidos en cuenta hasta entonces. Pero, además, el **almacenamiento** de estos datos en bases de datos públicas también está creciendo exponencialmente (Figura 1B), lo que ofrece un recurso invaluable, a la mano de cualquier investigador, con el que poder realizar multitud de análisis bioinformáticos diferentes usando cohortes grandes de datos y obtener así nuevos hallazgos científicos. Como ejemplo de estas bases de datos encontramos consorcios como *The Cancer Genome Atlas Program*, donde se almacenan datos de pacientes de cáncer de múltiples ómicas. Si nos centramos en transcriptómica, NCBI GEO y ArrayExpress son con diferencia los máximos exponentes, almacenando por separado datos de más de 2 millones de muestras (13). Estos recursos públicos adquieren más valor si cabe para el estudio de enfermedades raras, los cuales normalmente están limitados por las dificultades en el reclutamiento de pacientes debido a la baja prevalencia de éste tipo de patologías, por lo que esta limitación se traduce a tamaños muestrales pequeños.

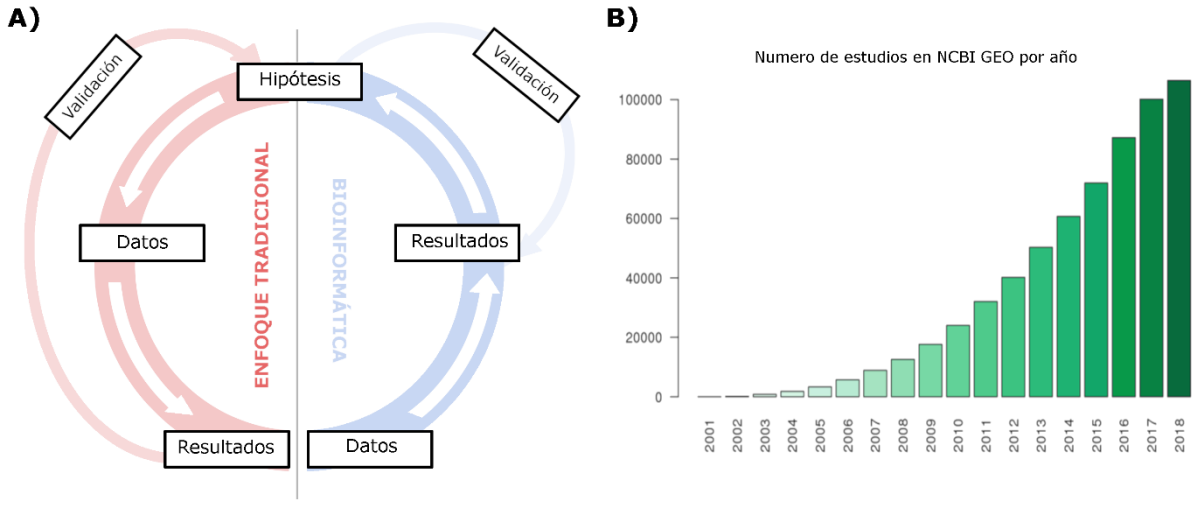


Figura 1: Bioinformática y bases de datos. La figura 1A muestra un esquema del orden seguido en los estudios tradicionales frente al orden de los estudios bioinformáticos que parten de datos masivos. La figura 1B muestra el incremento de almacenamiento de datos en NCBI GEO.

En esta tesis doctoral se han aprovechado las ventajas que nos ofrecen la transcriptómica y la bioinformática a la hora de generar nuevo conocimiento en enfermedades autoinmunes, así como la disponibilidad de usar datos almacenados en bases de datos públicas.

1.4. BASES DE DATOS PÚBLICAS

En esta sección hablaremos de los diferentes recursos públicos que usaremos a lo largo de la tesis.

- *Gene Expression Omnibus (GEO)*

GEO (13) es una base de datos pública creada y mantenida por el *National Center for Biotechnology Information* (NCBI), que almacena datos de expresión pertenecientes a una gran cantidad de estudios diferentes, y que han sido generados mediante tecnologías de *microarrays* o RNA-Seq. En total, NCBI GEO contiene más de cien mil estudios de expresión diferentes, lo

que suma alrededor de 3 millones de muestras. Estos datos están organizados y estructurados en 3 capas, las cuales pueden localizarse y extraerse mediante una serie de códigos. Los códigos GSM hacen referencia a muestras concretas. Su nomenclatura consiste en las letras **GSM** seguidas de un número único representante para cada muestra, por ejemplo, el código GSM260920 equivale a una muestra de linfocitos B de un paciente de LES. El siguiente nivel de la jerarquía son los códigos **GSE**, que engloban y representan al conjunto de muestras de un estudio determinado. Por ejemplo, el código GSE10325 representa un estudio con diferentes muestras de pacientes de LES y muestras de individuos sanos. Estos códigos GSE son realmente útiles a la hora de realizar diferentes análisis, ya que podemos extraer los datos en cómodas matrices donde los genes son representados en filas y las diferentes muestras en columnas, siendo los valores de expresión génica las entradas de la matriz. La tercera y última capa de organización en GEO son los códigos **GDS**. Éstos representan una colección curada por parte del personal de GEO e incluyen diferentes muestras o estudios que son biológicamente y estadísticamente comparables y analizables como un todo. Como ejemplo, GDS6247 representa un conjunto de muestras de un modelo de ratón de obesidad.

- LINCSCLOUD

Lincscloud, ahora llamado CLUE (14), es un repositorio y herramienta de exploración de datos perteneciente al Broad Institute (Cambridge, Massachussets, EEUU) y generada bajo el marco del proyecto denominado *The Library of Integrated Network-Based Cellular Signatures* (LINCS). Este repositorio surge con el objetivo de crear nuevo conocimiento dentro de la biología humana basándose en conexiones entre distintas áreas, como los cambios transcriptómicos y proteómicos, procesos de señalización, morfología celular y estados epigenéticos que son causados a las células ante la exposición de agentes perturbadores. En relación a la transcriptómica, esta base de datos contiene perfiles transcripcionales derivados o causados por más de 20000 compuestos o fármacos diferentes y alrededor de 7000 experimentos de alteraciones genéticas (sobre-expresión o silenciamiento de la expresión de un gen concreto), generados en una serie de células humanas, usando diferentes dosis y medidos a diferentes tiempos. Los datos se encuentran almacenados y organizados en diferentes niveles de procesamiento de los mismos, siendo el último nivel una biblioteca de listas de genes

ordenadas de acuerdo con la diferencia de expresión entre la muestra donde se realiza el experimento (por ejemplo, la aplicación de un determinado fármaco) y muestras control. Cada lista representa un único experimento, realizado con un determinado fármaco, en una línea celular, dosis y tiempo determinado. Estas listas nos muestran, por ejemplo, cuales son los genes más sobre-expresados e infra-expresados por un determinado fármaco y pueden ser utilizadas para buscar patrones genéticos inversos entre los perfiles causados por fármacos y las firmas presentes en una condición de interés, por ejemplo, una enfermedad concreta, en lo que se conoce como análisis de **reutilización de fármacos**, de lo que hablaremos en más detalle en la sección 4 de esta tesis.

El término *perturbagen* o agente perturbador, acuñado en esta base de datos, hace referencia a cada uno de los fármacos y experimentos de alteración de un gen con los cuales generaron las firmas transcripcionales. En total, la base de datos contiene más de 1300000 firmas de expresión genéticas diferentes. Esta enorme cantidad de datos es posible debido a un nuevo enfoque para medir la expresión genética que se desarrolló. Primero, analizaron de forma conjunta más de 2000 *microarrays* diferentes almacenados en la base de datos de GEO para determinar conjuntos de genes correlacionados, cuyas expresiones se comportan con una tendencia similar u opuesta entre los diferentes conjuntos de datos. Con este análisis, identificaron un grupo de 978 genes con los cuales, conociendo sus valores de expresión, era posible inferir la expresión del resto de genes del genoma mediante las reglas de correlación obtenidas previamente. Esto permitió reducir el número de genes a medir en los experimentos transcriptómicos, pasando de medir el genoma completo a un pequeño conjunto de genes, con lo cual se abarataron enormemente los costes en los experimentos y se permitió la generación y obtención masiva de firmas genéticas derivadas por los diferentes perturbagenes.

2. OBJETIVOS

Esta tesis doctoral se nutre de las premisas definidas dentro del marco del proyecto europeo PRECISESADS, recogidas en la sección de *Introducción* de este documento, y profundiza principalmente en 2 conceptos desde el punto de vista transcriptómico: 1; la homogeneidad dentro y entre enfermedades autoinmunes y 2; la heterogeneidad implícita dentro de una misma enfermedad. Los objetivos abordados durante la tesis doctoral son los siguientes:

1. Identificar patrones genéticos, mecanismos moleculares y rutas biológicas alteradas con respecto a individuos sanos de manera común y homogénea a través 3 enfermedades autoinmunes diferentes; LES, AR y SjS, mediante el uso de técnicas de meta-análisis basado en expresión génica.
2. Determinar nuevos fármacos candidatos para tratar el LES mediante análisis de reutilización de fármacos basándose en la capacidad de los fármacos de revertir los patrones genéticos homogéneos, comunes y compartidos a través de los diferentes pacientes.
3. Estratificar los pacientes de LES dentro de subgrupos homogéneos de acuerdo a cómo varía la expresión génica de manera longitudinal cuando aumenta la actividad de la enfermedad.

4. Caracterizar molecular, celular y clínicamente los subgrupos de LES obtenidos para determinar si existe una relación directa entre el subgrupo al que pertenece un paciente y un fenotipo patogénico específico.

El abordaje de estos objetivos ha sido desarrollado en una serie de publicaciones, las cuales, para clarificar y ordenar la lectura de esta tesis, se recogen en los siguientes 3 capítulos: **1**; *Meta-análisis basado en transcriptómica*, donde se trabaja sobre el objetivo número 1. **2**; *Análisis de reutilización de fármacos*, donde tratamos el objetivo 2. **3**; *Estratificación de pacientes de lupus*, que cubre los objetivos 3 y 4.

3. META-ANÁLYSIS BASADO EN TRANSCRIPTÓMICA

3.1. INTRODUCCIÓN AL META-ANÁLISIS

El término **meta-análisis** hace referencia a un conjunto de métodos estadísticos que abordan la combinación de datos pertenecientes a diferentes estudios con el objetivo de obtener un único resultado común significativo. Concretamente, los meta-análisis de expresión génica permiten combinar múltiples estudios de transcriptómica, generalmente para la identificación de genes o perfiles de expresión desregulados entre dos condiciones (por ejemplo, entre individuos sanos y pacientes de una determinada enfermedad) investigadas a través de los diferentes estudios (15). Estas técnicas junto al incremento del almacenamiento de datos en repositorios públicos ofrecen una poderosa sinergia con la que poder profundizar en el conocimiento de las enfermedades y la búsqueda e identificación de biomarcadores.

En los últimos años, encontramos multitud de publicaciones científicas donde usan técnicas de meta-análisis de expresión génica. Por ejemplo, Tianxiao Huan et al. (16), combinaron un total de 7017 pacientes con tensión arterial de 6 estudios diferentes e identificaron conjuntos de genes relacionados con diferentes manifestaciones clínicas dentro de la enfermedad. En el estudio de João Pedro de Magalhães et al. (17), usaron 27 *datasets* o conjuntos de datos de diferentes

estudios descargados de bases de datos públicas para buscar patrones genéticos relacionados con la edad, identificando un conjunto de 56 genes sobre-expresados con el incremento de la edad de forma consistente en todos ellos. Los meta-análisis han sido ampliamente usados en múltiples tipos de cáncer (18), para la identificación de biomarcadores de actividad y de progresión de la enfermedad. Pero además de esto, los meta-análisis resultan realmente útiles en el caso de las enfermedades raras o poco frecuentes, en las que generalmente los estudios son realizados con cohortes pequeñas de datos debido a las dificultades para el reclutamiento de pacientes. Los meta-análisis permiten aumentar enormemente el tamaño muestral incorporando varios estudios del mismo campo y de este modo, aumentar el poder estadístico y la robustez de los resultados. Por ejemplo, Ignazio S Piras et al. realizaron un meta-análisis usando 3 cohortes de esquizofrenia que descargaron de NCBI GEO, donde identificaron dos potenciales biomarcadores de la enfermedad (19) o el estudio de Jessica M Winkler et al., con el que revelaron disfunciones de expresión relacionada con procesos de homeostasis en Alzheimer (20).

En cuanto a enfermedades autoinmunes encontramos también algunos estudios publicados previos, como por ejemplo, Song et al. analizaron 3 conjuntos de datos de SjS (21), Arasappan et al. realizaron un meta-análisis basado en la rutas biológicas de conjuntos de datos de LES donde identificaron una firma de 37 genes asociada con la enfermedad (22), o el estudio realizado por Olsen et al. en el que encontraron funciones desreguladas en AR como la vía de señalización de IFN I (23). Además, también se han utilizado las técnicas de meta-análisis para integrar datos de diferentes enfermedades con el objetivo de descubrir patrones genéticos compartidos. En este contexto, Tuller et al. analizaron datos públicos sobre Esclerosis Múltiple, LES, Artritis juvenil, enfermedad de Crohn, colitis ulcerosa y diabetes tipo 1 (24). Silva et al. combinaron datos de LES y AR (25) o Higgs et al., que identificaron una firma común entre LES, AR y Escleroderma, nuevamente relacionada con los genes de IFN I (26).

3.2. APLICACIONES DEL META-ANÁLISIS

En el párrafo anterior ya hemos podido ver someramente las aplicaciones más comunes de los meta-análisis mediante ejemplos de trabajos publicados. De manera resumida, los meta-análisis son usados para 3 distintos fines en el contexto de la transcriptómica (Figura 2).

La primera aplicación es **aumentar el poder estadístico**, aumentando así la robustez y fiabilidad de los resultados. Esto se usa cuando tenemos una misma condición, por ejemplo, una misma enfermedad, en diferentes estudios. Como ejemplo, pongamos que tenemos varios estudios diferentes con muestras de control de individuos sanos y muestras de una enfermedad A, con el uso del meta-análisis aumentamos el tamaño muestral tanto de controles como de casos, teniendo en cuenta los datos de todos los estudios, lo que nos da una mayor fuerza estadística y nos permite extraer genes cuya expresión es consistentemente diferente entre los casos de la enfermedad A y los controles a través de los diferentes estudios. Por tanto, esta funcionalidad del meta-análisis es útil para extraer los perfiles de expresión consistentes para una determinada condición, así como biomarcadores específicos de tal condición. Este fin se utiliza también en genómica, aumentando el poder estadístico en estudios de asociación genética (27).

La segunda de las aplicaciones consiste en buscar **patrones genéticos comunes** entre diferentes condiciones. Por ejemplo, tenemos diferentes estudios con muestras de controles y casos de diferentes enfermedades, y queremos buscar que genes son diferencialmente expresados de manera compartida en todas esas enfermedades respecto a los controles sanos. Esta aplicación es útil para el estudio de enfermedades entre las cuales se supone cierto grado de similitud, para buscar mecanismos moleculares y rutas biológicas que intervienen en ambas patologías.

La tercera aplicación se basa en justo lo contrario que la anterior, la búsqueda de **patrones genéticos inversos** entre diferentes condiciones o enfermedades. Esto es realmente útil para extraer las diferencias a nivel transcriptómico que separan o distinguen dos patologías, por ejemplo, se puede buscar los genes que están significativamente sobre-expresados en una enfermedad A y al mismo tiempo, infra-expresados en una enfermedad B con respecto a controles. Como ejemplo de esta aplicabilidad, Ibáñez et al. meta-analizaron varios conjuntos de datos procedentes de GEO para identificar patrones genéticos inversos entre Alzheimer y

cáncer, enfermedades las cuales poseen una comorbidad inversa (28). Esto significa que los pacientes que padecen una de las dos enfermedades, tienen un menor riesgo de padecer la otra. Pero esta aplicación no sólo permite la búsqueda de comportamientos genéticos o rutas biológicas diferenciales entre dos enfermedades, sino que puede ser usada en análisis de reutilización de fármacos, de los que hablaremos más adelante en más detalle. Brevemente, un análisis de reutilización de fármacos para una enfermedad determinada basado en meta-análisis de expresión genética compara datos de expresión de dicha enfermedad y perfiles genéticos de expresión inducidos por fármacos, hipotetizando que, si un fármaco induce un perfil genético inverso al de la enfermedad, éste podría ser capaz de revertir los patrones genéticos de la enfermedad, y por ende sus fenotipos patogénicos. Este tipo de métodos ayuda a la creación de hipótesis que requieren ser validadas experimentalmente.

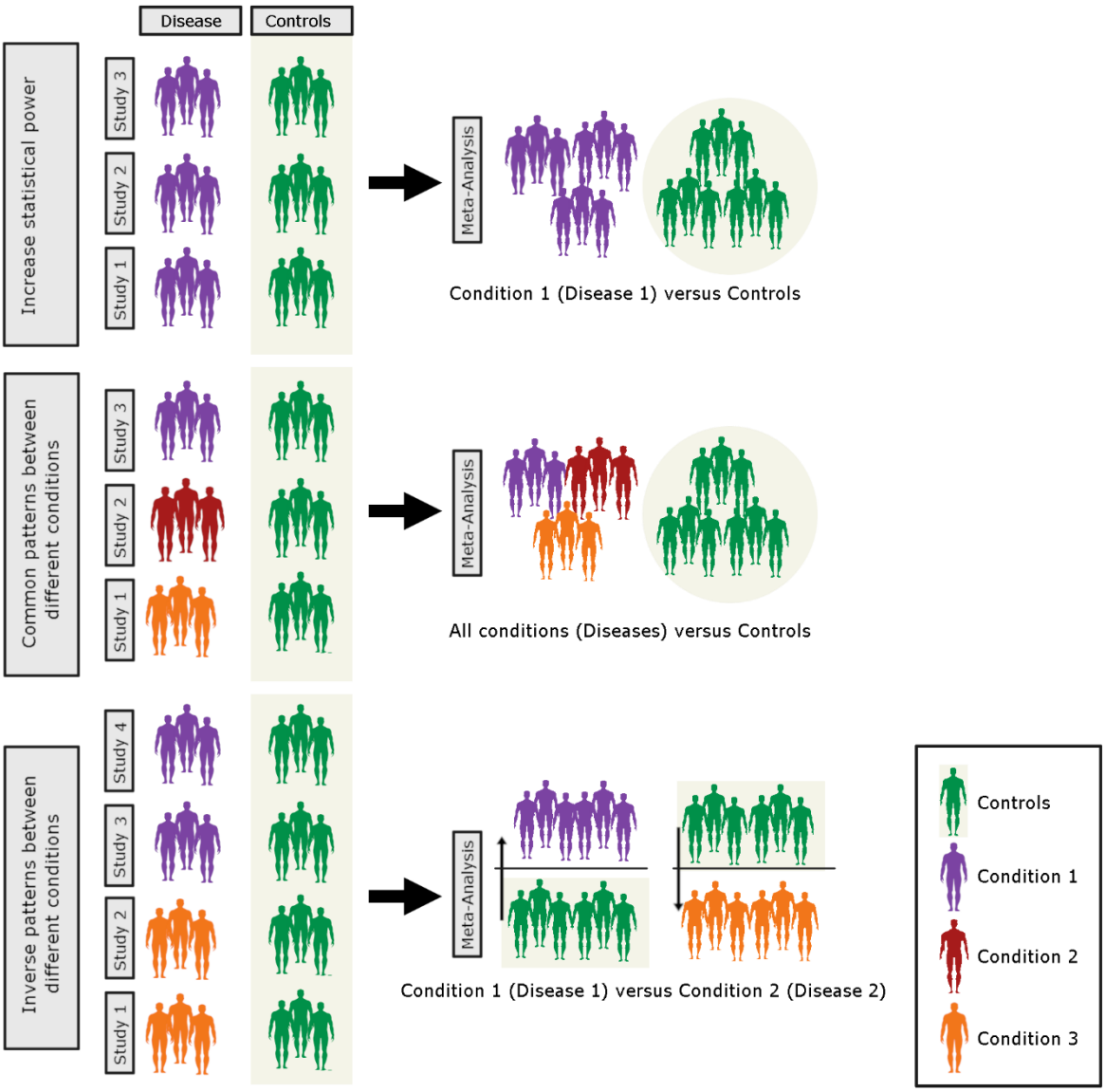


Figura 2: Resumen de las principales aplicaciones de los meta-análisis basados en expresión génica.

3.3. MÉTODOS

Entre los métodos de meta-análisis más usados encontramos dos principales ramas, los meta-análisis basados en tamaños de efectos y los meta-análisis basados en combinación de p-valores

(15). Además de éstos, existen otras técnicas como los métodos no paramétricos, que explicaremos en esta sección debido a que son utilizados en los trabajos descritos en esta tesis.

3.3.1. META-ANÁLISIS BASADOS EN TAMAÑO DE EFECTO

Definimos **tamaño de efecto** como la magnitud de un fenómeno dentro de un estudio, siendo la naturaleza de éste variable dependiendo del tipo de datos y análisis. Las técnicas de meta-análisis basadas en tamaño de efecto se basan en explicar la fuerza de este fenómeno o efecto entre diferentes clases, por ejemplo, entre muestras de pacientes sanos y muestras de una determinada enfermedad en un estudio concreto, para posteriormente combinar los efectos de los diferentes estudios y calcular el efecto combinado, que generalmente es una media ponderada. En el caso de los meta-análisis basados en expresión genética, el efecto que se calcula por estudio es la diferencia de expresión entre las dos clases para cada gen. Esta diferencia de expresión es calculada mediante una media tipificada, la cual se define como la diferencia de las medias de expresión de las muestras de una clase y de la otra clase, dividida por la desviación típica calculada tomando los valores de expresión para todas las muestras. Posteriormente, usando los efectos calculados en cada estudio y las varianzas de los mismos, se calcularía el efecto combinado para cada gen. Dependiendo de la **homogeneidad** de los datos entre los diferentes estudios, el cálculo del efecto combinado se realiza mediante dos técnicas diferentes. Cuando la homogeneidad es alta se utiliza un modelo de efectos fijos, mientras que entre estudios heterogéneos es más correcto usar un modelo de efectos aleatorios.

- Modelo de efectos fijos.

Se aconseja este modelo en el análisis de datos altamente homogéneos, por ejemplo, en el estudio de varios conjuntos de datos de la misma condición o enfermedad. En este caso, se supone que todos los estudios comparten un efecto común por gen y a cada uno de los estudios se le es asignado un peso relacionado directamente a su tamaño muestral, que consiste en el inverso de su varianza, premiando así o dando más importancia a los estudios con más muestras. Posteriormente se realiza el cálculo del efecto combinado mediante una media ponderada, o el sumatorio de los pesos por los efectos de un gen en todos los estudios, dividido por el sumatorio de los pesos. Por último, se representan los valores de los efectos combinados en una

distribución normal mediante el cálculo de los z-scores y a partir de éstos valores se calcula un nivel de significancia para cada gen.

- Modelo de efectos aleatorios

Este modelo se usa para estudios de meta-análisis con conjuntos de datos heterogéneos, por ejemplo, cuando se analizan *datasets* de diferentes enfermedades o *datasets* generados con distintas tecnologías. La principal diferencia con el modelo de efectos fijos es que en este caso no se asume un efecto común a todos los estudios, sino que el efecto es variable de un estudio a otro y, por tanto, en lugar de usar un efecto común, el efecto combinado representa la media de los efectos de los diferentes estudios. El cálculo del efecto combinado para cada gen se realiza del mismo modo que el caso anterior, con la salvedad de que en el peso que se le asigna a cada estudio se tiene en cuenta tanto la varianza en el estudio de manera individual como la varianza entre todos los estudios.

3.3.2. META-ANÁLISIS BASADOS EN COMBINACIÓN DE P-VALOR

Esta serie de técnicas se basan en la combinación de los p valores extraídos individualmente en cada estudio para cada gen. El **valor de p** (o p-valor) mide la significancia estadística de una hipótesis, por ejemplo, si el gen está diferencialmente expresado entre dos clases, y posteriormente se realiza la combinación de éstos entre todos los estudios para obtener un único valor de significancia, que nos dice si la hipótesis se cumple de manera consistente en todos los estudios. Estas técnicas son aconsejables cuando se estudian condiciones o enfermedades muy distintas entre sí y heterogéneas, las cuales no se asume a priori una relación directa. Estos meta-análisis trabajan con una lista de p-valores por estudio para el conjunto de genes, por lo que requieren análisis previo para la obtención de tales estadísticos y, por tanto, son más laxos en este sentido. Generalmente cuando se trabaja con datos de expresión, los p-valores individuales suelen calcularse mediante t de *student* o alguna de sus variantes y representan la significancia de la expresión diferencial entre dos clases, por ejemplo, entre muestras de individuos sanos y muestras de pacientes de una determinada enfermedad. Hay diferentes aproximaciones que permiten combinar los p-valores entre diferentes estudios, entre los más usados encontramos los siguientes.

- Método de Fisher

Este método se basa en el sumatorio del logaritmo negativo de los p-valores en todos los estudios. Este método es muy poco restrictivo y es muy influenciable dependiendo del número de estudios, generando muchos falsos positivos, o genes tomados como significativos cuando en realidad no lo deberían ser. Por este motivo, no se recomienda para meta-análisis con 5 o más estudios diferentes.

- Método de Stouffer

Método similar al método de Fisher, pero basado en *z-scores* en lugar de p-valores, lo que permite la incorporación de los diferentes pesos a cada estudio. Estos pesos pueden ser añadidos por el investigador en base por ejemplo a controles de calidad realizados previamente para cada estudio, a modo de poder dar más importancia a un estudio u otro y su influencia en los resultados.

- Método de Tippet

Este es un método simple basado en la selección para cada gen del p-valor mínimo obtenido para un estudio concreto.

- Método de Wilkinson

Esta técnica es similar a la anterior, pero selecciona para cada gen el p-valor máximo obtenido entre todos los estudios, siendo por tanto un método muy restrictivo en sus resultados, forzando a que los p-valores obtenidos de manera individual en cada estudio deban ser muy bajos.

3.3.3. META-ANÁLISIS NO PARAMÉTRICOS

Dentro de esta tercera categoría englobamos dos métodos muy usados que permiten tanto la combinación de p-valores entre diferentes estudios, como la combinación de cualquier otro estadístico o dato ordenable. Generalmente suele usarse un *fold change*, o magnitud que mide la diferencia de expresión, en nuestro caso, entre dos clases. Tras el cálculo del fold change, sus valores son sustituidos por los rangos. Esto es, el gen con el valor más pequeño de *fold change*

obtendría la posición o el valor 1 del *ranking* y el gen con el mayor valor obtendría la posición *n* (número total de genes en el estudio). De este modo, pasaríamos de listas con valores de *fold change* a listas de rangos con las posiciones que esos *fold changes* representan frente al total. Una lista es generada por estudio, combinándose todas en una matriz donde tenemos los genes en filas, y los diferentes estudios en columnas. Posteriormente se realiza su combinación y la obtención de un p-valor para cada gen.

- Producto de rangos (29,30)

La combinación de los rangos para cada gen es tan simple como el cálculo de una media geométrica de sus rangos entre los diferentes estudios. Posteriormente, se calcula un p-valor empírico mediante la aleatorización de los valores dentro de la matriz de rangos y volviendo a calcular el producto de rangos para los nuevos valores. Si el rango obtenido aleatoriamente en un gen es mayor que el original, se suma 1 al error de ese gen. Este proceso es permutado o repetido *k* veces, siendo finalmente el p-valor para un gen el error total obtenido dividido entre el número de permutaciones k.

- Suma de rangos (30)

Este método es similar al anterior, con la diferencia de que la combinación de rangos es calculada por media aritmética. Entre las ventajas con respecto al anterior se encuentra la facilidad de cómputo en análisis con un alto número de estudios, evitándose trabajar con números grandes resultado de la multiplicación de rangos.

3.4. PRIMER ARTÍCULO

Título: *Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis.*

Dirección web: https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-014-0489-x

En este artículo se realizó un meta-análisis usando 4 *datasets* de la base de datos NCBI GEO con datos de expresión de células sanguíneas de pacientes de LES, RA y SjS, con el objetivo de identificar patrones de expresión y rutas biológicas compartidas entre las tres enfermedades. La descarga, normalización, pre-procesamiento y filtrado de los datos fue realizado usando el lenguaje de programación R y posteriormente, usando los datasets procesados, utilizamos una herramienta web llamada Inmex (31) para realizar un meta-análisis basado en tamaño de efectos con el modelo de efectos aleatorios.

Los resultados nos revelaron un conjunto de 371 genes diferencialmente expresados en pacientes respecto a individuos sanos, los cuales son conservados en cada una de las 3 enfermedades y compartidos por las mismas. Muchos de los genes obtenidos han sido descritos previamente como biomarcadores de estas enfermedades de manera individual, aunque no de forma conjunta, lo que nos hace suponer que juegan un papel fundamental en el desarrollo o mantenimiento de la autoinmunidad más allá de las enfermedades, como genes relacionados con procesos inflamatorios, genes relacionados con la ruta de señalización mediante citoquinas o con la vía de señalización del IFN I. Pero otros genes y rutas biológicas asociadas no habían sido descritos o estudiados en detalle y, por tanto, nos proporcionan nuevas claves para entender los mecanismos patogénicos que dirigen este tipo de enfermedades, como la alteración en la expresión de ciertos genes implicados en procesos reguladores de apoptosis y ciclo celular en las células sanguíneas o la infra-expresión de genes ribosomales involucrados en la síntesis proteica.

Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis

Daniel Toro-Domínguez[1,2], Pedro Carmona-Sáez*[2] & Marta E Alarcón-Riquelme*[1,3]

[1]Area of Medical Genomics, Pfizer–Universidad de Granada–Junta de Andalucía de Genómica e Investigación Oncológica (GENyO), Parque Tenológico de la Salud Fundación (PTS) Granada, Avenida de la Ilustración, Granada, 114-18016, Spain

[2]Bioinformatics Unit, Pfize-Universidad de Granada-Junta de Andalucia, Centro de Genómica e investigación Oncológica (GENyO), Parque Tenológico de la Salud Fundación (PTS) Granada, Avenida de la Ilustración, Granada, 114-18016, Spain

[3]Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation, 825 NE 13th Street, Oklahoma City, 73104, OK, USA

*Corresponding authors

Abstract

Introduction: Systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) and Sjögren's syndrome (SjS) are inflammatory systemic autoimmune diseases (SADs) that share several clinical and pathological features. The shared biological mechanisms are not yet fully characterized. The objective of this study was to perform a meta-analysis using publicly available gene expression data about the three diseases to identify shared gene expression signatures and overlapping biological processes.

Methods: Previously reported gene expression datasets were selected and downloaded from the Gene Expression Omnibus database. Normalization and initial preprocessing were performed using the statistical programming language R and random effects model–based meta-analysis was carried out using INMEX software. Functional analysis of over- and underexpressed genes was done using the GeneCodis tool.

Results: The gene expression meta-analysis revealed a SAD signature composed of 371 differentially expressed genes in patients and healthy controls, 187 of which were underexpressed and 184 overexpressed. Many of these genes have previously been reported as significant biomarkers for individual diseases, but others provide new clues to the shared pathological state. Functional analysis showed that overexpressed genes were involved mainly in immune and inflammatory responses, mitotic cell cycles, cytokine-mediated signaling pathways, apoptotic processes, type I interferon–mediated signaling pathways and responses to viruses. Underexpressed genes were involved primarily in inhibition of protein synthesis.

Conclusions: We define a common gene expression signature for SLE, RA and SjS. The analysis of this signature revealed relevant biological processes that may play important roles in the shared development of these pathologies.

Introduction

Autoimmunity refers to the failure of the immune system to recognize its own constituent parts, eliciting an immune response against the tissues themselves. Currently, there are more than 80 clinically distinct autoimmune diseases [1], and the biological mechanisms that cause them are not clearly understood. It has been suggested that both genetic and environmental factors influence the development of autoimmune diseases [2]. Three of these inflammatory, autoimmune diseases are systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) and Sjögren's syndrome (SjS).

Although at first appearance these disorders have different phenotypes, they all are heterogeneous, multifactorial disorders that share molecular mechanisms which elicit similar clinical and pathogenic features. In fact, a differential diagnosis between these immune disorders at an early stage is not always reliable, and treatments are similar for all three, except when organ damage ensues or features of one dominate over another. Therefore, one of the major aims in this field is to discover similarities and differences at the molecular level between these diseases and between groups of patients across diseases. This will lead to a better understanding of the specific biological mechanisms and the development of more efficient and personalized treatments.

In this context, the analysis of gene expression patterns can provide useful information for understanding the molecular mechanisms by defining specific gene expression signatures that underlie these disorders. These studies are becoming more plentiful as a result of the development of high-throughput technologies such as microarrays and next-generation sequencing. These methods allow us to measure gene expression on a genome-wide scale, including for autoimmune diseases, and have been widely used during the past decade (see, for example, [3]-[5]). In this field, meta-analysis techniques offer the potential to integrate and jointly analyze data from different sources. In previous meta-analyses, investigators have combined data related to the same disease from different studies to get more consistent and reliable results. For example, Song et al. [6] performed a meta-analysis integrating three public SjS datasets. Arasappan et al. [7] conducted a pathway-based meta-analysis of four SLE datasets and identified a 37-gene signature associated with this disease. Olsen et al. [8],[9] analyzed different RA datasets and found several genes related to pathways such as type I interferon (IFN), apoptotic processes and cell cycles.

Moreover, meta-analytic techniques have also been used to integrate data from different diseases to uncover similar patterns. In this context, Tuller et al. [4] analyzed public data on six different autoimmune diseases (multiple sclerosis, SLE, juvenile RA, Crohn's disease, ulcerative colitis and type 1 diabetes) from peripheral blood mononuclear cells (PBMCs). Silva et al. [10] combined SLE and RA data to uncover coexpression patterns, and Higgs et al. [11],[12] integrated data on SLE, myositis, RA and scleroderma and defined a common type I IFN-related signature.

In this study, we performed a gene expression meta-analysis using publicly available gene expression data from PBMC samples of SLE, RA and SjS patients and controls. To the best of our knowledge, this

is the first study in which gene expression data from these three diseases have been integrated, together with analysis of the common gene expression signatures with respect to healthy controls.

Methods

Search and selection of datasets

We mined the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database [13] to find all publicly available gene expression datasets related to SLE, RA and SjS. From among all published studies, we selected for our analysis those that fit the following criteria: (1) They had to include control and case samples in separate arrays (one-channel arrays); (2) they had to have been performed with human PBMC samples; and (3) the samples had to have been obtained without any type of treatment.

Processing of the datasets and meta-analysis

Initial processing of the data was carried out using the R statistical programming language. Each dataset was downloaded from the NCBI GEO database using the GEOquery R package [14], and probes were annotated with the Entrez Gene identifiers, which were used to merge data from different platforms for further analysis. In each dataset, gene expression profiles were averaged for duplicate genes by computing the median values. Genes with missing values in more than 10% of samples were filtered out, and the remaining missing values were imputed using the average expression values within the group (case or control). The integration of different datasets and gene expression meta-analysis was performed using the INMEX software package [15]. Gene expression values were log-transformed and normalized by applying quantile normalization. The dataset for identification of genes specifically overexpressed in lupus CD4 T and B cells [GEO:GSE4588] contains samples from SLE and RA patients; therefore, these two subpopulations were treated as two different datasets.

Differential expression meta-analysis across diseases and healthy controls was carried out by using a random effects model (REM) [16],[17], which is based on combining the effect sizes (ESs) or changes of gene expression from different studies and obtaining an overall mean.

Functional analysis

In order to obtain biological information from the list of differentially expressed genes, we performed a functional analysis using the GeneCodis tool [18]-[21]. This software allows evaluation of which annotations are significantly enriched in a gene list, which can be used as functional descriptors of the biological processes that are acting in experimental conditions [19],[20]. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway annotations were evaluated in the analysis.

Results

Studies selected for the meta-analysis

After conducting a thorough search, we identified 19 datasets related to SLE, RA and SjS. From among this initial set, we dismissed three datasets that included treated samples, four datasets without control samples, five with too many missing values (more than 50% of missing values) and three generated with two-channel arrays. Thus, our final meta-analysis included four datasets (Table 1). The selected datasets comprise a total of 371 samples with a breakdown of 94 controls, 190 SjS samples, 54 SLE samples and 33 RA samples.

Identifying a common gene expression signature among systemic lupus erythematosus, rheumatoid arthritis and Sjögren's syndrome

For meta-analysis, the processed data were loaded into the INMEX web tool, and ES statistical analysis was used to find genes that were differentially expressed among diseases and healthy controls across different studies. Rather than requiring us to merge the original datasets, this method allowed us to combine them through high-level summary statistics, thus avoiding the problem of interstudy variation.

We identified 412 genes that were consistently differentially expressed ($P < 0.05$). Among the different studies (see Additional file 1), 210 genes were overexpressed and 202 were underexpressed.

Among this initial set, we found 187 overexpressed genes in all diseases with respect to healthy controls and 184 that were underexpressed. Additional file 1 gives the average fold changes of each gene in all datasets. This set comprises the common gene expression signature—that is, genes that are significantly differentially expressed in all diseases with respect to healthy controls. Figure 1 shows the top 50 over- and underexpressed genes.

Some of the most differentially expressed genes were HERC6, which belongs to the HERC family of E3 ubiquitin ligases, and RTP4, which encodes a receptor (chemosensory) transporter protein related to "chemotaxis" and was previously described as an IFN-inducible gene [25]. In addition, the most overexpressed genes with the largest ESs were RSAD2 and IFI44L. IFI44L encodes the IFN-induced protein 44-like, which has been described in several autoimmune diseases in conjunction with other genes involved in the type I IFN signaling pathway, such as IFIT1, IFI27 and IFITM1. We also found overexpression of these genes in our results.

The gene with the lowest ES (ES = −1.2545) was eukaryotic translation elongation factor 2 (EEF2), a biomarker protein of some types of cancer [26] that plays an important role in protein synthesis. This was, in fact, the most relevant pathway that was associated with underexpressed genes.

Meta-analytic techniques have been also used to evaluate reproducibility and bias across microarray studies. This is especially important when comparing replicated samples or samples of the same condition or phenotype. In this sense, there are different methods that can be used for this purpose [27]. In this context, we also evaluated the meta-analysis results with those obtained from individual analyses of studies and/or diseases. We found 132 gained genes and 2,168 lost genes in our meta-analysis (see Additional file 1). Gained genes are the differentially expressed genes identified only in the meta-analysis and not in the individual analysis, because they show weak signals but consistent expression patterns across the different datasets. Lost genes are genes identified as differentially expressed genes in any individual analysis, but not in the meta-analysis. These genes show either conflicting changes in expression profiles or very large variations across different studies [6],[15]. Additional file 2 contains a detailed study of the different datasets and the analysis used in our study.

Functional and pathway analysis

For the analysis of biological processes associated with the differentially expressed genes, we evaluated the enrichment of functional annotations using the GeneCodis tool [18]. GO annotations for biological processes were significantly overrepresented in the gene list if they showed a P-value <0.05. Results for biological pathways of overexpressed and underexpressed genes are shown in Figure 2 and Additional file 1. Functions such as “mitotic cell cycle,” “cytokine-mediated signaling pathway,” “response to virus” or “type I IFN-mediated signaling pathway” or “immune response” were significantly associated with this set of genes. This is in agreement with previous work that has associated these pathways with each of the diseases [12],[28]-[31].

Similarly, the most significant GO categories or pathways in the analysis of underexpressed genes were “gene expression” and others related to protein biosynthesis mechanisms previously reported [32].

Discussion

In this study, we define a signature of differentially expressed genes for SLE, RA and SjS using a gene expression meta-analytic strategy showing common biological mechanisms across three otherwise clinically separate entities. The combined ES and REM meta-analytic method was chosen because it allowed us to integrate microarray datasets from different platforms consistently, without the obstacle of the batch effects that we clearly observed when we began our analyses. We performed the meta-analysis using four publicly available datasets and defined a common signature composed of 187 overexpressed genes and 184 underexpressed genes in all diseases compared to healthy controls. We found significant pathways related to overexpressed genes, such as “immune response,” “type I IFN-mediated signaling pathway,” “cytokine-mediated signaling pathway,” “mitotic cycle” and “response to virus,” as well as pathways related to underexpressed genes that highlight gene expression and metabolic processes.

In this context, independent studies of SLE, RA and/or SjS have shown that overexpression of type I IFN-related genes is very consistent [28]-[31],[33],[34], and it appears that these genes roughly form a common pattern in all inflammatory autoimmune diseases [4],[6],[7],[9]-[12]. The cytokine signaling pathway-related genes behave similarly. In fact, both routes are tightly related. IFN proteins regulate several signaling pathways normally in response to pathogens, such as apoptosis or immune stimulation. In the gene expression signature, we identified several type I IFN-related genes as the most significantly overexpressed genes in all diseases, such as IFI44L, IFI44, IFI27 and IFIT1. We also found other genes, such as JAK2, involved in the Janus kinase/signal transducer and activator of transcription (JAK/STAT) signaling pathway and previously related to immune disorders [35]. JAK/STAT signaling promotes IFN-stimulated gene transcription. OASL is and IFN-inducible protein related to antiviral activity [36]. In addition, we found genes related to apoptosis, with genes such as FAS, which has been described as a risk allele in some autoimmune diseases [37],[38]; TNFSF10; and CASP1. We also identified several proteasome subunits, such as PSM2, PSM6 and PSMC2. Apoptotic processes have been related to autoimmune diseases in different studies [39]. Apoptotic cells are not immunologically neutral, and the accumulation of apoptotic material not properly phagocytosed can be an important source of autoimmune antigens, enabling the development of autoimmune disorders [40]. In this context, there are some hypotheses focused on the increase in apoptosis, lazy phagocytes or the interaction between apoptotic material and antigen-presenting cells as potential triggers for these diseases.

In addition, some of these genes have been described previously as biomarkers of one or a variety of autoimmune diseases, such as IFN-induced protein 44-like, chemokine receptor 1 and FAS. Therefore, our results are consistent with previously published data for each of the three disorders, but show, for the first time to our knowledge, and formally, their shared genetic signatures.

Moreover, we found interesting results, such as the overexpression of EIF2AK2 gene (or PKR) (see Additional file 1). This gene is initially related to the response to virus and the innate immune response and encodes a serine/threonine protein kinase that is activated by autophosphorylation after binding to double-stranded RNA [41]. The activated form can phosphorylate multiple substrates, including several translation initiation factors, such as eIF2A, eIF3F eIF2S1 and eEF2, impairing the recycling of these factors between successive rounds of initiation and leading to inhibition of translation, which eventually results in shutdown of cellular protein synthesis and a reduction in cell proliferation [42],[43]. This is in agreement with our finding of biological processes related to inhibition of gene expression and protein synthesis in the list of underexpressed genes.

In a previous study of SLE, Groulleau et al. [32] described the relationship between PKR and the phosphorylation of the eIF2A translation initiation factor, but this action was attributed only to SLE, whereas researchers in other studies independently related PKR with RA [44],[45]. In addition, PKR phosphorylates p38, JNK and nuclear factor κB (NF-κB), which are proteins of the mitogen-activated protein kinase signaling pathway related to the production of cytokines and tumor necrosis factor [46]. These in turn intervene in apoptotic processes, regulation of signal transduction or cell proliferation

and differentiation. PKR has direct influences on the production of IFNs [47],[48]. We also found other genes related to the immune system, such as MYD88, which is an adaptor protein of Toll-like receptors that activates NF-κB and translocates to the nucleus to stimulate the expression of certain genes for the production of cytokines and IFN proteins [49].

Regarding the pathways related to underexpressed genes, the two most relevant processes were "gene expression" and "cellular protein metabolic process". Analysis of genes associated with these annotations revealed that many were genes involved in translation, such as ribosomal protein-encoding genes, and different eukaryotic translation initiation factor subunits, such as eEF2 or eIF3F mentioned above. The relationship between the underexpression of these genes and autoimmune disease is largely undefined. We also found genes involved in translation and cell growth, which are underexpressed, as mentioned above.

Conclusions

We performed a gene expression meta-analysis using previously published datasets obtained from PBMCs of SLE, RA and SjS patients. A common gene expression signature was defined, comprising many genes that have been previously related to one, two or each of the three diseases. Although there are previous gene expression meta-analyses of immune-related diseases, our present study is the first one, to our knowledge, in which data on these three specific disorders have been integrated, which allowed us to define common biological processes. We found that pathways in our results, such as "type I IFN-mediated signaling pathway," apoptotic processes, "immune response," reduction in translation processes and "response to virus." This suggests that a majority of these pathways are related to the action of the IFN proteins. However, we found other pathways, such as mitotic cell cycles, whose relationship to the IFN pathway, or even to the diseases themselves, has not been described. Future functional and specific studies of the genes we identified are needed to define the roles of these genes in the pathogenesis of SADs.

References

1. Karopka T, Fluck J, Mevissen HT, Glass Ä: The Autoimmune Disease Database: a dynamically compiled literature-derived database. BMC Bioinformatics. 2006, 7: 325-10.1186/1471-2105-7-325.

2. Salaman MR: A two-step hypothesis for the appearance of autoimmune disease. Autoimmunity. 2003, 36: 57-61. 10.1080/0891693031000068637.

3. Burska AN, Roget K, Blits M, Soto Gomez L, van de Loo F, Hazelwood LD, Verweij CL, Rowe A, Goulielmos GN, van Baarsen LGM, Ponchel F: Gene expression analysis in RA: towards personalized medicine. Pharmacogenomics J. 2014, 14: 93-106. 10.1038/tpj.2013.48.

4. Tuller T, Atar S, Ruppin E, Gurevich M, Achiron A: Common and specific signatures of gene expression and protein–protein interactions in autoimmune diseases. Genes Immun. 2013, 14: 67-82. 10.1038/gene.2012.55.

5. van 't Veer LJ, Bernards R: Enabling personalized cancer medicine through analysis of gene-expression patterns.Nature 2008, 452:564–570.,

6. Song GG, Kim JH, Seo YH, Choi SJ, Ji JD, Lee YH: Meta-analysis of differentially expressed genes in primary Sjogren's syndrome by using microarray. Hum Immunol. 2014, 75: 98-104. 10.1016/j.humimm.2013.09.012.

7. Arasappan D, Tong W, Mummaneni P, Fang H, Amur S: Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. BMC Med. 2011, 9: 65-10.1186/1741-7015-9-65.

8. Olsen N, Sokka T, Seehorn CL, Kraft B, Maas K, Moore J, Aune TM: A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. Ann Rheum Dis. 2004, 63: 1387-1392. 10.1136/ard.2003.017194.

9. Olsen NJ, Moore JH, Aune TM: Gene expression signatures for autoimmune disease in peripheral blood mononuclear cells. Arthritis Res Ther. 2004, 6: 120-128. 10.1186/ar1190.

10. Silva GL, Junta CM, Mello SS, Garcia PS, Rassi DM, Sakamoto-Hojo ET, Donadi EA, Passos GAS: Profiling meta-analysis reveals primarily gene coexpression concordance between systemic lupus erythematosus and rheumatoid arthritis. Ann N Y Acad Sci. 2007, 1110: 33-46. 10.1196/annals.1423.005.

11. Higgs BW, Liu Z, White B, Zhu W, White WI, Morehouse C, Brohawn P, Kiener PA, Richman L, Fiorentino D, Greenberg SA, Jallal B, Yao Y: Patients with systemic lupus erythematosus, myositis, rheumatoid arthritis and scleroderma share activation of a common type I interferon pathway. Ann Rheum Dis. 2011, 70: 2029-2036. 10.1136/ard.2011.150326.

12. Higgs BW, Zhu W, Richman L, Fiorentino DF, Greenberg SA, Jallal B, Yao Y: Identification of activated cytokine pathways in the blood of systemic lupus erythematosus, myositis, rheumatoid arthritis, and scleroderma patients. Int J Rheum Dis. 2012, 15: 25-35. 10.1111/j.1756-185X.2011.01654.x.

13. National Center for Biotechnology Information: Gene Expression Omnibus. [http://www.ncbi.nlm.nih.gov/geo/] (accessed 17 December 2014).

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

15. Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock REW: INMEX—a web-based tool for integrative meta-analysis of expression data.Nucleic Acids Res 2013, 41(Web server issue):W63–W70.,

16.Choi JK, Yu U, Kim S, Yoo OJ: Combining multiple microarray studies and modeling interstudy variation. Bioinformatics. 2003, 19: i84-i90. 10.1093/bioinformatics/btg1010.

17. Marot G, Foulley JL, Mayer CD, Jaffrézic F: Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics. 2009, 25: 2692-2699. 10.1093/bioinformatics/btp444.

18. GeneCodis: [http://genecodis.cnb.csic.es/] (accessed 17 December 2014).

19. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. 2007, 8: R3-10.1186/gb-2007-8-1-r3.

20. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A: GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information.Nucleic Acids Res 2009, 37(Web server issue):W317–W322.,

21. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A: GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics.Nucleic Acids Res 2012, 40(Webserver issue):W478–W483.,

22. Hutcheson J, Scatizzi JC, Siddiqui AM, Haines GK, Wu T, Li QZ, Davis LS, Mohan C, Perlman H: Combined deficiency of proapoptotic regulators Bim and Fas results in the early onset of systemic autoimmunity. Immunity. 2008, 28: 206-217. 10.1016/j.immuni.2007.12.015.

23. Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR, Hilliquin P, Gut I, Cornelis F, Petit-Teixeira E: Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. PLoS One. 2009, 4: e6803-10.1371/journal.pone.0006803.

24. Lessard CJ, Li H, Adrianto I, Ice JA, Rasmussen A, Grundahl KM, Kelly JA, Dozmorov MG, Miceli-Richard C, Bowman S, Lester S, Eriksson P, Eloranta ML, Brun JG, Gøransson LG, Harboe E, Guthridge JM, Kaufman KM, Kvarnström M, Jazebi H, Graham DSC, Grandits ME, Nazmul-Hossain ANM, Patel K, Adler AJ, Maier-Moore JS, Farris AD, Brennan MT, Lessard JA, Chodosh J: Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. Nat Genet. 2013, 45: 1284-1292. 10.1038/ng.2792.

25. Yao Y, Higgs BW, Morehouse C, de los Reyes M, Trigona W, Brohawn P, White W, Zhang J, White B, Coyle AJ, Kiener PA, Jallal B: Development of potential pharmacodynamic and diagnostic markers for anti-IFN-α monoclonal antibody trials in systemic lupus erythematosus. Hum Genomics Proteomics. 2009, 1: 374312-

26. Sun HG, Dong XJ, Lu T, Yang MF, Wang XM: Clinical value of eukaryotic elongation factor 2 (eEF2) in non-small cell lung cancer patients. Asian Pac J Cancer Prev. 2014, 14: 6533-6535. 10.7314/APJCP.2013.14.11.6533.

27. Tseng GC, Ghosh D, Feingold E: Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012, 40: 3785-3799. 10.1093/nar/gkr1265.

28. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, Shark KB, Grande WJ, Hughes KM, Kapur V, Gregersen PK, Behrens TW: Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc Natl Acad Sci U S A. 2003, 100: 2610-2615. 10.1073/pnas.0337679100.

29. Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, Pascual V: Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med. 2003, 197: 711-723. 10.1084/jem.20021553.

30. Bezalel S, Guri KM, Elbirt D, Asher I, Sthoeger ZM: Type I interferon signature in systemic lupus erythematosus. Isr Med Assoc J. 2014, 16: 246-249.

31. Salloum R, Niewold TB: Interferon regulatory factors in human lupus pathogenesis. Transl Res. 2011, 157: 326-331. 10.1016/j.trsl.2011.01.006.

32. Grolleau A, Kaplan MJ, Hanash SM, Beretta L, Richardson B: Impaired translational response and increased protein kinase PKR expression in T cells from lupus patients. J Clin Invest. 2000, 106: 1561-1568. 10.1172/JCI9352.

33. Brkic Z, Versnel MA: Type I IFN signature in primary Sjögren's syndrome patients. Expert Rev Clin Immunol. 2014, 10: 457-467. 10.1586/1744666X.2014.876364.

34. Le Page C, Génin P, Baines MG, Hiscott J: Interferon activation and innate immunity. Rev Immunogenet. 2000, 2: 374-386.

35. Malemud CJ: Intracellular signaling pathways in rheumatoid arthritis. J Clin Cell Immunol. 2013, 4: 160-10.4172/2155-9899.1000160.

36. Zhu J, Zhang Y, Ghosh A, Cuevas RA, Forero A, Dhar J, Ibsen MS, Schmid-Burgk JL, Schmidt T, Ganapathiraju MK, Fujita T, Hartmann R, Barik S, Hornung V, Coyne CB, Sarkar SN: Antiviral activity of human OASL protein is mediated by enhancing signaling of the RIG-I RNA sensor. Immunity. 2014, 40: 936-948. 10.1016/j.immuni.2014.05.007.

37. Lu MM, Ye QL, Feng CC, Yang J, Zhang T, Li J, Leng RX, Pan HF, Yuan H, Ye DQ: Association of FAS gene polymorphisms with systemic lupus erythematosus: a case–control study and meta-analysis. Exp Ther Med. 2012, 4: 497-502.

38. Treviño-Talavera BA, Palafox-Sánchez CA, Muñoz-Valle JF, Orozco-Barocio G, Navarro-Hernández RE, Vázquez-Del Mercado M, García de la Torre I, Oregon-Romero E: FAS −670A>G promoter polymorphism is associated with soluble Fas levels in primary Sjögren's syndrome. Genet Mol Res. 2014, 13: 4831-4838. 10.4238/2014.July.2.12.

39. Colonna L, Lood C, Elkon KB: Beyond apoptosis in lupus. Curr Opin Rheumatol. 2014, 26: 459-466. 10.1097/BOR.0000000000000083.

40. Biermann MH, Veissi S, Maueröder C, Chaurio R, Berens C, Herrmann M, Munoz LE: The role of dead cell clearance in the etiology and pathogenesis of systemic lupus erythematosus: dendritic cells as potential targets. Expert Rev Clin Immunol. 2014, 10: 1151-1164. 10.1586/1744666X.2014.944162.

41. Zhang S, Sun Y, Chen H, Dai Y, Zhan Y, Yu S, Qiu X, Tan L, Song C, Ding C: Activation of the PKR/eIF2α signaling cascade inhibits replication of Newcastle disease virus. Virol J. 2014, 11: 62-10.1186/1743-422X-11-62.

42. Bullido MJ, Martínez-García A, Tenorio R, Sastre I, Muñoz DG, Frank A, Valdivieso F: Double stranded RNA activated EIF2α kinase (EIF2AK2; PKR) is associated with Alzheimer's disease. Neurobiol Aging. 2008, 29: 1160-1166. 10.1016/j.neurobiolaging.2007.02.023.

43. Sadler AJ, Williams BRG: Structure and function of the protein kinase R. Curr Top Microbiol Immunol. 2007, 316: 253-292.

44. Gilbert SJ, Duance VC, Mason DJ: Does protein kinase R mediate TNF-α- and ceramide-induced increases in expression and activation of matrix metalloproteinases in articular cartilage by a novel mechanism?. Arthritis Res Ther. 2003, 6: R46-10.1186/ar1024.

45. Gilbert SJ, Duance VC, Mason DJ: Protein kinase R: a novel mediator of articular cartilage degradation in arthritis. Curr Rheumatol Rev. 2006, 2: 9-21. 10.2174/157339706775697026.

46. Yeung MC, Liu J, Lau AS: An essential role for the interferon-inducible, double-stranded RNA-activated protein kinase PKR in the tumor necrosis factor-induced apoptosis in U937 cells. Proc Natl Acad Sci U S A. 1996, 93: 12451-12455. 10.1073/pnas.93.22.12451.

47. Ward SV, Samuel CE: The PKR kinase promoter binds both Sp1 and Sp3, but only Sp3 functions as part of the interferon-inducible complex with ISGF-3 proteins. Virology. 2003, 313: 553-566. 10.1016/S0042-6822(03)00347-7.

48. McAllister CS, Taghavi N, Samuel CE: Protein kinase PKR amplification of interferon β induction occurs through initiation factor eIF-2α-mediated translational control. J Biol Chem. 2012, 287: 36384-36392. 10.1074/jbc.M112.390039.

49. Kumar H, Kawai T, Akira S: Toll-like receptors and innate immunity. Biochem Biophys Res Commun. 2009, 388: 621-625. 10.1016/j.bbrc.2009.08.062.

Abbreviations

ES: Effect size, *GEO*: Gene Expression Omnibus, *GO*: Gene ontology, *IFN*: Interferon, *JAK/STAT*: Janus kinase/signal transducer and activator of transcription, *NCBI*: National Center for Biotechnology

Information, *NF-κB*: Nuclear factor κB, *PBMC*: Peripheral blood mononuclear cell, *RA*: Rheumatoid arthritis, *REM*: Random effects model, *SAD*: Inflammatory and systemic autoimmune disease, *SjS*: Sjögren's síndrome, *SLE*: Systemic lupus erythematosus.

Acknowledgements

This study was possible thanks to grants from the Instituto de Salud Carlos III (IP12/02558), supported in part by European Regional Development Fund (to MAR), and from Innovative Medicines Initiative (grant GA-115565, coordinated by MAR).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PCS and MAR conceived of and supervised the project. PCS and DTD designed the study and the data analysis pipeline. DTD implemented the meta-analysis workflow and performed the analysis. PCS, MAR and DTD contributed to the interpretation of data. All authors wrote and revised the manuscript critically. All authors read and approved the final manuscript.

Figures and Tables

Figure 1: Heatmap of top differentially expressed genes. The heatmap represents the log2- transformed expression values with the top 50 overexpressed (top) and top 50 underexpressed (below) genes.

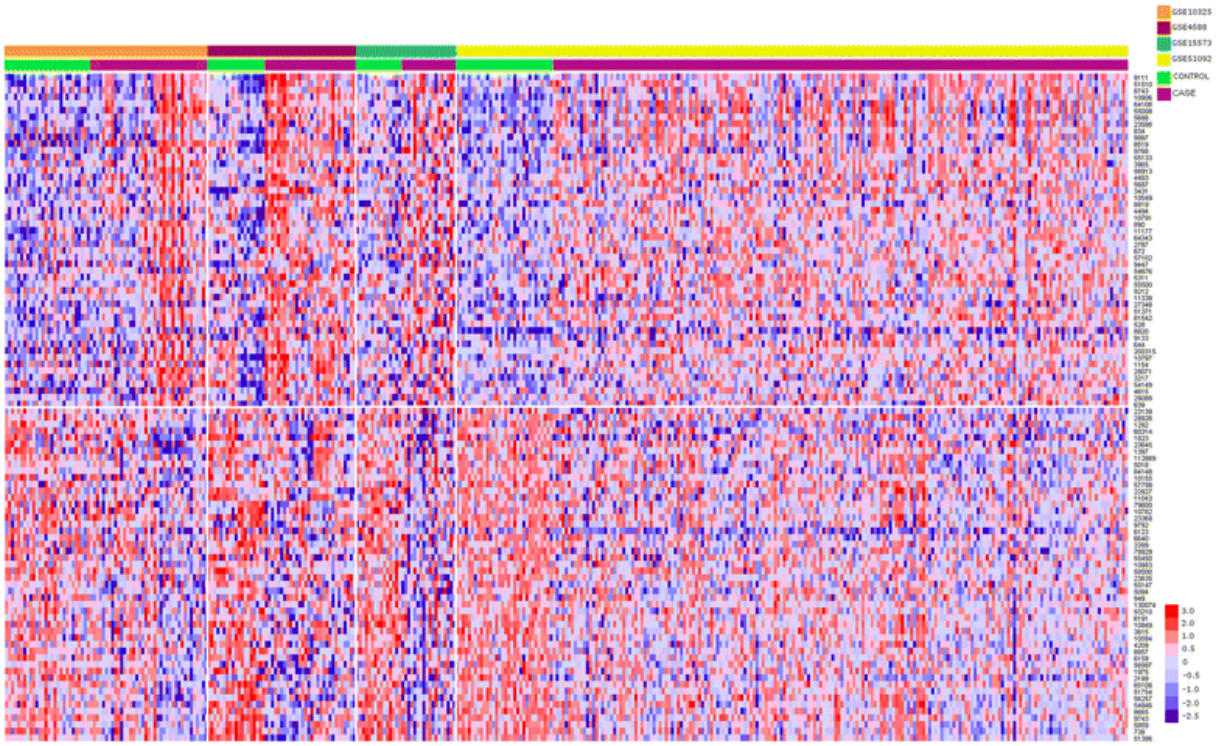


Figure 2: Biological function related to the differentially expressed genes. This graphic shows the main Gene Ontology (GO) biological functions identified and related to overexpressed genes (A) and underexpressed genes (B). GO annotations were considered significantly enriched in the list of genes if they had a P-value <0.01 and were associated with at least ten genes. The x-axis represents the number of genes, and the y-axis shows the names of the significant GO categories sorted by decreasing P-values.

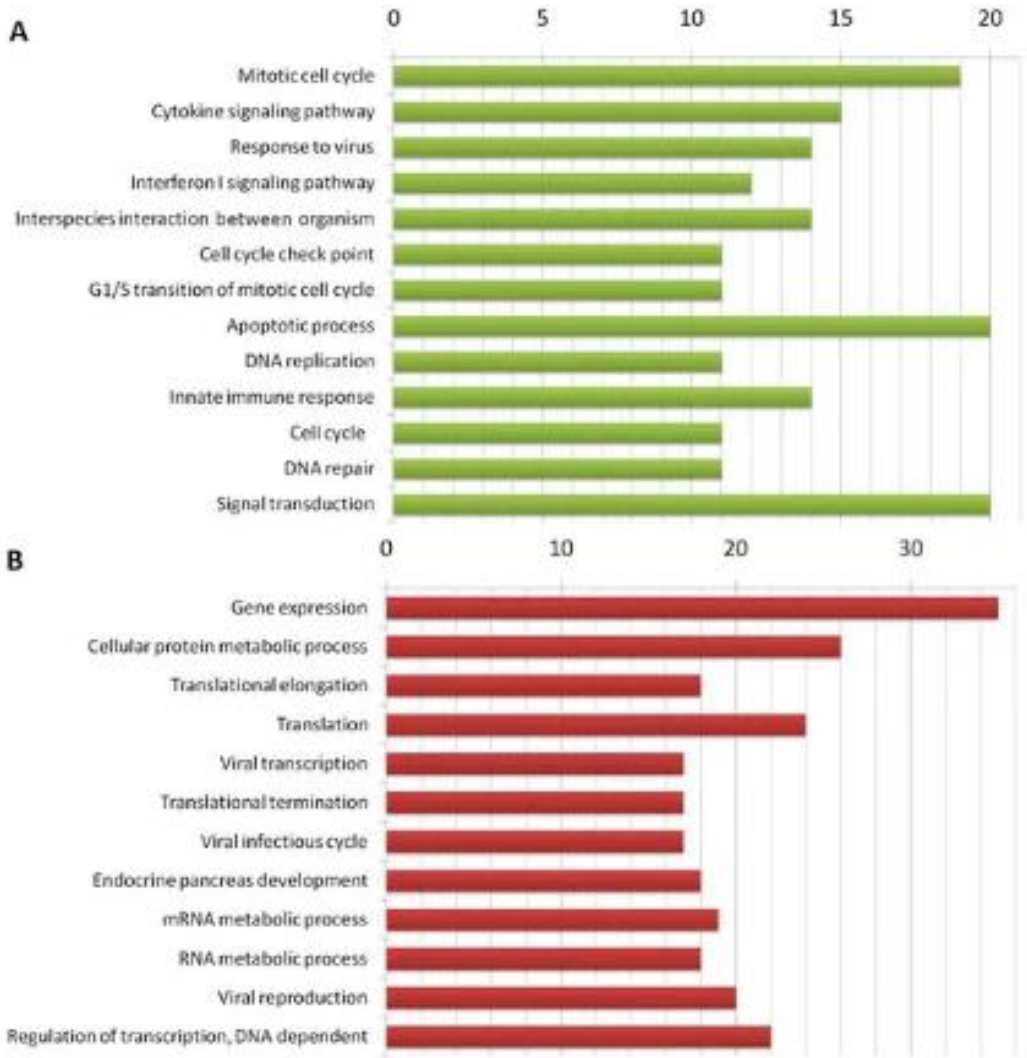


Table 1: Datasets used in the study. NCBI GEO ID: Unique gene expression series identifier for each dataset in the National Center for Biotechnology Information Gene Expression Omnibus database; Platform: Microarray platform; Disease: Type of disease and number of cases/controls in each dataset;

Table 1 Datasets used in the study[a]

GEO ID	Platform	Disease	Cases/controls	Description	Reference	Key findings
[GEO:GSE10325]	Affymetrix Human Genome U133A Array	SLE	39/28	Expression data from human peripheral blood subsets	[22]	Some apoptotic genes are related to SLE phenotype
[GEO:GSE15573]	Affymetrix Human Genome U133 Plus 2.0 Array	RA	18/15	Immunity and defense genes in PBMCs of RA patients	[23]	283 underexpressed genes (involved in metabolic processes), 101 overexpressed (immunity, calcium transport)
[GEO:GSE4588]	Illumina Human 6 v2.0 expression BeadChip	SLE, RA	15 SLE, 15 RA/19	Identification of genes specifically overexpressed in SLE CD4 T and B cells	-	-
[GEO:GSE51092]	Illumina Human WG-6 v3.0 expression BeadChip	SS	190/32	Variants at multiple loci implicated in both innate and adaptive immune responses are associated with SS	[24]	Risk loci for SS (related to IFN pathway, STAT4, chemokine receptors, interleukins proteins or in BLK

Description: Brief description of the study; Reference: Publication; Key findings: Main findings in the original studies. BLK, B Lymphocyte Kinase Protein; IFN, Interferon; PBMCs, Peripheral blood mononuclear cells; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; SjS, Sjögren's syndrome; STAT, Signal transducers and activators of transcription.

Additional Material

Additional file 1: Four Excel spreadsheets with results from the meta-analysis. Spreadsheet 1: Differential gene expression analysis results with list of the names of the genes and their Entrez ID, the value of the combined ESs among the different experiments for each gene and the associated P-values. The left column shows the upexpressed genes, and the right column shows the underexpressed genes. Spreadsheet 2: The fold changes of each gene in all datasets. Genes marked with an asterisk were removed from the gene expression signature, as they did not show a consistent pattern of overexpression or underexpression in all diseases. Spreadsheet 3: Gain and loss gene lists. Spreadsheet 4: All GO categories enriched in the gene signature are listed, with the minimum number of genes set to three. The table contains the number of genes in each pathway, the corrected P-values (Hyp*), the GO annotation numbers and the names of the pathways and the genes related to each pathway.

Additional file 2: Detailed study of the different datasets used and their influence in the final results of the meta-analysis.

4. ANÁLISIS DE REUTILIZACIÓN DE FÁRMACOS

4.1. INTRODUCCIÓN AL ANÁLISIS DE REUTILIZACIÓN DE FÁRMACOS

La estrategia principal para el desarrollo y descubrimiento de fármacos durante las últimas décadas se ha basado en el análisis masivo de miles de moléculas simultáneamente para identificar aquellos compuestos que muestran actividad contra los objetivos terapéuticos estudiados (32). Estos métodos están sujetos a que cuanto más medicamentos o moléculas se analicen, más probable es que se encuentre un resultado alentador, pero este enfoque de ensayo y error supone un enorme coste económico y una eficacia baja. Motivo que afirma la inviabilidad del establecimiento de esta técnica como doctrina en el campo y evoca a la necesidad de uso de técnicas alternativas más baratas y eficientes. Si el sistema de estudio está bien caracterizado molecularmente, por ejemplo, si se desea desarrollar un fármaco para una determinada enfermedad y se tiene información robusta y suficiente sobre objetivos genéticos implicados en la misma, se puede reducir el rango de medicamentos que se testean a aquellos relacionados previamente de algún modo con el sistema, lo que aumenta la ratio eficiencia / costo. Sin embargo, se perderá el posible descubrimiento de nuevas y quizás más eficientes terapias con fármacos o moléculas no relacionadas previamente con el sistema de estudio. Además, la eficiencia de estos enfoques orientados a objetivos genéticos concretos, o enfoques locales, es más baja aún para las enfermedades poligénicas y complejas, debido a que éstas a

menudo alteran el sistema biológico en diferentes puntos y por diferentes vías. A esto se suma que la heterogeneidad genética puede contribuir a que un fármaco candidato sea inefectivo en la mayoría de los casos. Por lo tanto, debe tenerse en cuenta la heterogeneidad de la población, algo que apenas se hace en las primeras etapas de las pruebas *in vitro* clásicas. Por último, en estos enfoques locales se pierde la información de posibles efectos adversos ya que suele medirse el impacto de un medicamento sobre un sistema concreto, por ejemplo, una ruta patogénica, sin evaluar el efecto del mismo a nivel global.

En este contexto, los **análisis de reutilización de fármacos** son una potencial alternativa para el descubrimiento de medicamentos potenciales, siendo algunas de sus técnicas capaces de solucionar desde algunos a todos los problemas mencionados anteriormente. Un análisis de reutilización de fármacos consiste en buscar nuevas aplicaciones para fármacos ya existentes (33). Históricamente, el concepto de reutilización de fármacos surge de la serendipia o la casualidad (34), siendo algunos ejemplos clásicos la penicilina o el Sildenafil, actualmente utilizado contra la disfunción eréctil, pero inicialmente utilizado para tratar la hipertensión (35). Estas técnicas reducen drásticamente el tiempo y el coste con respecto a los métodos tradicionales debido a que, al ser medicamentos conocidos y a menudo ya aprobados para su uso clínico, no es necesario la realización posterior de los ensayos clínicos de Fase I (36) y los perfiles farmacocinéticos, farmacodinámicos y de toxicidad de los medicamentos ya se conocen. A esto debe sumarse el desarrollo de las técnicas de generación masiva de datos ómicos, que han permitido a los investigadores desarrollar nuevos enfoques computacionales aún más económicos y eficientes. Algunos de estos métodos tienen un enfoque local y se basan en conocimientos previos para explorar las propiedades compartidas entre compuestos, por ejemplo datos de estructura química (37), información sobre efectos secundarios (38,39) o información relacionada con los mecanismos de acción o vías terapéuticas (40,41). Sin embargo, estos métodos se centran principalmente en encontrar nuevos compuestos potenciales que compartan algunas propiedades con respecto a un fármaco de referencia usado en un contexto determinado, pero no pueden proporcionar las conexiones entre estos nuevos fármacos y los fenotipos de las enfermedades.

También encontramos técnicas con un **enfoque global** del efecto del fármaco y es aquí donde entra en juego el uso de la transcriptómica en los análisis de reutilización de fármacos. La

aplicación de un fármaco dado en una célula induce una firma específica de expresión génica, es decir, un conjunto de genes que se sobre-expresan o se infra-expresan (42), reflejando el efecto del fármaco en la célula. Por lo tanto, comparar las **firmas de expresión** génica nos permitirá establecer conexiones de efectos moleculares tanto entre diferentes fármacos como con las firmas genéticas de enfermedades.

Partiendo de una firma de expresión génica causada por una enfermedad, es decir, el conjunto de genes expresados diferencialmente en el estado patológico respecto al fenotipo normal, podríamos compararla con firmas de expresión génica inducidas por una gran multitud de compuestos y medir una correlación entre ellas. Una correlación inversa o negativa entre los perfiles de un fármaco y una enfermedad significa que los genes sobre-expresados por una se encuentran infra-expresados por la otra y viceversa; es decir, el fármaco causa una firma transcripcional opuesta a la de la enfermedad. Esto nos da pie a hipotetizar que un fármaco *x* puede revertir la firma de expresión derivada de la enfermedad y, por tanto, el fenotipo patogénico, más allá de objetivos genéticos concretos o rutas biológicas aisladas (43). Por otro lado, una correlación positiva podría implicar que el fármaco particular es un potencial inductor de la enfermedad, lo que tiene aplicaciones tanto en la generación de modelos *in vitro* o *in vivo* como además en la profundización del conocimiento en los mecanismos que desarrollan la patogenia. Igualmente, correlaciones positivas entre distintos fármacos nos permite encontrar fármacos con efectos similares a nivel transcripcional. De esta manera se abren multitud de hipótesis a probar a nivel experimental.

El uso de estas técnicas ha supuesto un gran avance en el campo, cosa que podemos ver reflejada en el número creciente de publicaciones relacionadas. Por ejemplo, con la búsqueda "*drug repurposing*" en PubMed, el número de artículos ha aumentado de 15 artículos publicados en 2009 a 848 publicados durante el 2018. En relación al uso de la transcriptómica para la reutilización de fármacos, encontramos el trabajo pionero de Hughes et al. donde generaron un compendio de perfiles de expresión de fármacos en Saccharomyces cerevisiae y mostraron el potencial de explorar similitudes entre las firmas de expresión génica como alternativa a los métodos tradicionales (44). Algunos años después, Lamb et al. (45) desarrollaron la herramienta Connectivity Map (CMAP). Esta herramienta contenía una colección de referencia de perfiles de expresión génica de células humanas cultivadas tratadas con pequeñas moléculas bioactivas

que abrieron un escenario totalmente nuevo para los análisis de descubrimiento y reutilización de fármacos. Posteriormente esta herramienta derivó en CLUE (14), la cual describíamos en la sección de introducción, que posee más de 1300000 firmas genéticas derivadas de agentes perturbadores.

4.2. MÉTODOS BASADOS EN TRANSCRIPTÓMICA

Existen diferentes clasificaciones para los diferentes tipos de análisis de reutilización de fármacos según los criterios que se tomen en cuenta, por ejemplo, en relación al objetivo, podríamos agruparlos en métodos enfocados en fármacos y métodos enfocados en enfermedad (46), dependiendo de si el objetivo principal es encontrar similitudes entre fármacos o similitudes entre fármacos y enfermedades. La clasificación seguida en esta tesis se basa en la **similitud metodológica** o algorítmica (34). En este contexto, hemos agrupado los diferentes algoritmos en 2 grupos principales, los métodos basados en la similitud y los métodos basados en *machine learning* o aprendizaje automático, donde también incluimos técnicas de análisis en redes de interacción.

4.2.1. MÉTODOS BASADOS EN SIMILITUD

Los métodos basados en similitud utilizan la comparación directa entre firmas de expresión genética diferentes, ordenadas en base al valor de expresión diferencial con respecto a controles con el objetivo de medir un nivel de relación o similitud entre ellas. Por ejemplo, entre una firma de una enfermedad y firmas derivadas de fármacos. Este grado de similitud o correlación está basado en el orden en el que se encuentran los genes en ambas listas y nos revela si las firmas tienen patrones genéticos comunes o conjuntos de genes sobre-expresados e infra-expresados compartidos o patrones inversos o grupos de genes sobre-expresados en una firma e infra-expresados en la otra. Entre las técnicas más usadas encontramos las siguientes:

- *Gene Set Enrichment Analysis* (GSEA).

El análisis de enriquecimiento de genes es un método no-paramétrico basado en la estadística de Kolmogorov-Smirnov inicialmente desarrollado para realizar análisis funcionales con el objetivo de medir el nivel de relación entre un conjunto de genes diferencialmente expresados y una biblioteca de conjuntos de genes que representan funciones biológicas específicas (47). Posteriormente, se implementó dentro de CMAP como aplicación en los análisis de reutilización de fármacos siendo a día de hoy uno de los métodos más usados y el que usamos para el segundo trabajo publicado que recoge esta tesis.

Brevemente, la aplicación del método comienza con la generación de listas ordenadas de genes asociados a un fenotipo o condición determinada, en nuestro caso, las firmas genéticas derivadas de fármacos. Los genes dentro de las firmas se ordenan de mayor a menor según una métrica, que suele ser el *fold change* o la diferencia del nivel de expresión entre dos condiciones, muestras tratadas y muestras no tratadas. Estas firmas serán usadas como biblioteca de referencias sobre las que comparar a continuación. De igual modo, necesitamos obtener una firma de estudio, por ejemplo, la firma genética de una enfermedad. En éste caso, la firma no contendrá el total de los genes, sino que es resumida en un conjunto de genes, los genes significativa y diferencialmente expresados o incluso un subgrupo de los mismos. Posteriormente se calcula un valor estadístico de enriquecimiento entre la firma de la enfermedad y cada una de las firmas derivadas por fármacos que es reflejo de las posiciones de los genes de la firma de la enfermedad dentro de la firma de un fármaco. Si todos los genes sobre-expresados en la enfermedad se encuentran arriba, o muy sobre-expresados en la firma del fármaco, el valor de enriquecimiento será cercano a 1 y refleja una similitud o correlación positiva entre ambas firmas. En el caso contrario, si los genes sobre-expresados se encuentran abajo o muy infra-expresados en la firma del fármaco se obtendrá un valor cercano a -1, reflejo de que las firmas son inversas. El cálculo se realiza recorriendo los genes de la firma del fármaco, si en cada paso, el gen está contenido en la firma de la enfermedad, se incrementa el valor de enriquecimiento y en caso contrario se reduce el mismo.

Existen variaciones dentro del método de GSEA, entre las que podemos destacar el PGSEA o análisis paramétrico de enriquecimiento de genes (48), el cual mejora la eficiencia computacional.

- Enriquecimiento basado en co-expresión de genes.

Este tipo de técnicas se basan en extraer conjuntos de genes cuya expresión está altamente correlacionada a través de un conjunto de diferentes muestras de una condición. Estos grupos de genes suelen actuar de forma conjunta y por tanto pueden ser considerados como módulos de regulación. Estos módulos proporcionan una alternativa a las comparaciones basadas en genes que puede aportar ciertas ventajas, al establecer una conexión entre un fenotipo y un módulo de regulación implicado en ella.

Su aplicación dentro del análisis de reutilización de fármacos ha sido usada dentro de algunas herramientas (49) y puede dividirse en 3 pasos principales. Primero deben identificarse los módulos de co-expresión, por ejemplo, mediante análisis de agrupamiento (*clustering*) usando las matrices de expresión genética. El siguiente paso es realizar un análisis funcional sobre esos módulos para anotarlos con las rutas biológicas las cuales regulan y, por último, se realiza un segundo análisis de enriquecimiento esta vez entre los módulos significativa y diferencialmente expresados y las firmas derivadas de fármacos. Este análisis proporciona una serie de fármacos los cuales están relacionados con una serie de rutas funcionales las cuales a su vez se encuentran alteradas en la condición de estudio o enfermedad con respecto a sus controles.

4.2.2. MÉTODOS BASADOS EN *MACHINE LEARNING*

Los algoritmos basados en *machine learning* son técnicas basadas en la construcción de un conjunto de reglas de clasificación para distinguir asociaciones o relaciones entre dos condiciones (50). Estas técnicas son una poderosa forma de predecir nuevas asociaciones e indicaciones de fármacos, aunque necesitan conocimiento y relaciones previamente establecidas.

- Análisis basados en conexiones en redes.

Debido al carácter multifactorial de las enfermedades complejas es lógico hipotetizar que los análisis de redes de interacción que integran múltiples capas de información pueden obtener

resultados más cercanos a la realidad que los análisis dirigidos a objetivos genéticos individuales. Un análisis de interacciones puede ser entendido como un gráfico en el que tenemos una serie de nodos y conexiones entre ellos formando una compleja red. Cada nodo representa una unidad dentro del estudio, por ejemplo, un gen, una proteína o un fármaco, mientras que cada eje representa una conexión o relación entre dos nodos y su magnitud. Hay muchos enfoques diferentes dentro de esta categoría de técnicas con las que abordar un análisis de reutilización de fármacos (51) pero todos están basados en la construcción de diferentes capas de interacción, por ejemplo, de interacciones en expresión genética fármaco-fármaco, enfermedad-fármaco, fármaco-enfermedad-fenotipos o incluso capas con otro tipo de información como toxicidad o efectos secundarios. Las diferentes capas son conectadas posteriormente para descubrir nuevas conexiones entre capas no conectadas previamente. Vemos entonces que la formación de las redes de conexión de cada capa puede contener diferente nivel y magnitud de información y puede estar sujeta tanto a información local previamente descrita, como por ejemplo interacciones DNA-proteína, proteína-proteína o fármaco-proteína, como a información más laxa y amplia derivada de datos ómicos, como las conexiones entre firmas genéticas o rutas biológicas y los fármacos o las enfermedades (52).

Como ejemplos de integración de redes encontramos algoritmos ampliamente usados como *random walk* (53,54) donde se miden la similitud o relación entre dos o más nodos teniendo en cuenta la posición de cada uno, la longitud y dirección de los ejes que los conectan, pasando o no por nodos intermedios o vecinos, y se hace recorriendo la red de procedural y aleatoriamente. Entre otros algoritmos encontramos DTINet, para integrar de una manera fácilmente escalable a diferentes capas de información (55) o los métodos basados en identificación de núcleos de información o *kernel-based methods*. Estos últimos se basan en la estructura propia de la red, que suele formar núcleos de relaciones, por ejemplo, si generamos una red genes-rutas biológicas, todos los genes pertenecientes a una ruta van a situarse cerca dentro de la red formando un agrupamiento. Una vez generada la red, al incluir nueva información, por ejemplo, los genes pertenecientes a la firma de un fármaco, estos genes se distribuirán dentro de los diferentes núcleos, lo que permite asignar un peso a cada uno y medir así la importancia o la magnitud de relación entre el fármaco y los diferentes núcleos de la red (56).

- Modelos de factorización de matrices.

La factorización de matrices es un método ampliamente usado en diferentes contextos para reducir la dimensionalidad en los datos y extraer las características importantes o patrones más explicativos en los datos. Recientemente está siendo utilizado en análisis de reutilización de fármacos basados en transcriptómica, como en el estudio de Yang et al. (57), donde generaron una matriz que resumía de manera probabilística una serie de patrones genéticos compartidos entre fármacos y enfermedades, o el estudio de Dai et al. (58), en el que aplicaron esta técnica para integrar en un mismo espacio genes, fármacos y enfermedades con el objetivo de descubrir nuevas indicaciones para los fármacos.

- Clasificadores supervisados

Estos enfoques son usados comúnmente para clasificar un elemento dentro de una clase conocida, por ejemplo, clasificar una muestra en el grupo de pacientes o en el grupo de controles en base a la expresión de ciertos genes, pero de igual manera pueden usarse en reutilización de fármacos como veremos a continuación. Estos métodos están basados en la generación primeramente de un **modelo de clasificación**, para lo cual se usa un conjunto de datos de entrenamiento donde se tienen definidas todas las interacciones y todas las clases posibles. Un conjunto de entrenamiento puede ser una matriz con la expresión de fármacos los cuales se conoce sobre qué enfermedad se usan, siendo las enfermedades, las clases. De este modo se construye el modelo de interacciones conocidas sobre el cual posteriormente se analizarán las nuevas muestras sobre las cuales se quiere obtener un resultado nuevo.

Entre los métodos más usados encontramos el método local de bipartito (BLM), que es un método basado en *kernel* o en la búsqueda y pesado de agrupamientos dentro de una red de interacciones. La información de la nueva muestra, por ejemplo, los genes significativos y diferencialmente expresados, son introducidos en la red modelo, y cada módulo, que puede representar el uso para una determinada enfermedad, es pesado de acuerdo al número y la conexión con los genes de la muestra de estudio. Los módulos con mayor peso o importancia nos revelarán cuales son los posibles mejores usos para el fármaco de entre todas las enfermedades introducidas en el modelo.

La figura 3 muestra un resumen de los métodos y aplicaciones de los análisis de reutilización de fármacos basados en transcriptómica, así como de los repositorios públicos más usados (figura adaptada de *In Silico Drug Design: Repurposing Techniques and Methodologies* (59)).

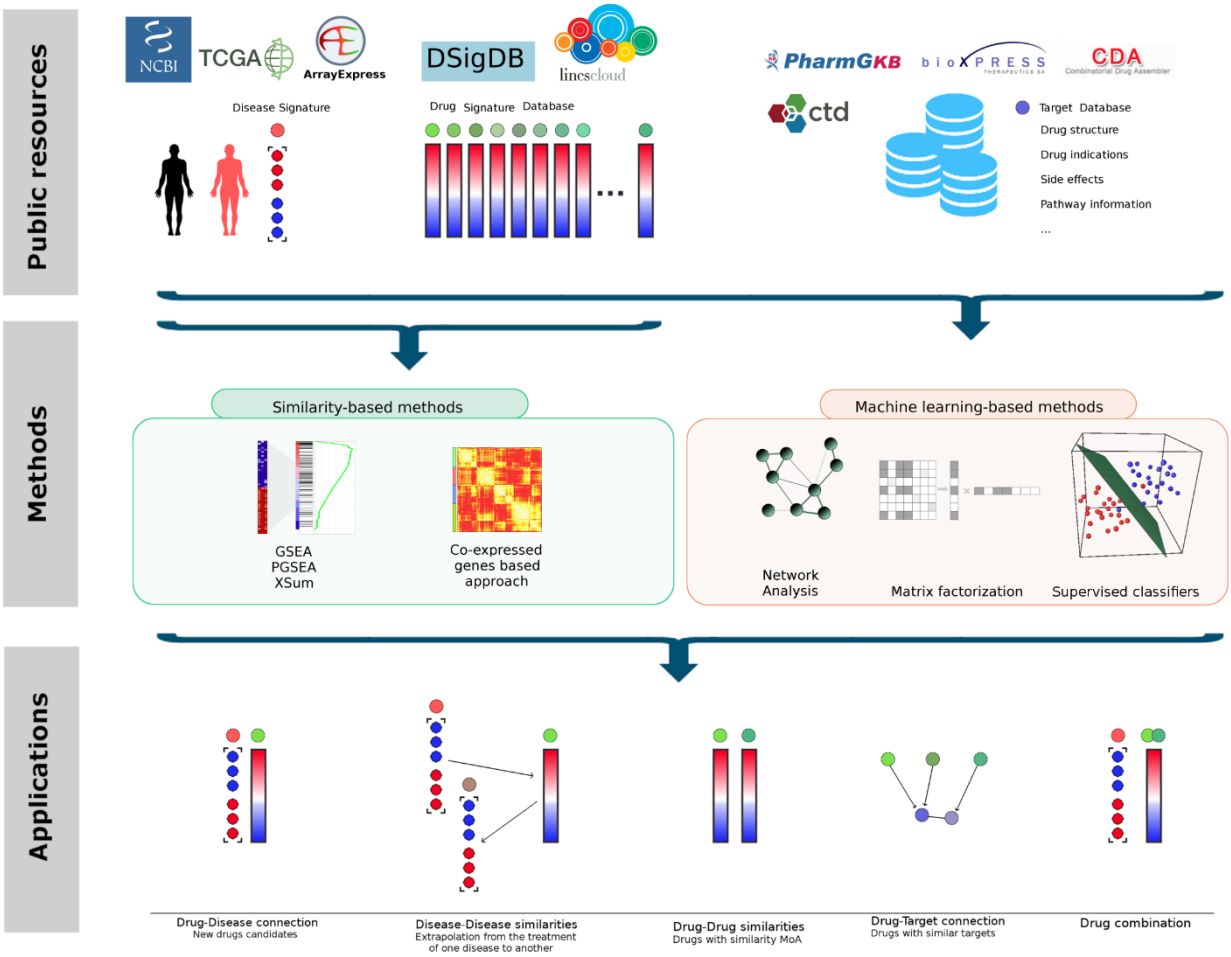


Figura 3: Resumen de recursos públicos, métodos y aplicaciones de los análisis de reutilización de fármacos.

4.3. SEGUNDO ARTÍCULO

Título: *Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis.*

Dirección web: https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-017-1263-7

El LES es una enfermedad autoinmune con pocas opciones de tratamiento. Las terapias actuales no son completamente efectivas y muestran respuestas altamente variables además de severos efectos secundarios. En este sentido, los grandes esfuerzos de la comunidad investigadora se han centrado en el desarrollo de nuevas estrategias terapéuticas más efectivas y seguras. En este trabajo, realizamos un análisis de reutilización de fármacos para el LES con el objetivo de proporcionar nuevas alternativas de estudio de posibles terapias más eficientes. El análisis de reutilización de fármacos basado en la comparación de firmas de expresión génica es una técnica exploratoria eficaz para la identificación de nuevos enfoques terapéuticos que permite obtener fármacos candidatos para tratar una enfermedad basándose en los patrones genéticos desregulados en la misma.

Para ello, recopilamos un compendio de estudios de expresión o conjuntos de datos de diferentes estudios de LES realizados en sangre y en diferentes tipos celulares sanguíneos. Una vez extraídas las firmas genéticas de cada estudio, usamos la base de datos Lincscloud (ahora llamada CLUE), que contiene más de 20000 firmas genéticas inducidas por fármacos, para obtener las medidas de similitud entre cada firma de cada estudio y cada fármaco. Con el hecho de seleccionar algunos estudios realizados en sangre y otros en líneas celulares concretas buscábamos crear una cohorte inicial de partida lo más heterogénea posible, debido a que posteriormente buscamos los patrones consistentes entre todos los casos y los fármacos que puedan revertir el fenotipo patogénico a través de toda esa heterogeneidad. Una vez obtenida una lista ordenada de fármacos en base a la similitud para cada firma genética de LES, se realizó un meta-análisis por producto de rangos y seleccionamos los fármacos con medidas de similitud muy positivas y muy negativas a través de todos los estudios.

Los resultados proporcionaron una lista de fármacos inductores de los patrones genéticos conservados en LES (similaridad positiva entre firmas fármaco-LES) así como una lista con posibles fármacos candidatos para revertir la enfermedad (similitud negativa). Sobre estos resultados, profundizamos posteriormente en los mecanismos de acción y las rutas biológicas sobre las que actúan cada fármaco, obteniendo como los candidatos más significativos aquellos fármacos inhibidores de la ruta de la PI3K y el mTOR.

Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis

Daniel Toro-Domínguez,[1,2] Pedro Carmona-Sáez*[1,2] and Marta E. Alarcón-Riquelme *[1,3].

[1]Area of Medical Genomics, Pfizer–University of Granada–Andalusian Government Centre for Genomics and Oncological Research (GENYO), Avda. de la Ilustración 114, PTS-18016 Granada, Spain

[2]Bioinformatics Unit, Pfizer–University of Granada–Andalusian Government Centre of Genomics and Oncological Research (GENYO), Avda. de la Ilustración 114, PTS-18016 Granada, Spain

[3]Unit of Chronic Inflammatory Diseases, Institute of Environmental Medicine, Karolinska Institute, Stockholm, 17177 Sweden

Daniel Toro-Domínguez, Email: se.oyneg@orot.leinad.

Contributor Information.

*Corresponding author.

Abstract

Background: Systemic lupus erythematosus (SLE) is an autoimmune disease with few treatment options. Current therapies are not fully effective and show highly variable responses. In this regard, large efforts have focused on developing more effective therapeutic strategies. Drug repurposing based on the comparison of gene expression signatures is an effective technique for the identification of new therapeutic approaches. Here we present a drug-repurposing exploratory analysis using gene expression signatures from SLE patients to discover potential new drug candidates and target genes.

Methods: We collected a compendium of gene expression signatures comprising peripheral blood cells and different separate blood cell types from SLE patients. The Lincscloud database was mined to link SLE signatures with drugs, gene knock-down, and knock-in expression signatures. The derived dataset was analyzed in order to identify compounds, genes, and pathways that were significantly correlated with SLE gene expression signatures.

Results: We obtained a list of drugs that showed an inverse correlation with SLE gene expression signatures as well as a set of potential target genes and their associated biological

pathways. The list includes drugs never or little studied in the context of SLE treatment, as well as recently studied compounds.

Conclusion: Our exploratory analysis provides evidence that phosphoinositol 3 kinase and mammalian target of rapamycin (mTOR) inhibitors could be potential therapeutic options in SLE worth further future testing.

Keywords: Autoimmunity; Drug discovery; Drug repurposing; Gene expression; Lincscloud; Systemic lupus erythematosus

Background

Systemic lupus erythematosus (SLE) is an autoimmune disorder in which the immune system produces autoantibodies against its own cells and tissues leading to chronic inflammation and organ damage. Although some biological pathways are well known to be altered in lupus, such as the type I interferon (IFN) pathway [1], the biological mechanisms behind disease development are poorly understood in general and it has been proposed that genetic and environmental factors are involved [2]. There are many classes of drugs commonly used for SLE treatment, such as corticosteroids, immunosuppressants, nonsteroidal anti-inflammatory drugs, or specific monoclonal antibodies directed against cell surface receptors or cytokines [3]. Nevertheless, the multifactorial nature and the undefined etiology of this disease contribute to the absence of efficient treatments [4].

In the last decade, the widespread use of high-throughput technologies such as gene expression microarrays has enabled access to large collections of gene expression databases that can be exploited for a wide range of applications. In this context, in-silico drug-repurposing analysis based on gene expression data allows us to identify new therapeutic applications for drugs used in other contexts. This technique compares the disease gene expression signature against a large collection of profiles derived from different compounds, measuring the degree of similarity among them. A positive similarity score means that the compound produces a similar gene expression pattern to that of the disease. In the same way, a negative similarity score represents the opposite; that is, the overexpressed genes in the disease appear underexpressed in the drug signature and vice versa. This evidences that the effect of the drug on transcription is opposite to the effect of the disease, and it is reasonable to hypothesize that the drug might be able to reverse the disease gene expression program and the phenotype itself [5].

The Connectivity Map [6] was a pioneer tool that implemented this approach. Since its publication, many studies have proven the potential of this type of analysis to discover new

treatments for different diseases such as several types of cancer, muscle atrophy, or inflammatory bowel disease, among others [7].

In this context, Lincscloud [8] has been deployed recently as the successor to the Connectivity Map. This database contains genetic profiles derived from a larger number of drugs and also includes knock-down and knock-in gene experiments, where whole gene expression profiles are measured after inhibiting or overexpressing a single gene. During the last few years there has been an increasing interest in the application of this approach for drug repurposing or target predictions. For example, Johannessen et al. [9] explored the transcriptional connections between cAMP signaling and GPCR pathway-associated drug resistance candidates. Santagata et al. [10] revealed a strong connection between the HSF1 gene and compounds that inhibit protein translation, while Siavelis et al. [11] proposed new treatments for Alzheimer's disease.

In this work we performed a drug-repurposing analysis using a collection of gene expression signatures derived from previously published studies of SLE patients and gene expression signatures derived from Lincscloud. This analysis allowed us to establish a set of drug candidates that reverse the SLE signatures and a set of genetic targets, as well as new pharmacological paths in SLE.

Methods

Processing gene expression data

We mined the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database [12] to retrieve gene expression datasets from SLE patients. We selected experiments performed in any blood tissue, with case and healthy samples, without any treatment applied in the case of in-vitro samples, and each experiment with more than four replicates. To purposely obtain a heterogeneous dataset we searched for gene expression data from adult and juvenile SLE performed in different microarray platforms. By doing this we considered the patterns conserved across all SLE cases removing differences between SLE clinical types or microarray platform-dependent biases.

Each gene expression dataset was downloaded and processed independently using the R statistical environment. Genes with a high percentage of missing values (more than 15% across samples) were filtered out and remaining missing values were imputed using the average expression values within each group (case or control) of each dataset. We annotated probes to gene symbol identifiers, data were transformed to a logarithm scale, and the median expression value was computed for probes corresponding to the same gene. Differential expression analysis was performed between controls and cases for each dataset using the limma R package. Next we discarded genes presenting $p > 0.05$, and the top 500 most overexpressed and

underexpressed genes were selected as the SLE genetic signature from each dataset to be used for further analysis.

Drug-repurposing analysis

For each independent SLE signature we performed a query on the Lincscloud database and retrieved the list of drugs and knock-in and knock-down genes with high similarity scores. We used as a similarity score the "best score 4" value, which is the proposed threshold in Lincscloud and is calculated as the mean connectivity score across the four cell lines in which the drug or perturbagen connected most strongly to the query.

To integrate the results from each independent SLE signature, a unique dataset was created where rows represent drugs and columns represent SLE signatures, and each entry of the matrix is the similarity score (best score 4 values) between drugs and SLE signatures. For each drug (row) we calculated the median similarity score across all SLE signatures. To evaluate whether equal or better scores could be obtained by chance, an empirical p value was calculated generating 10,000 random datasets permuting rows and columns in the original set of data. We then computed the p value as the fraction of permutations having a similarity score equal to or higher than (in absolute value) the observed score. Significant drugs were then selected if they presented $p < 0.05$ and showed a median similarity score > 80. The same procedure was applied to knock-in and knock-down gene expression signatures (see Fig. 1). The results obtained are therefore independent of the cell lines' inherent gene expression patterns but are consistent with the patterns that are common to all of the SLE signatures.

Drug-target enrichment analysis

To evaluate whether some drug targets were significantly enriched in the list of obtained drugs we downloaded drug-target information from DrugBank [13], ChEBI [14], and Therapeutic Target Database [15]. Data files from these three databases were parsed and an annotation file was created with information for 131,162 drugs (including synonymous names) and their biological targets. With this information, we associated target genes to the list of drugs in Lincscloud and our list of significant drugs. For drugs without target information in these resources we carefully revised the information available from compound manufacturer catalogs and the associated literature. Drugs without any information in the literature or in databases were discarded from the drug-target analysis.

Fisher's exact test was applied to evaluate what target genes were statistically overrepresented in the list of significant drugs with respect to the total set of annotated drugs.

Results

Analysis of gene expression signatures

After careful exploration we found 10 datasets of SLE in the NCBI GEO, two of which contained samples from juvenile SLE patients. Some of the datasets contained samples from different tissues, which we treated as independent datasets in our analysis. Thus, we identified 14 different tissue-specific datasets that passed the initial filters (see Additional file 1: Sheets 1 and 3). These datasets comprised a total of 327 SLE samples and 173 healthy controls. Each dataset was subjected to quality control and processed as described in Methods, generating 14 individual signatures including different blood tissues (see Additional file 1: Sheet 2).

Connections between SLE and drug gene expression signatures

Our analysis yielded 61 drugs that were significantly associated with the SLE signatures, 40 with similar gene expression patterns and 21 with opposite patterns (see Fig. 2 and Additional file 1: Sheet 4). Some of these compounds have been associated previously with SLE but some others have not been described in this context and hence could be new potential drug candidates (see Discussion). We used the information from DrugBank, ChEBI, and Therapeutic Target Database to annotate target genes for each drug and classify these compounds into groups with the same target.

The analysis of targets common across the list of drugs yielded three sets with similar gene expression signatures that showed significant p values, including topoisomerase II inhibitors, histone deacetylase (HDAC) inhibitors, and PKC activators, as well as three groups with negative scores, where we found phosphoinositol 3 kinase (PI3K) inhibitors, cyclin-dependent kinase (CDK) inhibitors, and mammalian target of rapamycin (mTOR) inhibitors (see Table 1). Five different compounds were PI3K inhibitors, providing the most significant p value in the enrichment analysis.

To further explore this result, we used information from the KEGG database [16] to construct a network of the PI3K signaling pathway (see Additional file 2). Interestingly, we found that most of the other drug targets, such as IGF, Rho, mTOR, or CDK, were also playing important roles in the PI3K signaling pathway. PI3K regulates important processes such as cell survival, immune proliferation, anti-apoptotic pathways of immune cells, and immune response linked to interferon signaling and cytokine signaling pathways [17], all important and impaired in SLE. We also obtained dual inhibitors of PI3K and mTOR such as NVP-BEZ235 [18]. Other drug targets of the PI3K signaling pathway have been related with SLE or other SLE-like disorders, such as CDK inhibitors, recently proposed to be used for treatment of some autoimmune disorders [19], or inhibitors of the mitogen-activated protein kinase (MAPK) signaling pathway [20].

Study of gene effect-caused profiles

We obtained seven knock-in and 90 knock-down genes with a positive similarity score that produce an SLE-like profile, and 50 knock-down genes with a negative similarity score (see Table 2 and Additional file 1: Sheet 4) that reverse the SLE profile (genes up-regulated in the disease signature are down-regulated in the drug signature, and vice versa). Many genes have been already described in SLE, such as CD40 [21], interferon-related genes, and translation initiation factors, such as EIEF4 [22, 23]. Additional functional analyses with these genes are described in Additional file 3. Interestingly and in agreement with our previous analysis, we found that the gene expression signature associated with knock-down genes such as PI3K or IGF1R show a negative similarity score. That is, the inhibition of these genes could reverse the gene expression profile induced by SLE. This is consistent with the fact that gene expression profiles of drugs which inhibit these genes showed a negative score with respect to the SLE signatures.

Discussion

In this study we performed a systematic screening for drugs or genes that induced similar or opposite gene expression programs to signatures from SLE patients. We integrated signatures from different blood cell populations and SLE subtypes in order to identify consistent and conserved profiles, reducing considerably the false positive ratio. In this analysis, we found 40 drugs (see Additional file 1: Sheet 4) with a positive similarity score, which induces changes similar to the SLE phenotype. In this set of compounds, HDAC, topoisomerase II, and PKC were the more significant targets. Many of these drug targets are key factors in biological processes that are altered in SLE. For example, HDAC inhibitors have been related to impairment of immune processes described in lupus, such as autophagy [24], although there is contradictory information about the role of some isoforms of HDAC in the immune system [25, 26]. A recent study shows that HDAC inhibitors may be suitable for treatment of autoimmunity, while primary responses to the same inhibitors were greatly impaired, probably explaining the contradiction between the positive similarity score we obtained and the potential use of HDAC inhibitors in SLE [27]. In addition, Lohman et al. [28] showed that HDAC inhibitors have anti-inflammatory activity which is inversely correlated with dose, amplifying the production of inflammatory mediators at concentration > 3 μM. In another context, treatment of human cells with topoisomerase II inhibitors such as etoposide has been shown to induce interferon-stimulated genes [29].

Other positively correlated compounds are phorbol-12-myristate-13-acetate and ingenol, the former of which has been used to stimulate the immune response and the interferon signaling pathway [30]. These drugs are protein kinase C (PKC) activators, a protein with some isoforms

associated with SLE. In this context, the use of PKC inhibitors has been proposed as treatment for autoimmune disorders [31, 32] due to their induced increase in proliferation of regulatory T cells (Tregs). In addition, deficient MEK/ERK signaling pathway is related to SLE and cytokine generation [33] through impaired PKC activation. This pathway has also been proposed as a potential therapeutic target for rheumatoid arthritis [34]. Another compound with a high positive similarity score was LE-135, which is a retinoic acid receptor inhibitor. The use of retinoic acid has been also related to an improvement in SLE recovering the Treg balance [35, 36].

Attending to drugs with negative similarity scores, we identified 21 compounds that induce opposite gene expression programs with respect to SLE signatures (see Additional file 1: Sheet 4). Almost all of them act in the same processes, down-regulating the immune response and the proliferation of immune cells. PI3K was the most significant in the target enrichment analysis, due to a set of PI3K inhibitors. PI3K inhibitors have been reported to ameliorate the effects of SLE and other autoimmune disorders in animal models [37–39]. In addition, mTOR was also found as a significantly enriched target associated with mTOR inhibitors such as NVP-BEZ235, AZD8055, TGX115, or Ku0063794.

Recent experimental evidence suggests that mTOR inhibitors may provide a new therapeutic strategy for the treatment of SLE patients [40]. Indeed, PI3K and mTOR act in the same signaling cascade [38] promoting the interferon and cytokine signaling pathways [41].

Complementarily, the analysis of gene-caused profiles defined a set of genes – both described and not previously described in SLE – that could play an important role in the development of the disease. Some of these were interferon-related genes, transcriptional and translational factors, and a set of biological pathways related to these genes including the PI3K and the insulin signaling pathways, immune response, or transcriptional and translational processes (Additional file 2). These results are highly consistent with the analyzed list of drugs and also support that the inhibition of PI3K signaling could improve the SLE phenotype. The evidence presented here should lead not only to testing of PI3K inhibitors as potential SLE treatment, but also to actively testing any other compound obtained, such as the insulin growth factor receptor inhibitors that crosstalk with the PI3K and mTOR pathways or the Rho kinase inhibitors.

Although the Lincscloud database contains mostly experiments carried out in cancer cell lines, the integration of different SLE signatures and the inclusion of summarized drug signatures from different cell populations enable one to establish global associations based on ubiquitous expression across different cell lines. In-silico analyses are often exploratory studies and should be confirmed by in-vitro or in-vivo experiments. In this sense, previous experiments already provide evidence that PI3K inhibitors ameliorate the SLE phenotype in animal models [37–39], and that of other autoimmune disorders, although these drugs are not used clinically. Our results would therefore provide further support for the inhibition of the PI3K signaling pathway to treat SLE.

Conclusions

We performed an integrative in-silico drug-repurposing exploratory analysis based on comparing gene expression data of SLE against gene expression profiles produced by perturbagens from the Lincscloud database. Our analysis is designed to reduce the biases of using different microarray platforms and the heterogeneity of SLE, leading to discovery of conserved genetic patterns across different disease states or cell types. We identified a set of pathways related to biological processes impaired in SLE, compounds, and drug targets with potential therapeutic interest for SLE treatment. Based on the results, we highlighted PI3K and mTOR as good candidates and PI3K signaling pathway inhibitors as potential treatment options that are interesting enough to be further explored, although we described other targets that could also be further evaluated to test their effect in improving the phenotype of SLE, such as PKC, MAPK, or other specific kinases. This type of analysis has seldom been performed for autoimmune diseases and can provide novel therapeutic approaches for heterogeneous and multifactorial disorders, such as SLE.

References

1. Crow MK. Type I, interferon in the pathogenesis of lupus. J Immunol. 2014;192:5459–5468. doi: 10.4049/jimmunol.1002795.

2. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015;47:1457–1464. doi: 10.1038/ng.3434.

3. Chambers SA, Rahman A, Isenberg DA. Treatment adherence and clinical outcome in systemic lupus erythematosus. Rheumatology. 2007;46:895–898. doi: 10.1093/rheumatology/kem016.

4. Muangchan C, van Vollenhoven RF, Bernatsky SR, Smith CD, Hudson M, Inanç M, et al. Treatment algorithms in systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2015;67:1237–1245. doi: 10.1002/acr.22589.

5. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? Drug Discov Today. 2013;18:350–357. doi: 10.1016/j.drudis.2012.07.014.

6. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–1935. doi: 10.1126/science.1132939.

7. Vilar S, Hripcsak G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. Brief Bioinform. 2016; bbw048.

8. LINCS Consortium. Lincscloud. http://lincsproject.org/.

9. Johannessen CM, Johnson LA, Piccioni F, Townes A, Frederick DT, Donahue MK, et al. A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. Nature. 2013;504:138–142. doi: 10.1038/nature12688.

10. Santagata S, Mendillo ML, Tang Y, Subramanian A, Perley CC, Roche SP, et al. Tight coordination of protein translation and HSF1 activation supports the anabolic malignant state. Science. 2013;341:1238303. doi: 10.1126/science.1238303.

11. Siavelis JC, Bourdakou MM, Athanasiadis EI, Spyrou GM, Nikita KS. Bioinformatics methods in drug repurposing for Alzheimer's disease. Brief Bioinform. 2015;17:322–35.

12. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41(Database issue):D991–D995. doi: 10.1093/nar/gks1193.

13. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–D672. doi: 10.1093/nar/gkj067.

14. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucl Acids Res. 2013;41:D456–D463. doi: 10.1093/nar/gks1146.

15. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 2012;40(Database issue):D1128–D1136. doi: 10.1093/nar/gkr797.

16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30. doi: 10.1093/nar/28.1.27.

17. Guiducci C, Ghirelli C, Marloie-Provost M-A, Matray T, Coffman RL, Liu Y-J, 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;205:315–322. doi: 10.1084/jem.20070763.

18. Mukherjee B, Tomimatsu N, Amancherla K, Camacho CV, Pichamoorthy N, Burma S. The dual PI3K/mTOR inhibitor NVP-BEZ235 is a potent inhibitor of ATM- and DNA-PKCs-mediated DNA damage responses. Neoplasia. 2012;14:34–43. doi: 10.1593/neo.111512

19. Xia Y, Lin L-Y, Liu M-L, Wang Z, Hong H-H, Guo X-G, et al. Selective inhibition of CDK7 ameliorates experimental arthritis in mice. Clin Exp Med. 2014;15:269–275. doi: 10.1007/s10238-014-0305-6.

20. Qu H, Bian W, Xu Y. A novel NF-κB inhibitor, DHMEQ, ameliorates pristane-induced lupus in mice. Exp Ther Med. 2014;8:100–104

21. Desai-Mehta A, Lu L, Ramsey-Goldman R, Datta SK. Hyperexpression of CD40 ligand by B and T cells in human lupus and its role in pathogenic autoantibody production. J Clin Invest. 1996;97:2063–2073. doi: 10.1172/JCI118643.

22. Wu Y-Y, Kumar R, Haque MS, Castillejo-López C, Alarcón-Riquelme ME. BANK1 controls CpG-induced IL-6 secretion via a p38 and MNK1/2/eIF4E translation initiation pathway. J Immunol. 2013;191:6110–6116. doi: 10.4049/jimmunol.1301203.

23. 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;16:489. doi: 10.1186/s13075-014-0489-x.

24. Shao Y, Gao Z, Marks PA, Jiang X. Apoptotic and autophagic cell death induced by histone deacetylase inhibitors. PNAS. 2004;101:18030–18035. doi: 10.1073/pnas.0408345102.

25. Choi SW, Gatza E, Hou G, Sun Y, Whitfield J, Song Y, et al. Histone deacetylase inhibition regulates inflammation and enhances Tregs after allogeneic hematopoietic cell transplantation in humans. Blood. 2015;125:815–819. doi: 10.1182/blood-2014-10-605238.

26. Wang L, Liu Y, Han R, Beier UH, Bhatti TR, Akimova T, et al. FOXP3+ regulatory T cell development and function require histone/protein deacetylase 3. J Clin Invest. 2015;125:1111–1123. doi: 10.1172/JCI77088.

27. Waibel M, Christiansen AJ, Hibbs ML, Shortt J, Jones SA, Simpson I, et al. Manipulation of B-cell responses with histone deacetylase inhibitors. Nat Commun. 2015;6:6838. doi: 10.1038/ncomms7838.

28. Lohman R-J, Iyer A, Fairlie TJ, Cotterell A, Gupta P, Reid RC, et al. Differential anti-inflammatory activity of HDAC inhibitors in human macrophages and rat arthritis. J Pharmacol Exp Ther. 2016;356:387–396. doi: 10.1124/jpet.115.229328.

29. Brzostek-Racine S, Gordon C, Van Scoy S, Reich NC. The DNA damage response induces interferon. J Immunol. 2011;187:5336–5345. doi: 10.4049/jimmunol.1100040.

30. Lund ME, To J, O'Brien BA, Donnelly S. The choice of phorbol 12-myristate 13-acetate differentiation protocol influences the response of THP-1 macrophages to a pro-inflammatory stimulus. J Immunol Methods. 2016;430:64–70. doi: 10.1016/j.jim.2016.01.012.

31. Oleksyn D, Pulvino M, Zhao J, Misra R, Vosoughi A, Jenks S, et al. Protein kinase Cβ is required for lupus development in Sle mice. Arthritis Rheum. 2013;65:1022–1031. doi: 10.1002/art.37825.

32. Gray RD, Lucas CD, MacKellar A, Li F, Hiersemenzel K, Haslett C, et al. Activation of conventional protein kinase C (PKC) is critical in the generation of human neutrophil extracellular traps. J Inflamm (Lond) 2013;10:12. doi: 10.1186/1476-9255-10-12.

33. Gorelik G, Fang JY, Wu A, Sawalha AH, Richardson B. Impaired T cell protein kinase Cδ activation decreases ERK pathway signaling in idiopathic and hydralazine-induced lupus. J Immunol. 2007;179:5553–5563. doi: 10.4049/jimmunol.179.8.5553.

34. Thiel MJ, Schaefer CJ, Lesch ME, Mobley JL, Dudley DT, Tecle H, et al. Central role of the MEK/ERK MAP kinase pathway in a mouse model of rheumatoid arthritis: potential proinflammatory mechanisms. Arthritis Rheum. 2007;56:3347–3357. doi: 10.1002/art.22869.

35. Kinoshita K, Yoo B-S, Nozaki Y, Sugiyama M, Ikoma S, Ohno M, et al. Retinoic acid reduces autoimmune renal injury and increases survival in NZB/W F1 mice. J Immunol. 2003;170:5793–5798. doi: 10.4049/jimmunol.170.11.5793.

36. Xiao S, Jin H, Korn T, Liu SM, Oukka M, Lim B, et al. Retinoic acid increases Foxp3+ regulatory T cells and inhibits development of Th17 cells by enhancing TGF-β-driven Smad3 signaling and inhibiting IL-6 and IL-23 receptor expression. J Immunol. 2008;181:2277–2284. doi: 10.4049/jimmunol.181.4.2277.

37. Winkler DG, Faia KL, DiNitto JP, Ali JA, White KF, Brophy EE, et al. PI3K-δ and PI3K-γ Inhibition by IPI-145 Abrogates immune responses and suppresses activity in autoimmune and inflammatory disease models. Chem Biol. 2013;20:1364–1374. doi: 10.1016/j.chembiol.2013.09.017.

38. Weichhart T, Säemann MD. The PI3K/Akt/mTOR pathway in innate immune cells: emerging therapeutic applications. Ann Rheum Dis. 2008;67(Suppl 3):iii70–iii74. doi: 10.1136/ard.2008.098459.

39. Yin CC, Trigunaite A, Dinh D, Yiu Y, Lannutti BJ, Stein PL. Targeting the phosphatidylinositol 3-kinase signaling pathway to inhibit pathogenic B cells in systemic lupus erythematosus (SLE) J Immunol. 2016;196(1 Supplement):210.

40. Gu Z, Tan W, Ji J, Feng G, Meng Y, Da Z, et al. Rapamycin reverses the senescent phenotype and improves immuno-regulation of mesenchymal stem cells from MRL/lpr mice

and systemic lupus erythematosus patients through inhibition of the mTOR signaling pathway. Aging (Albany NY) 2016;8:1102–1114. doi: 10.18632/aging.100925.

41. Perl A, Fernandez DR, Telarico T, Doherty E, Francis L, Phillips PE. T-cell and B-cell signaling biomarkers and treatment targets in lupus. Curr Opin Rheumatol. 2009;21:454–464. doi: 10.1097/BOR.0b013e32832e977c.

Acknowledgements

The authors thank the Swedish Association Against Rheumatism (Reumatikerförbundet) and the Gustaf den Ve: 80th-year Foundation of Sweden for support.

Funding

Funding for the work presented received support from the Innovative Medicines Initiative Joint Undertaking (IMI-JU) under grant agreement n°115565, resources for which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution.

Authors' contributions

PC-S and MEA-R conceived and supervised the project. PC-S and DT-D designed the study and the data analysis pipeline. DT-D implemented the drug-repurposing analysis workflow and performed all of the analysis. PC-S, MEA-R, and DT-D contributed to the interpretation of the data. All authors wrote and revised the manuscript critically. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Abbreviations

CDK: Cyclin-dependent kinase, *HDAC*: Histone deacetylase, *IFN*: Interferon, *MAPK*: Mitogen-activated protein kinase, *mTOR*: Mammalian target of Rapamycin, *NCBI GEO*: National Centre

for Biotechnology Information Gene Expression Omnibus database, *PI3K:* Phosphoinositol 3 kinase, *PKC:* Protein kinase C, *SLE:* Systemic lupus erythematosus, *Treg*: Regulatory T cell.

Figures

Figure 1: Integrative drug-repurposing analysis. Fourteen signatures of SLE were obtained from 14 different datasets. Each signature was queried on the Lincscloud database and a set of drugs and knock-down and knock-in genes was obtained with similarity scores. The median similarity score and empirical p values were calculated to select significant results across all datasets. Bottom: summary interpretation of the positively and negatively correlated results. NCBI GEO National Center for Biotechnology Information Gene Expression Omnibus, SLE systemic lupus erythematosus.

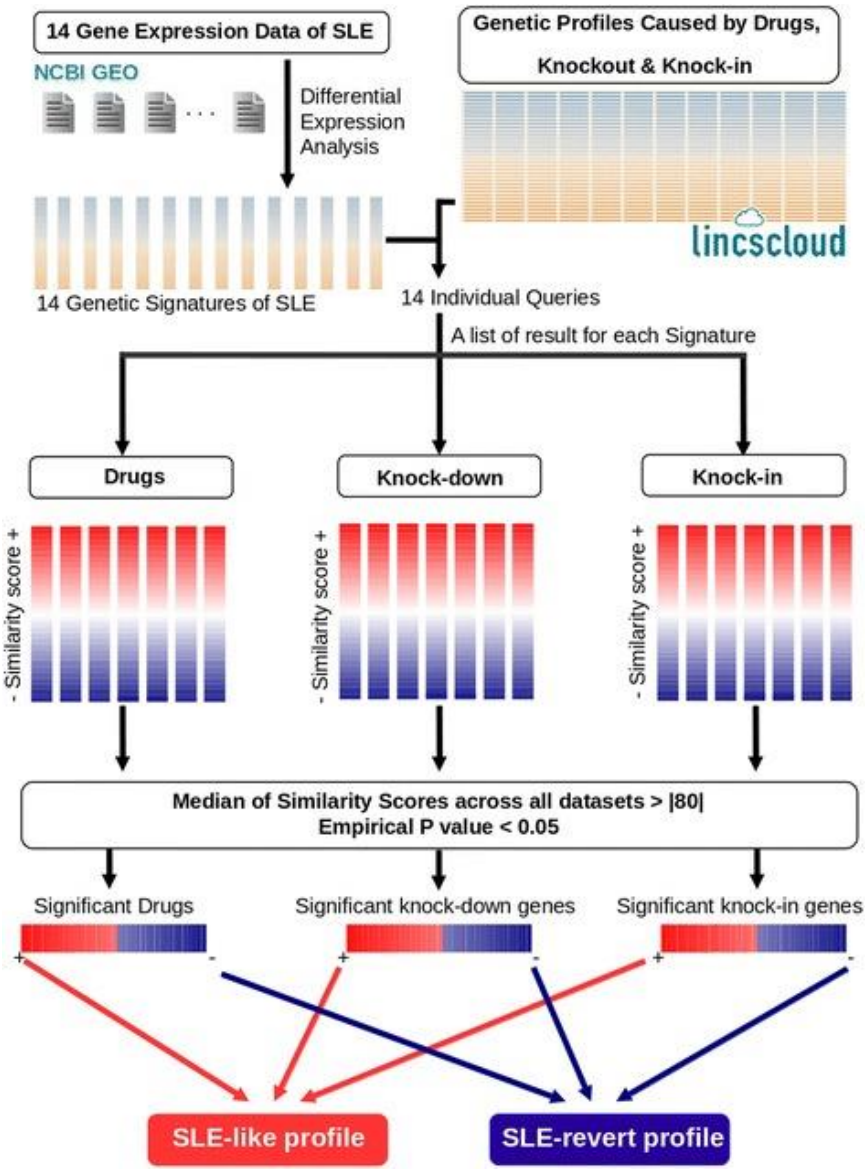
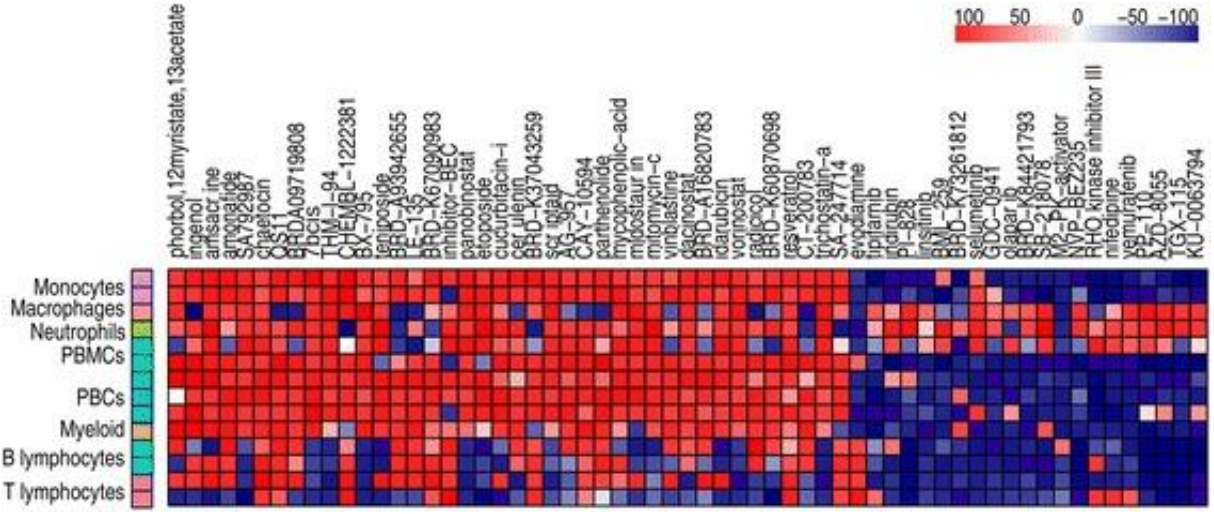


Figure 2: Heatmap of significant drugs representing similarity scores for each drug in the results of each dataset. *Rows*: results of the different datasets used for the analysis. Datasets classified according to the blood cell type (see key). *Columns*: different drugs sorted decreasingly by the median of similarity scores, from left to right.



Tables

Table 1: Drugs obtained and their biological targets. Table presents significant drugs with their biological target and their mechanism of action. p value calculated for groups of drugs with the same target using Fisher's exact test.

CDK cyclin-dependent kinase, HDAC histone deacetylase, mTOR mammalian target of rapamycin, PI3K phosphoinositol 3 kinase.

a + drugs with positive similarity score, – drugs with negative similarity score in regard to SLE signatures.

Score[a]	Biological target	Action	Drugs	*p* value
+	Topoisomerase II	Inhibitor	Amsacrine, amonafide, teniposide, etoposide, idarubicin	2.829×10^{-4}
+	HDAC	Inhibitor	Panobinostat, scriptaid, dacinostat, vorinostat, trichostatin A	1.451×10^{-4}
+	Protein kinase C delta	Activator	Phorbol-12-myristate-13-acetate, ingenol	3.172×10^{-3}
+	Histone lysine methyltransferase	Inhibitor	Chaetocin	
+	ARFGAP1	Inhibitor	QS11	
+	PDK1	Inhibitor	BX795	
+	Retinoic acid receptor beta	Inhibitor	Le135	
+	Arginase	Inhibitor	Inhibitor Bec	
+	JAK2/STAT3	Inhibitor	Cucurbitacin I	
+	Fatty acid synthetase	Inhibitor	Cerulenin	
+	Src, Bcr-Abl tyrosine kinase	Inhibitor	AG957	
+	PLD2	Inhibitor	CAY10594	
+	IKKβ	Inhibitor	Parthenolide	
+	IMPDH1	Inhibitor	Mycophenolic acid	
+	FTL3	Inhibitor	Midostaurin	
+	DNA	Crosslinker	Mitomycin C	
+	Tubulin	Inhibitor	Vinblastine	
+	Hsp90	Inhibitor	Radicol	
+	Multiple targets	Inhibitor	Resveratrol	
–	PI3K	Inhibitor	PI828, GDC0941, NVP-BEZ235,PP110, TGX115	4.915×10^{-6}
–	mTOR	Inhibitor	NVP-BEZ235, AZD8055, TGX115, Ku0063794	1.792×10^{-5}
–	CDK	Inhibitor	BML259, indirubin	1.463×10^{-2}
–	IKBalfa	Inhibitor	Evodiamine	
–	Farnesyltransferase	Inhibitor	Tipifamib	
–	IGF1R	Inhibitor	Linsitinib	
–	MAP2K1	Inhibitor	Selumetinib	
–	CHK1	Inhibitor	SB218078	
–	Piruvate kinase	Inhibitor	M2PK activator	
–	Rho kinase	Inhibitor	Rho kinase inhibitor III	
–	Voltage-dependent calcium channel	Inhibitor	Nifedipine	
–	Braf	Inhibitor	Vemurafenib	

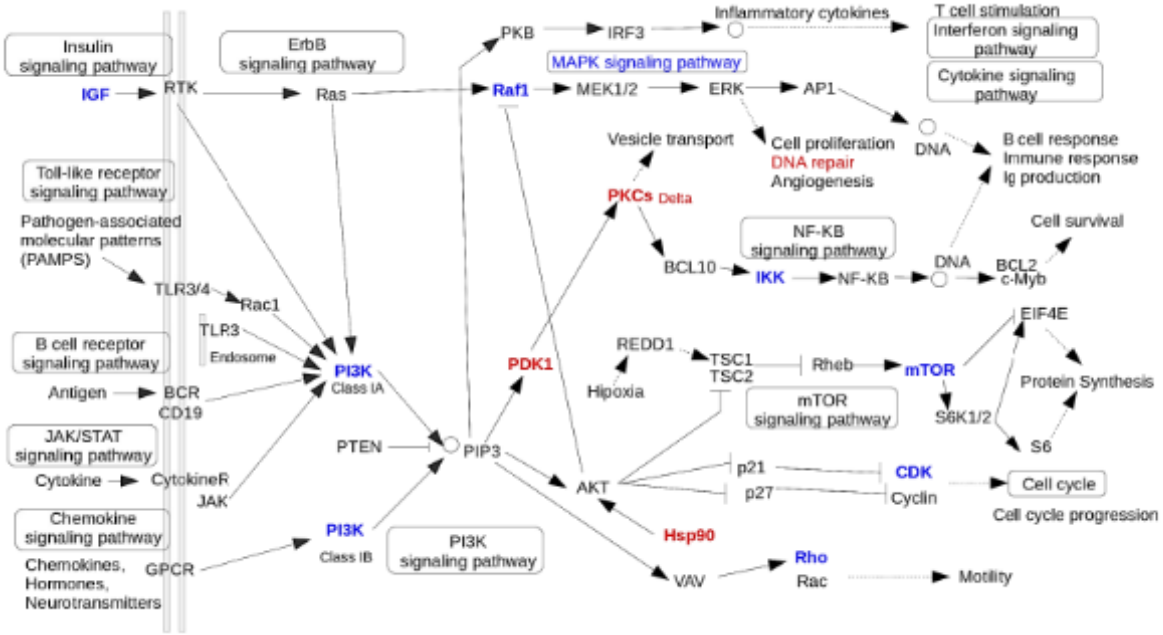
Table 2: Significant knock-down and knock-in genes obtained. Table presents knock-in genes with positive similarity score (score +), and knock-down genes with positive and negative similarity score (score –). The genes are sorted into each list by median of similarity scores across all dataset. No knock-in signatures were found with significant negative similarity score.

Score	Type of experiment	Genes
+	Knock-in	*IFNB1, IFNG, CD40, BCL10, KLF6, LYN, TIRAP*
+	Knock-down	*CLCN3, PPP1R14B, LMNB2, TBX2, PMM2, MYC, ATP6V1F, MAX, PEPD, PUF60, PHB2, AKR1A1, BTG1, ABHD2, TFDP1, PAX8, FOSL2, NTSE, RRM1, NR2F6, RAMP1, RYK, CISH, PPP2R1A, CD14, UFD1L, HTRA1, SLC35A1, TWF2, NNT, HOMER2, HS2ST1, ZNF768, GGT1, DFFB, HSPA2, PRKDC, ARPC5, NFKBIA, SLC39A8, THAP11, GSTP1, ETV1, GCAT, KIAA0907, DLX3, ELK1, PIAS4, MEOX2, GPER, NRAS, TCEB3C, KIF2C, POLR2F, CTBP2, CHAF1B, CEP55, HOOK2, ZNF8, NDUFB7, NISCH, HOXC10, AQP12A, YES1, PSMD5, JAG1, MDH2, POLR2I, DDF1, HRAS, HDAC10, SLC25A14, MED7, HMGCR, PDXP, FDX1, NIPBL, PRKAG3, PPIA, EIF2AK3, B4GALT1, UCK2, JUN, MED4, YBX1, BUB1B, CRCP, MED1, HDAC11, S8NO1*
–	Knock-down	*MITF, ETFA, PIP4K2B, VRK2, SPEN, NSDHL, ZNF586, GNPDA1, SIX4, PARN, DUSP14, IQGAP1, LRRK2, GPR123, SF1, FEZ2, IPMK, SAT1, ELF4, RPTOR, EIF4E, ARL3, KARS, CSNK1A1, SPTLC2, MEN1, SNX17, VEGFC, PPP3CA, BNIP3, ERBB3, ERO1L, COPB2, SERPINC1, AK4, HLA_A, PIK3CA, PIK3C2A, IGF2R, LYPLA1, STX4, ATM, ESPL1, IGF1R, ST3GAL5, MTOR, GRN, HSP90AA1, PRPF4B, TM9SF3*

Additional files

Additional file 1: presents information about quality control of the data, SLE signatures, and significant results obtained: Sheet 1 shows information about datasets including their GEO identifiers, the SLE state, the cell type, and the microarray platform for each, number of control and case samples, PubMed identifier, and date of publication; Sheet 2 shows results of the quality control based on the percentage of missing values and the number of significant genes in each dataset of SLE; Sheet 3 shows the signatures of each dataset used to query on Lincscloud, significant genes sorted by fold change in each signature; and Sheet 4 shows the lists of drugs and knock-in and knockdown genes with similarity scores, median of similarity scores across datasets, and significance values.

Additional file 2: is a figure showing the PI3K molecular signaling pathway. Plot constructed based on the information of different PI3K interaction graphs from the KEGG database. Red, drug targets with positive similarity scores; blue, drug targets with negative similarity scores.



Additional file 3: shows a description of the functional analysis of gene targets obtained and their significance, and is divided into three sections: methods, results, and references.

5. ESTRATIFICACIÓN DE PACIENTES DE LUPUS

5.1. EL PROBLEMA DE LA HETEROGENEIDAD EN LUPUS

Aunque hemos observado que hay mecanismos moleculares no sólo conservados en todo el conjunto de pacientes de LES sino además a través de diferentes enfermedades autoinmunes, el LES es una enfermedad altamente heterogénea. La **heterogeneidad clínica** del LES se refleja en una gran diversidad de anormalidades a diferentes niveles tanto serológicos, celulares como fenotípicos (60). No todos los pacientes comparten las mismas anormalidades, lo que sugiere que debe haber diferentes rutas biológicas que están dirigiendo la patogenia de la enfermedad en los diferentes pacientes. Esta heterogeneidad se refleja además en una **respuesta a fármacos diferente** entre diferentes pacientes, lo que resulta que un determinado fármaco pueda funcionar mejor para uno grupo de pacientes que para otro, siendo la terapia inefectiva si tomamos al conjunto de pacientes como un todo (61). Por tanto, es urgente la estratificación de pacientes de LES, o, dicho de otro modo, la búsqueda y creación de subgrupos de pacientes más homogéneos molecularmente, con el fin de abrir las puertas a una medicina más personalizada que estudie la eficiencia de los fármacos dentro de grupos más homogéneos y diferenciados de pacientes. Este tipo de análisis de estratificación usando datos de expresión génica no es algo nuevo, sino que ha sido usado ampliamente en otras patologías como el cáncer, donde se ha conseguido establecer una nueva clasificación de la enfermedad en base a

las similitudes de los perfiles de expresión de los pacientes (62,63). Aunque este tipo de investigación dentro del contexto de enfermedades autoinmunes es aún un campo yermo.

Sin embargo, la mayoría de los algoritmos de estratificación de pacientes están diseñados para agrupar muestras a partir de mediciones independientes o mediciones tomadas en un solo tiempo de la enfermedad para cada paciente. Para conseguir una estratificación real y efectiva del LES, hay que tener en cuenta además que el LES es una enfermedad dinámica y variable, caracterizada por sufrir de manera impredecible en el tiempo recaídas o brotes y remisiones, siendo estos patrones de actividad de la enfermedad diferentes entre pacientes e incluso dentro del propio paciente. Esto añade una nueva capa de heterogeneidad en el comportamiento de la enfermedad. Durante los periodos de *flares* o de alta actividad de la enfermedad es cuando se producen la mayoría de los síntomas, pudiendo aparecer desde erupciones cutáneas a fallos multi-orgánicos y alteraciones del sistema nervioso (64). La actividad de la enfermedad es medida mayormente mediante el denominado *SLE Disease Activity Index* (SLEDAI) (65), que nos indica de un modo *cuasi*-cuantitativo si un paciente está más o menos grave mediante un sistema que pondera numéricamente las diferentes sintomatologías y suma los valores de las cuales se encuentren presentes en el paciente. Aunque a ser un sistema que usa un sumatorio de elementos, pacientes con un mismo valor de SLEDAI pueden tener diferente pronóstico y diferentes manifestaciones clínicas, por lo que es un valor útil dentro de un mismo paciente, pero no directamente comparable entre diferentes pacientes, más aún si se tiene en cuenta un único punto de la enfermedad. Por ejemplo, un valor de 6 para el SLEDAI puede venir desde una remisión en un paciente y de un *flare* en otro, lo que probablemente refleje diferencias a nivel molecular.

5.2. ANÁLISIS LONGITUDINAL

Una manera de abordar todos los problemas de heterogeneidad mencionados anteriormente es mediante el análisis longitudinal de los pacientes de LES. Con **datos longitudinales** se hace referencia a que para cada paciente se toman diferentes muestras a distintos tiempos o distintos estados de la enfermedad, de tal manera que podemos extraer cuales son los mecanismos moleculares que están dirigiendo la actividad de la enfermedad en cada paciente. De este modo, podemos comparar y estratificar pacientes en base a esos mecanismos de progresión de la enfermedad.

Hay diferentes algoritmos para agrupar o estratificar en base a datos longitudinales o series temporales, comúnmente usados para buscar conjuntos de genes con comportamientos similares a través del tiempo. Por ejemplo, el paquete de R llamado TSclust (66) implementa 20 diferentes métricas y enfoques metodológicos para hacer agrupamiento con datos de series temporales. Aunque aquí surge un nuevo problema y es que, al contrario que la mayoría de las enfermedades, el LES y las enfermedades autoinmunes en general, no tienen patrones de progresión en el tiempo conocidos y fijos, por lo que no pueden definirse estados temporales dentro de la enfermedad. Esto nos lleva a la incapacidad de poder usar el tiempo como medida común entre pacientes, ya que, por ejemplo, el tiempo denominado como tiempo 1 para un paciente no tiene porqué reflejar las mismas condiciones clínicas, fenotípicas y moleculares, que el tiempo 1 para otro paciente. Y, por tanto, los métodos de agrupamiento longitudinal comúnmente utilizados no son válidos para este tipo de patologías independientes del tiempo.

Debemos destacar el artículo de Banchereau et al. (67) como trabajo pionero de estratificación longitudinal de pacientes de LES. En este trabajo, el primer paso fue extraer módulos de genes correlacionados entre sí en los diferentes puntos de tiempo para cada paciente. Posteriormente, seleccionaron el módulo que más correlacionaba con la variación del SLEDAI como única representación para cada paciente y finalmente, proyectaban estos módulos seleccionados dentro de un espacio de anotación funcional que usaban para agrupar los pacientes. De este modo, identificaron 7 grupos de pacientes en los cuales se activaban unas rutas biológicas específicas u otras cuando la actividad de la enfermedad aumentaba.

Aunque el potencial y el alcance de este artículo es innegable, durante esta tesis observamos en éste una serie de limitaciones y consideraciones que no habían sido tenidas en cuenta, lo que nos llevó a desarrollar un nuevo trabajo sobre estratificación de LES más técnica y biológicamente correcto, a nuestro ver. Entre estas consideraciones a tener en cuenta tenemos que el tamaño de los módulos seleccionados varía enormemente entre pacientes (desde 31 a 3434 genes), lo que causa un sesgo en la comparación entre pacientes. Los genes seleccionados para cada paciente pueden ser diferentes, por lo que no hay un espacio común de genes en el que comparalos. Este problema puede verse en el siguiente caso de ejemplo. Supongamos que seleccionamos el módulo 1 para un paciente, módulo relacionado con la ruta del IFN I, y para otro paciente seleccionamos el módulo 3, relacionado con otras funciones. Al proyectar estos módulos en el espacio funcional donde se agrupa, estos dos pacientes no se agruparán juntos debido a que sus módulos representan funciones diferentes, pero no podemos afirmar que el paciente 2 no esté relacionado con la ruta del IFN I. Podría darse el caso de que el módulo relacionado con el IFN I sea el segundo módulo más correlacionado para el paciente 2. Bajo este criterio de selección, por tanto, se están eliminando señales que pueden ser importantes dentro de cada paciente, además de que, en muchos casos, los valores de correlación con los que se seleccionaba un determinado módulo eran muy cercanos o incluso iguales a los de otros módulos. Además, como la selección de genes se hace de manera individual dentro de cada paciente, puede haber sesgos por agentes externos que alteren la expresión de forma individual que no son corregidos y pueden estar dirigiendo parte de la estratificación, por ejemplo, desde el uso de determinados fármacos a un simple resfriado.

Con **nuestro enfoque** de estratificación longitudinal para enfermedades complejas abordamos esta serie de limitaciones, seleccionando los genes más informativos para agrupar a través de todos los pacientes y creando un espacio común y único de genes donde realizar el análisis de agrupamiento de pacientes.

5.3 TERCER ARTÍCULO

Título: *Stratification of Systemic Lupus Erythematosus Patients IntoThree Groups of Disease Activity Progression According to Longitudinal Gene Expression*.

Dirección web: https://onlinelibrary.wiley.com/doi/full/10.1002/art.40653

El objetivo de este artículo es realizar un análisis de estratificación longitudinal para extraer subgrupos de pacientes homogéneos desde el punto de vista molecular, sujetos a los mismos mecanismos genéticos que dirigen la progresión de la enfermedad. Con este tipo de estudios de estratificación se busca reducir la heterogeneidad dentro de todo el conjunto de pacientes, abriendo las puertas a una medicina más personalizada que permita el estudio de la respuesta a los fármacos dentro de cada uno de los grupos.

Para ello, usamos dos cohortes longitudinales de pacientes de LES. Para cada paciente tenemos un numero variable de visitas en cada cual se midieron datos de expresión y un gran abanico de variables clínicas. Una cohorte será usada como cohorte de descubrimiento y la segunda será usada independientemente para intentar replicar los resultados obtenidos. El método o enfoque desarrollado y aplicado consiste en medir la correlación entre el valor de SLEDAI y los valores de expresión de cada gen de forma individual a través de las diferentes visitas de cada paciente. De este modo, pasamos de trabajar con 3 dimensiones (genes, pacientes y visitas) a generar una matriz bi-dimensional que contiene los genes en filas, los pacientes en columnas y los valores de correlación como entradas de la matriz. El valor de correlación resume el comportamiento del gen a medida que la actividad de la enfermedad aumenta o disminuye. Por ejemplo, una correlación fuertemente positiva nos dará la información de que el gen se sobre-expresa cuando la enfermedad está activa y viceversa. Posteriormente, seleccionamos los genes más informativos para agrupar priorizándolos mediante un método basado en suma de rangos, con algunas modificaciones, y creamos un espacio común con los genes seleccionados teniendo en cuenta el conjunto de los pacientes.

Como resultados obtuvimos 3 subgrupos robustos y estables los cuales fueron replicados en las 2 cohortes independientes. La caracterización clínica nos mostró que dos de los grupos mostraban una correlación positiva con neutrófilos, o lo que es lo mismo, que el porcentaje de neutrófilos aumenta cuando la actividad de la enfermedad aumenta, mientras que un tercer

grupo mostró una correlación positiva con linfocitos, es decir, en este caso es el porcentaje de linfocitos el que aumenta con la actividad de la enfermedad. Este hallazgo representa dos mecanismos de progresión de la enfermedad totalmente distintos. Además, estos grupos obtenidos mostraban diferencias significativas en cuanto a sintomatología clínica. Por ejemplo, entre otras, los grupos relacionados con neutrófilos mostraban un mayor número de casos de desarrollo de nefritis proliferativa, mientras que el grupo dirigido con linfocitos se asociaba a una mayor comorbidad con otras enfermedades autoinmunes, como el SjS.

Nuestros resultados tienen importantes implicaciones en las opciones de tratamientos, pudiendo usarse la estratificación propuesta para el estudio de fármacos y su eficacia dentro de cada grupo específicamente.

Debido a las políticas de copyright de la revista *Arthritis and Rheumatology*, en este documento recogemos el artículo en su versión *pre-print*, o versión previa al ajuste de formato por parte de la revista.

Arthritis & Rheumatology

Longitudinal Stratification of Gene Expression Reveals Three SLE Groups of Disease Activity Progression

Journal:	*Arthritis & Rheumatology*
Manuscript ID	ar-18-0419.R1
Wiley - Manuscript type:	Full Length
Date Submitted by the Author:	24-May-2018
Complete List of Authors:	Toro Dominguez, Daniel; Centre Pfizer–University of Granada–Andalusian Government for Genomics and Oncological Research (GENYO), Area of Medical Genomics; Centre Pfizer–University of Granada–Andalusian Government for Genomics and Oncological Research (GENYO), Bioinformatics Unit Martorell-Marugán, Jordi; Centre Pfizer–University of Granada–Andalusian Government for Genomics and Oncological Research (GENYO), Unit of Bioinformatics Goldman, Daniel; Johns Hopkins School of Medicine, Medicine Petri, Michelle; Johns Hopkins School of Medicine, Medicine Carmona-Saez, Pedro; Pfizer–University of Granada–Andalusian Government Centre for Genomics and Oncological Research (GENYO), Bioinformatics Unit Alarcón-Riquelme, Marta E.; Centro de Genómica e Investigaciones Oncológicas (GENYO) Pfizer-Universidad de Granada-Junta de Andalucía,
Keywords:	Systemic lupus erythematosus (SLE), Disease Activity, Gene Expression, Longitudinal Studies, stratification
<B>Disease Category: Please select the category from the list below that best describes the content of your manuscript.:	Systemic Lupus Erythematosus

John Wiley & Sons

Arthritis & Rheumatology

Longitudinal Stratification of Gene Expression Reveals Three SLE Groups of Disease Activity Progression

Daniel Toro-Domínguez, BSc[1], Jordi Martorell-Marugán, BSc[1], Daniel Goldman, PhD, MSC[2], Michelle Petri, MD, MPH[2], Pedro Carmona-Sáez, PhD*[1], and Marta E. Alarcón-Riquelme, MD, PhD*[1, 3]

[1]GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS, 18016, Granada, Spain

[2]Division of Rheumatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[3]Unit of Chronic Inflammatory Diseases, Institute of Environmental Medicine, Karolinska Institutet, 17177, Stockholm, Sweden

*Marta Eugenia Alarcón-Riquelme, M.D., PhD., e-mail: marta.alarcon@genyo.es. (ORCID Id: 0000-0002-7632-4154), and Pedro Carmona Sáez, PhD, e-mail: pedro.carmona@genyo.es. (ORCID Id: 0000-0002-6173-7255). GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS, 18016, Granada, Spain.

Running Title: Stratification of Lupus with Longitudinal Gene Expression Correlations

Keywords: Systemic lupus erythematosus, stratification, clustering, gene expression, activity scores, longitudinal

John Wiley & Sons

Arthritis & Rheumatology

Funding: Daniel Toro-Domínguez is supported through the grant GA#115565 from the Innovative Medicines Initiative Joint Undertaking of the European Union.

Abstract

Objectives: The highly heterogeneous clinical presentation of lupus is characterized by the unpredictable appearance of flares of disease activity and important organ damage. Attempts to stratify lupus patients have been limited to clinical information, leading to unsuccessful clinical trials and controversial research results. Our aim was to develop and validate a robust method to stratify patients with lupus according to longitudinal disease activity and whole-genome gene expression data in order to establish subgroups of patients who share disease progression mechanisms.

Methods: We applied a clustering-based approach to stratify SLE patients based on the correlation between disease activity scores and longitudinal gene expression information. Clustering robustness was evaluated by bootstrapping and the clusters were characterized in terms of clinical and functional features.

Results: Using two independent sets of patients, one pediatric and another adult, our results show a clear partition into three different disease clusters not influenced by treatment, race or other source of bias. Two of the clusters differentiate into a neutrophil correlated disease group and a lymphocyte correlated disease group, while the third that correlated to a lesser extent with neutrophils, was functionally more heterogeneous. The neutrophil-driven clusters were associated with increased development towards proliferative nephritis.

Conclusions: We found three subgroups of patients that show different mechanisms of disease progression and are clinically differentiated. Our results have important implications for

John Wiley & Sons

Arthritis & Rheumatology

treatment options, the design of clinical trials, the etiology of the disease, and the prediction of severe glomerulonephritis.

John Wiley & Sons

Arthritis & Rheumatology

Introduction

SLE disease activity varies unpredictably over time and this variation is heterogeneous between patients and within patients. Patients go through periods of flaring, with both disease activity and treatment with corticosteroids resulting in organ damage and premature death (1). The Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) is the most used scoring system for disease activity (2). However, patients with similar disease activity by SLEDAI may have different prognosis and treatment responses (3). Therefore, there is urgent need to establish new lupus patient stratification. The clinical heterogeneity of lupus manifests itself also in the diversity of abnormalities that have been described at cellular, serological, and other levels (4). Not all patients share the same abnormalities suggesting that specific pathways leading to active disease are different from patient to patient.

In gene expression studies, clustering analyses have been broadly used to discover sets of samples that share similarities in their gene expression programs. These have resulted in successful results establishing new disease classifications in cancer (5,6).

Nevertheless, most patient stratification algorithms are designed to cluster samples from independent measurements. Several clustering algorithms that deal with longitudinal data or time series are mainly used for defining gene clusters. For example, TSclust R package (7) implements 20 different metrics and approaches to time-series clustering. But unlike most diseases, autoimmune diseases have no known progression patterns over time and hence, defined stages of disease cannot be established. So, we cannot assume a similar time-dependent modulation for different patients. Therefore, the classical methods of patient stratification based on disease progression over time are not valid.

John Wiley & Sons

Arthritis & Rheumatology

Using two independent datasets, a previously reported and publicly available dataset of pediatric lupus patients (8) and a new adult dataset that we generated with the follow-up of the cohort reported in (9), we established three groups of patients from gene expression profiles correlated with disease activity progression in time. The clusters we report are robust and highly reproducible in both datasets differentiated by patterns of lymphocyte and neutrophil composition that occur with disease activity, the progression to proliferative nephritis, and presence of other clinical manifestations.

Materials and Methods

Population cohorts and design

We used two independent sets of SLE patients. As training set we used the public longitudinal data from Banchereau, et al (8). That study contained a unique set of pediatric SLE patients traced over time. Clinical variables and genome-wide gene expression levels were measured at different time points for every patient. Gene expression data was downloaded from NCBI GEO (ID GSE65391). Patients with less than three visits and whose SLEDAI magnitude does not change with time were discarded from analysis. A total of 80 patients were selected for further analysis, each one with a variable number of visits, continuous and categorical clinical variables, and gene expression data.

An independent dataset was generated from adult patients obtained from the SPARE (Study of biological Pathways, disease Activity and Response markers in patients with systemic lupus Erythematosus) (9) study protocol approved by Johns Hopkins University School of Medicine Institutional Review Board. Patients were enrolled from the Hopkins Lupus Cohort following informed consent. Adult patients were eligible if aged 18–75 years old and met the definition of

SLE by the revised American College of Rheumatology classification criteria from 1997. At entry (baseline), patient's medical history was reviewed, and information on current medications recorded. Visits were scheduled quarterly or more often if required for disease activity over a 2-year period. All patients were evaluated at entry and at all subsequent cohort visits (MP) by the same physician. A total of 306 SLE patients were enrolled. The demographics were 58.9% Caucasian, 33.9% African–American, 91.1% female, mean age 46.0±11.9 years. The number of visits per patient the following year ranged from 1 to 9. Six patients had 1 visit, 46 had 2–3 visits, 159 had 4 visits, and 81 had more than 4 visits.

Patients were treated according to standard clinical practice. To assess disease activity, the Safety of Estrogens in Lupus Erythematosus: National Assessment (SELENA) version of the SLEDAI as well as physician global assessment (PGA), were completed at each visit. C3, C4, anti-dsDNA (Crithidia), complete blood cell count and urinalysis were performed at every visit. Affymetrix GeneChip HT HG-U133+ arrays were used to measure gene expression profiles. The experimental protocol and gene expression data processing methods were reported (9).

Sixty-five adult patients with over three visits varying in their SLEDAI values were selected for analysis. We used this novel, unpublished longitudinal dataset to validate the results obtained with the pediatric dataset.

Processing of the data

Each set was processed independently. Transcripts with flat expression profiles and standard deviation below 0.1 across samples were removed. Retained transcripts were annotated to a gene symbol. Duplicated genes were merged assigning them their mean expression value. This yielded 15344 genes from 743 samples (80 pediatric patients) as the final gene expression training

dataset and 20741 genes from 288 samples (65 adult patients) as replication set. Healthy samples were left out from stratification analysis.

Stratification method

Because SLE is a disease with unknown progression, we cannot assume a temporal relationship between patients. For each patient we had gene expression data across a sparse and asynchronous number of visits (ranging from 3 to 22), with an associated SLEDAI score and clinical variables measured at each visit. Therefore, to stratify patients based on similarities in gene expression profiles and global disease progression, rather than to cluster individual time points, we calculated a gene by patient correlation matrix computing a stringent Pearson correlation coefficient between gene expression data and SLEDAI scores across each patients' visits (*Table S1*). Correlation values summarize behavior of expression of each gene in each patient in relation with disease activity, with positive and negative correlation values. Genes with highest absolute correlation values across samples were selected by applying the Rank Sum method (10). Briefly, genes were ranked by absolute correlation values and the sum of ranks from all patients was computed obtaining a unique score value per gene. Finally, gene scores were sorted and we calculated an empirical p-value for each gene against the probability to obtain genes with higher scores by chance. For this we randomly created 1000 bi-dimensional inter-patient matrices by altering rows and columns and comparing the new score obtained for each gene with their original score. The cut-off p-value corrected by false discovery rate was <0.05. With this approach, the possible alterations in gene expression in individual patients are not included thus selecting only genes that are highly correlated with SLEDAI. In this way the differential effect of treatment on gene expression would not influence stratification.

To scale the correlation values these were normalized for each patient from 1 to -1 as maximum and minimum values, respectively. This normalization removes the effect of the different visit numbers for different patients, as the probability to obtain higher correlations is increased when fewer points are compared. The process is summarized in Figure 1A.

Clustering analysis was performed using the normalized patient correlation matrix gene set. First, Bayesian Criterion Information (BIC) from mclust R package (11) was used to evaluate the optimal number of clusters and k-means based consensus clustering was used to stratify patients based on the optimal number of K. This analysis was performed using the ConsensusClusterPlus R package (12)(*Figure 1B*).

Cluster stability

Cluster robustness was evaluated by a bootstrap-based approach. Subsets of 75, 62 and 50 percent of samples were randomly selected from the original dataset and clustering analysis repeated. We calculated the stability of the clusters using the Jaccard coefficient by comparing our original clusters with the new clusters obtained in each subset and mean Jaccard coefficient across the cluster permutations. This analysis was performed using clusteval R package. We evaluated the robustness of the feature selection and clustering results with respect to the number of visits. For this, we randomly selected 3, 4 or 5 visits for each patient and re-calculated the correlation values. We constructed the bi-dimensional matrix resulting from the genes selected in the original feature selection but with new correlation values for each number of visits, and repeated the clustering procedure. We performed 1000 permutations and estimated the stability with Jaccard.

Functional analysis of the clusters

To characterize the main biological processes associated to each cluster we performed functional enrichment analysis of Gene Ontology terms (http://www.geneontology.org/) in the set of genes selected in the feature selection using enrichR (13). To evaluate broader groups of biological processes GO terms were re-annotated and grouped with terms from the common superior levels of GO hierarchy, and compared correlation values of the genes of each category between the different clusters obtained. This analysis determines the biological functions that are represented by genes increasing or decreasing in their transcription with respect to disease activity in the clusters, allowing us to discern differential biological mechanisms between clusters.

Tests for association of clinical variables

Fisher's exact test was used to evaluate if categorical variables (gender, race, and treatment) were enriched in a cluster with respect to the others. For continuous variables, we measured their correlation with SLEDAI scores for each patient and ANOVA test was applied to evaluate if clusters were enriched in samples with variables highly correlated with disease activity scores in pair-wise comparisons. With this analysis, we could measure not only the clinical variables that determine disease activity in each cluster, but also those differentially correlated with disease activity between clusters. Parameters measured on pediatric patients were downloaded from http://websle.com along with expression data.

Analysis of Treatment effects

To analyze if our selected stratification genes were modulated by treatment use we correlated gene expression of each patient with treatment doses for acetylsalicylic acid, cytotoxic drugs and prednisone, the treatments for which we had dose information in adults. No data was available for the pediatric set. We performed the rank sum method to select the top 100 positively and 100 negatively most drug-correlated genes for each treatment, that is, genes modified by treatment

doses. We compared these three lists of drug-correlated genes with the genes selected for stratification.

Finally, we analyzed if treatments were differentially affecting cell proportions between clusters. We measured the tendency of neutrophil proportions of each patient at the time point when treatment was applied and at the next visit and compared the results between clusters.

Imputation of Cell Proportions

Twelve adult patients had no cell proportions available. Missing neutrophil cell proportions were imputed using CiberSort (14). We used real cell proportions from the pediatric set as control measuring the correlation of real and imputed data. Missing lymphocyte proportions were not imputed because CiberSort has different lymphocyte subtypes obtained through different methods so comparison with whole lymphocyte data is not direct.

Modular Functional Analysis

To assess differences between clusters in terms of biological pathways we defined the set of genes differentially expressed in each cluster using limma [15] selecting those genes with corrected p-values < 0.05. Tmod R package [16] was used to determine functional pathways differentially over and under-represented in each cluster. This analysis was performed separately in samples with low and high SLEDAI categories (scores<3 or >8, respectively).

Results

Gene expression correlation with activity results in three clusters or subgroups of lupus patients

The gene selection process of the pediatric patient set yielded 777 significant genes (*Table S1)*. Not unexpectedly, a large number of genes belonged to the type I interferon signature (17, 18).

We found k=3 as the optimal number of clusters (*Figure 2A and 2B)*. We named the clusters P1 (P: pediatric), with 31, P2, with 21 and P3 with 28 patients (*Table S2*).

For the adult set, 1051 significant genes were selected (*Table S1*). Again, k=3 was obtained as best cluster number (*Figure 2C and Figure 2D*). The three clusters, A1 (A: adult), A2 and A3 grouped 20, 16, and 29 patients, respectively (*Table S2*).

The clusters are highly stable and not biased by demography or treatment

Cluster stability measurements are summarized in *Table S3*. Of the two methods used for cluster validation, only when sample size was halved did bootstrapping give low stability retaining still a high Jaccard coefficient value in the pediatric (0.77), and adult set (0.7). In the feature selection stability test, no individuals were miss-classified, demonstrating high reproducibility of the clusters. These results showed that the clusters are highly stable and the genes selected for stratification are maintained independently of the number of visits (for correlation, at least three visits must be measured).

Table S4 shows results comparing demographic or treatment variables and *Table S5* numbers of patients for each variable between clusters. In neither the three pediatric nor the three adult clusters did we observe statistical association with race, gender, or treatment, excluding these as drivers for stratification.

Behavior of numbers and proportions of neutrophils and lymphocytes in time differentiate the clusters

Figure 3A shows continuous variables significantly enriched in the pediatric clusters. As can be noted, there is differential distribution of neutrophil and lymphocyte proportions in the obtained clusters that correlate with disease activity, that is, increase or decrease with SLEDAI. The

Arthritis & Rheumatology

sharpest differences were observed in proportions of neutrophils and lymphocytes between clusters P2 and P3. The cellular proportion mean was not biased between clusters (*Table S6*) and SLEDAI values of patients from each cluster were within the same ranges. This means that the percentage of neutrophils increased with activity in clusters P2 and P1, and decreased with activity in P3 (*Figures S1*). Percentage of lymphocytes showed an opposite trend. Thus, proportions of these cellular populations had a completely different behavior between clusters despite having in average similar disease activity (see below). Other interesting differences between clusters were observed with C3 and C4 complement levels. The correlation values were strongly negative in patients from clusters P2 and P3 and less negatively correlated in P1, having this cluster, enrichment in patients that develop proliferative nephritis (see below). We found also correlation with increased aspartate aminotransferase, a hepatic function enzyme (19) in P3 and a somewhat higher erythrocyte sedimentation rate (ESR) in P1.

Figure 3B shows the significant continuous clinical variables of the adult set. We used pediatric set as control of the cell proportions imputation for the missing data (*Figure S2*). Neutrophil proportions decreased in cluster A3 when the SLEDAI increased and increased in clusters A2 and A1, with increased activity, and the lymphocytes go in the opposite direction as in the pediatric clusters. For the 12 patients whose lymphocyte data was missing, the correlation data with SLEDAI was set to zero. C3, C4 and ESR conserved the same pattern in both sets, but differences between clusters A1 and A2 were less marked. So, in two independent analyses applied to different datasets we found a similar partition of SLE patients associated with different cell behavior following disease activity, differentiating A3 and P3 as lymphocyte-driven clusters, and the remaining groups as neutrophil-driven clusters.

Treatment did not influence cluster formation

Treatment can modulate gene expression and blood cell proportions (20, 21). We tested whether expression of the selected genes could be modulated by doses of cytotoxic drugs, acetylsalicylic acid or by prednisone. The presence of treatment-modulated genes did not exceed 2% of our gene selection (*Figure S3*). Using trajectory analyses of the pediatric set we observed that treatment did not affect neutrophil proportions between clusters (*Figure S4*). Thus our stratification approach is not influenced by treatment, and treatment is not differentially influencing cell proportions between clusters.

Clusters are not influenced by disease activity

We then analyzed if there was differential distribution of disease activity scores between clusters. Pediatric clusters had mean SLEDAI scores of 6.45, 6.44 and 6.65, for P1, P2 and P3, respectively. Adult SLEDAI scores were lower and less variable, and had mean values of 3.29, 2.31, and 2.80, excluding such bias. Also, there was no difference in the overall magnitude of the change in SLEDAI across visits when comparing the clusters (*Table S6*). We also independently evaluated the clinical variables that compose the SLEDAI score. Differences were found only when the SLEDAI score was between 8-11 in the pediatric clusters (*Figure S5*). Specifically, the only SLEDAI related variables showing significant differences were pyuria and hematuria for cluster P2 ($P = 0.0064$ and $P = 0.0028$, respectively), and pyuria for P3 ($P = 0.0449$), as compared to the other clusters. Therefore the clusters were similar and clinically indistinguishable by SLEDAI parameters.

We analyzed the trajectory of severe proliferative nephritis development, the most common and serious organ affectation in pediatric lupus. We found that 65% of patients of cluster P1

developed nephritis with time compared to 53% and 45% of patients from clusters P2 and P3, respectively. A tendency towards nephritis for P1 and P1-P2 combination compared with P3 (Fisher´s exact test P=0.12 and p=0.16 respectively) was observed. In the adult set we observed the same pattern, where 45%, 42% and 13% of patients in clusters A1, A2 and A3, respectively, developed proliferative nephritis. Nephritis was significantly enriched in A1, A2 and A1-A2 combination compared with A3 (p=0.022, p=0.035, and p=0.014, respectively). So, clusters P1 and A1 show a high nephritis incidence followed by clusters P2 and A2, both with neutrophil-driven disease activity, suggesting a direct relationship between the neutrophil-driven clusters and risk to develop severe nephritis.

We had additional clinical variables for the adult set. *Figure 3C* shows those significantly enriched in each cluster. Cluster A1 was enriched in renal damage-related manifestations, A2 in lymphopenia, similar to the lymphocyte decrease observed for P2, lymphadenopathy and interstitial pulmonary fibrosis, and A3 was enriched in patients with secondary Sjögren's syndrome, photosensitive rash, and signs of anti-phospholipid syndrome, among others. Increased levels of aspartate aminotransferase activity correlated with cluster P3, showing that both adult and pediatric cluster 3 is related to abnormal hepatic function (*Figure 3C*).

Clusters P2/A2 and P3/A3 showed opposite correlation of activity with biological pathways, while P1 and A1 were heterogeneous

Figure 4A and *4B* show the fifteen top biological pathways represented by the selected genes stratifying pediatric and adult patients, respectively.

Interestingly, the genes forming the clusters and their biological pathways in P2 and A2 correlated with SLEDAI in a pattern opposite to that of P3 and A3. For example, cluster P2 had a

strong positive correlation with type I interferon, infection and cytokine-mediated signaling pathways, found also in A2, while in P3 this correlation was strongly negative. The same differences were observed in other pathways. In clusters P1 and A1, the biological pathways correlated in both directions. The genes in these clusters were heterogeneous in their response to disease activity. If we focus on individual pathways, some patients of clusters P1 and A1 could be classified into P2 or P3 and A2 or A3, respectively, but if we consider all pathways and correlated genes, the functional profiles of patients in P1 and A1 were totally different from the profiles of the other two clusters. So, we can differentiate 3 groups by their differential biological function and the behavior of the correlated genes.

Modular functional pathways are different for the different clusters

In order to specifically address functional differences between clusters, we performed a functional modular analysis comparing clusters at low and high SLEDAI scores and differential gene expression. *Figure 5A* and *5B* summarize the modular pathways with differential gene expression between clusters. A comprehensive modular analysis is shown in *Figure S6*. During high activity (>8) clusters P3 and A3 were over-represented with T lymphocyte-related functions and under-represented of neutrophil-related functions, while clusters P2 and A2 showed the opposite pattern. Interestingly, type I IFN-related pathways were over-expressed in clusters P3 and A3 at low SLEDAI values (*Figure S6*). Therefore, the clusters are differentially correlated in neutrophil and lymphocyte cell proportions, but also, the functions related with these cell types are differentially expressed between them.

Discussion

Arthritis & Rheumatology

We propose a lupus stratification based on longitudinal gene expression data that robustly correlates with disease activity and shows clear clinical, functional, and cellular differences. This stratification reveals parameters that may be used in predictive models of disease progression. Our results suggest that different immune system mechanisms occur during disease activity that may determine the predisposition towards developing different clinical manifestations.

Banchereau, et al., (8) established patient stratification applying a weighted gene co-expression network analysis (WGCNA) (22). For each patient they selected a gene module that best correlated with the SLEDAI score over time (called SLEDAI WGCNA module). Although useful, this strategy has drawbacks. The number of genes selected for each module shows large variation between patients (ranging from 31 to 3434 genes), which can bias patient comparison. The selected genes were different between patients, implying a lack of a common gene space for clustering patients. Thus, Banchereau proposed to stratify patients by comparing behavior of different genes between patients indirectly, in a functional common space. This space represented by the WGCNA modules selected for each patient projected by correlation into predefined functional and cellular gene modules obtained from blood (23) is from where seven groups were obtained. Therefore, patient stratification was subjected to a set of predefined modules, which could underestimate relevant relationships not found in the pre-established modules. By selecting one module for every patient, genes were discarded that might correlate with disease activity. Having selected the genes most correlated with the SLEDAI profile in each patient individually the global behavior of these genes was not evaluated in all patients and therefore, gene expression alterations caused by external factors such as treatment may have influenced their analysis.

John Wiley & Sons

Arthritis & Rheumatology

Our approach considers a common gene correlation space for all patients. The correlation space is constructed calculating the correlation of each gene and a continuous clinical variable for every patient (SLEDAI). The method is useful for complex diseases where samples have been taken at different times or disease states for a variable number of visits. In addition, data does not need to be corrected for treatment or other confounders that affect gene expression, as our feature selection approach considers as significant genes having a strong correlation value across all patients or homogeneous groups of patients, removing possible individual alterations. We validated the stability of our clusters.

We established three clearly differentiated clusters of SLE patients replicated in two independent sets. We obtained largely the same cellular behavior, clinical and functional patterns across pediatric and adult sets in spite of widely reported differences between pediatric and adult patients. Clinically the groups show interesting similarities, such as hepatic disease in cluster P3 and A3, and increased risk of proliferative nephritis in neutrophil-driven clusters.

The completely opposite behavior of neutrophils and lymphocytes between clusters 2 and 3 leads us to conclude that the involvement of specific cell types is key to differentiating SLE patients during disease activity and suggests a fundamental difference in the mechanisms driving disease activity. Those driven by lymphocytes had functional pathways related to all lymphocytic populations: B cells, T cells, and NK cells, while those driven by neutrophils also have monocyte related biological functions. Clusters A1 and P1 were, however heterogeneous, representing the most severe renal disease cases. Intuitively in these patients the disease appears to be driven mainly by neutrophils, but both cell types appear to play functional roles, with variable patterns of gene expression and cellular responses during disease activity. Why this is so, remains unexplained.

John Wiley & Sons

The SLEDAI is an activity index that detects disease activity with an overrepresentation towards nephritis, so it might be perceived that there is a bias when clusters separate nephritis cases in our data. However this is highly improbable for several reasons. The mean SLEDAI value between clusters and the magnitude of its components were very similar with small differences between clusters 2 and 3 only at indexes between 8-11. At this point we do not have a set of patients with other activity index, such as BILAG, which is broader in the components of activity it detects. However, differences as substantial and dependent on the activity-driving cell types suggests that most probably we would observe the same pattern as we do using SLEDAI. Our result therefore supports the role of neutrophils in severe nephritis (8) but not in other manifestations, such as Sjögren´s syndrome (24, 25).

That three gene expression samples with varying SLEDAIs per patient are required to estimate the clusters makes it clinically unfeasible to use this method to classify new patients. A classifier on a single time is necessary.

The molecular mechanisms behind the clusters are unknown. Evaluation of the cytokine signaling group provides insight (*Figure S7*). P2 showed increased expression of a precursor of LL-37, *CAMP* (26), and the necroptosis gene *RIPK3* (27). *STAT4* a T cell transcription factor and SLE susceptibility gene (28), *CARD11*, a B lymphocyte differentiation gene (29), and *LAG3*, a T and NK cell differentiation gene (30) were increased in P3, among others.

From the point of view of a disease like lupus, the clusters we identified or the parameters strongly associated with them could be used in future studies to improve therapy used and achieve greater efficacy. It may be possible to prevent severe nephritis having a selection of genes and cellular counts at hand to define the cluster to which they belong.

In short, we suggest three mechanisms of lupus progression influenced by cell proportions and their expressed gene having different behavior in time.

References

1. Pons-Estel G, Alarcón GS, Scofield S, Reinlib L, Cooper GS. Understanding the epidemiology and progression of SLE. *Semin Arthritis Rheum* 2010; 39: 257-268.

2. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI.A disease activity index for lupus patients.The Committee on Prognosis Studies in SLE. *Arthritis Rheum* 1992; 35: 630-640.

3. Chambers SA, Rahman A, Isenberg DA. Treatment adherence and clinical outcome in systemic lupus erythematosus. *Rheumatology* 2007; 46: 895–898.

4. Mohan C, Putterman C. Genetics and pathogenesis of systemic lupus erythematosus and lupus nephritis. *Nat Rev Nephrol*2015; 11: 329-341.

5. Sorlie T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur J Can* 2004; 40: 2667–2675.

6. Wang C, Machiraju R, Huanga K. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods* 2014; 67: 304-312.

7. Montero P, Vilar JA. TSclust: an R package for time series clustering. *J Statistic Soft* 2014; 62: 1-43.

8. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M *et al*. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 2016; 165: 551–565.

9. ZollarsE, CourtneyS, WolfB, AllaireN, RangerA, Hardiman Get al. Clinical application of a modular genomics technique in systemic lupus erythematosus. Progress towards precision medicine. *Int J Genomics* 2016; ID:7862962, 7 pages.

10. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol* 2005; 3: 1171-1189.

11. Fraley C, Raftery AE, Murphy TB, Scrucca L. Mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density Estimation. *J Am Stat Assoc* 2012; 97: 611-631.

12. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010; 26: 1572–1573.

13. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; 44: 90-97.

14. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Yet al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*2015;12, 453-457.

15. ME. Ritchie, B. Phipson, D. Wu, Y. Hu, CW. Law, W. Shi and GK. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 2015; 201543: pp. e47.

16. Weiner 3rd J, Domaszewska T. tmod: an R package for general and multivariate enrichment analysis. *PeerJ Preprints* 2016; 4: 22420v1.

17. Crow MK. Type I interferon in the pathogenesis of lupus. *J Immunol* 2014; 192: 5459–5468.

18. 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 ResTher* 2014;16: 489-x.

19. Liu Y, Yu J, Oaks Z, Marchena-Mendez I, Francis L, Bonilla Eet al. Liver injury correlates with biomarkers of autoimmunity and disease activity and represents an organ system involvement in patients with systemic lupus erythematosus. *Clin Immunol* 2015; 160: 319-27.

20. Tchetina EV, Pivanova AN, Markova GA, Lukina GV, Aleksandrova EN, Aleksankin AP et al. Rituximab Downregulates Gene Expression Associated with Cell Proliferation, Survival, and Proteolysis in the Peripheral Blood from Rheumatoid Arthritis Patients: A Link between High Baseline Autophagy-Related ULK1 Expression and Improved Pain Control. *Arthritis* 2016; 12.

21. Salmon JH, Cacoub P, Combe B, Sibilia J, Pallot-Prades B, FainOet al. Late-onset neutropenia after treatment with rituximab for rheumatoid arthritis and other autoimmune diseases: data from the AutoImmunity and Rituximab registry. *RMD Open*2015; 1:e000034.

22. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; 9: 559.

23. Chaussabel D, Quin C, Shen J, Patel P, Glaser C, Baldwin N *et al*. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* 2008; 29: 150-164.

24. Hochberg MC, Boyd RE, Ahearn JM, Arnett FC, Bias WB, Provost TT*et al*. Systemic lupus erythematosus: a review of clinico-laboratory features and immunogenetic markers in 150 patients with emphasis on demographic subsets. *Medicine (Baltimore)* 1985; 64:285-95.

25. Gustafsson JT, Herlitz Lindberg M, Gunnarsson I, Pettersson S, Elvin K, et al. Excess atherosclerosis in systemic lupus erythematosus,-A matter of renal involvement: Case control

study of 281 SLE patients and 281 individually matched population controls. *PlosONE* 2017; 12: e0174572.

26. Garcia-Romo GS, Caielli S, Vega B, Connolly J, Allantaz F, Xu Z, *et al*. Netting neutrophils are major inducers of type I IFN production in pediatric systemic lupus erythematosus. *Sci Trans Med* 2011; 73: 73ra20.

27. Sun L, Wang H, Wang Z, He S, Chen S, Liao D*et al*. Mixed lineage kinase domain-like protein mediates necrosis signaling downstream of RIP3 kinase. *Cell*2012; 148: 213-27.

28. Hagberg N, Joelsson M, Leonard D, Reid S, Eloranta ML, Mo J, et al. The *STAT4* SLE risk allele rs7574865[T] is associated with increased IL-12-induced IFN-γ production in T cells from patients with SLE. *Ann Rheum Dis* 2018; Epub Ahead of print.

29. Brohl AS, Stinson JR, Su HC, Badgett T, Jennings CD, Sukumar G*et al*. Germline CARD11 Mutation in a Patient with Severe Congenital B Cell Lymphocytosis. *J Clin Immunol* 2015; 35:32-46.

30. Triebel F, Jitsukawa S, Baixeras E, ,Genevee C, Viegas-Pequignot E et al. LAG-3, a novel lymphocyte activation gene closely related to CD4. *J Exp Med* 1990; 171:1393-405.

Acknowledgements: The data for analysis of the pediatric lupus patients was downloaded from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/) with the identifier GSE65391.

The authors would like to thank Ann Ranger, Normand Allaire, Chris Roberts and Huo Li, who contributed to the SPARE study at Biogen and produced the gene expression data for the adult SLE patients.

Author contributions: MEAR conceived the project, DT, PC and MEAR designed the project and supervised the analyses; MP and DG prepared the adult patient data and provided the

information; DT and JMM performed the analyses; DT, PC and MEAR wrote the manuscript and all authors approved its content.

Competing interests: None

Data and materials availability: Pediatric data is publicly available in GEO through ID: GSE65391.

Figure Legends

Figure 1. Summary of the clustering process. A) Obtaining the bi-dimensional correlation matrix. We part from a dataset with genes in rows, patients in columns and different visits for which we have gene expression and clinical variables for every patient. A correlation value for each gene and each patient is calculated throughout visits and a bi-dimensional matrix of genes, patients and correlation values is created. Feature selection is applied to select the best gene candidates to stratify and filter out the rest of genes. B) We performed consensus clustering on pediatric and adult bi-dimensional correlation matrices independently and then a functional and clinical characterization is performed.

Figure 2. Evaluation of the best number of clusters A) The plot shows the BIC values in y-axis and number of clusters in x-axis. The optimal number of clusters was three, value at which we found the highest BIC value. B) Stratification of pediatric patients using ConsensusClusterPlus R package. Rows and columns represent patients in the same order and color intensity represent the probability of that two patients clustering together. C) Estimation of the optimal number of clusters for the adult patient set, resulting in 3 as the best number of clusters. D) Stratification of the adult patients.

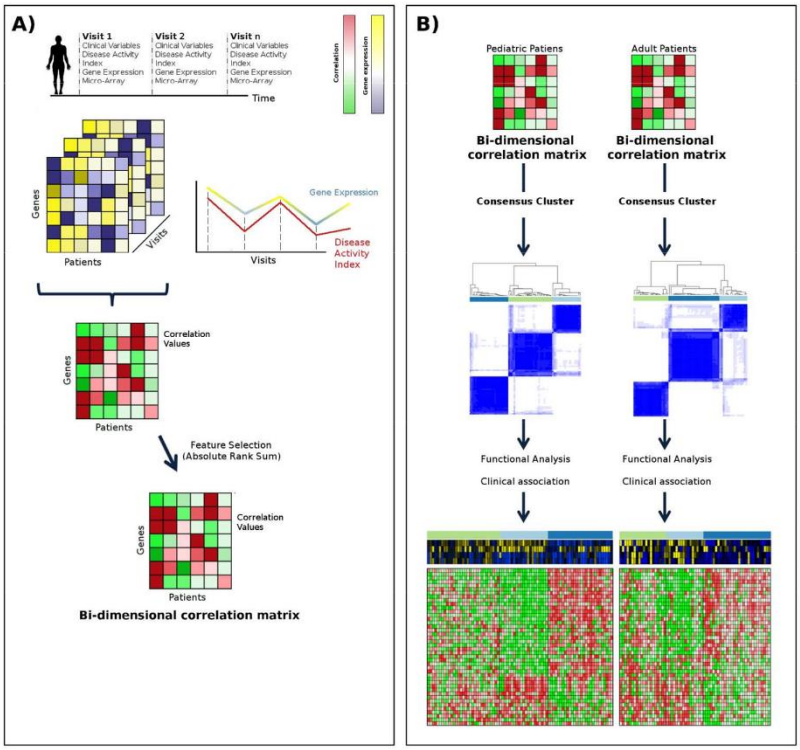
Figure 3. Clinical characterization of the clusters. A) Heatmap showing continuous significant clinical variables found in the pediatric set. Columns and rows represent the different patients and clinical variables, respectively. Summarized by the color scale is the correlation between SLEDAI and each clinical variable. The significance was obtained using ANOVA when comparing the correlation values of each cluster with respect to the others. B) Continuous significant variables found in adult set. C) Categorical clinical variables significantly enriched in adult set. The enrichment was calculated by Exact Fisher's test. Color represents the p-value of the enrichment.

Figure 4. Functional analysis of the genes selected to stratify. The color scale represents the correlation between gene expression and SLEDAI across visits. Significant Gene Ontology pathways were defined by EnrichR web tool and GO pathways were grouped according to the highest common hierarchy level (see methods).

Figure 5. Modular functional analysis according to disease activity. Summary of the significant modular pathways related with cluster's progression across sets**.** We selected patients from each set with SLEDAI higher than 8 and performed a differential gene expression analysis between clusters. The same analysis was performed selecting patients with SLEDAI less than 3. With the lists of significant genes resulting in each differential gene expression comparison, we performed functional modular analyses. A) Red and blue color intensity represents the percentage of the genes from each modular pathway that is significantly over or under-expressed in a cluster respect to the other clusters at the same SLEDAI range, respectively. All significant

pathways are shown in *Figure S5*. B) Red and blue color intensity represents the mean of the log2 fold change ofall significant genes that appears in each module for each subset.

Arthritis & Rheumatology



Fiure 1. Summary of the process. A) Obtaining the bi-dimensional correlation matrix. We part from a dataset of genes in rows, patients in columns and different visits for which we have gene expression and clinical variables for every patient. A correlation value for each gene and each patient is calculated throughout visits and a bi-dimensional matrix of genes, patients and correlation values is created. Feature selection is applied to select the best gene candidates to stratify and filter out the rest of genes. B) We performed consensus clustering on pediatric and adult bi-dimensional correlation matrices independently and then a functional and clinical characterization is performed.

149x138mm (300 x 300 DPI)

Arthritis & Rheumatology

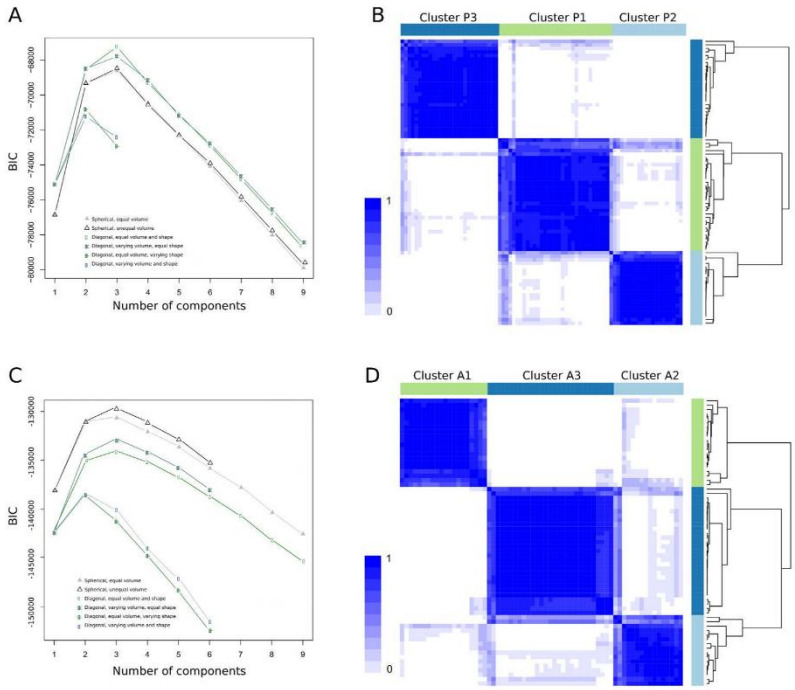


Figure 2.Evaluation of the best number of clusters A) The plot shows theBIC values in y-axis and number of clusters in x-axis. The optimal number of clusters was three, value at which we found the highest BICvalue. B) Stratification of pediatric patients using ConsensusClusterPlus R package. Rows and columns represent patients in the same order and color intensity represent the probability of that two patients clustering together. C) Estimation of the optimal number of clusters for the adult patient set, resulting in 3as the best number of clusters. D) Stratification of the adult patients.

185x169mm (300 x 300 DPI)

Arthritis & Rheumatology

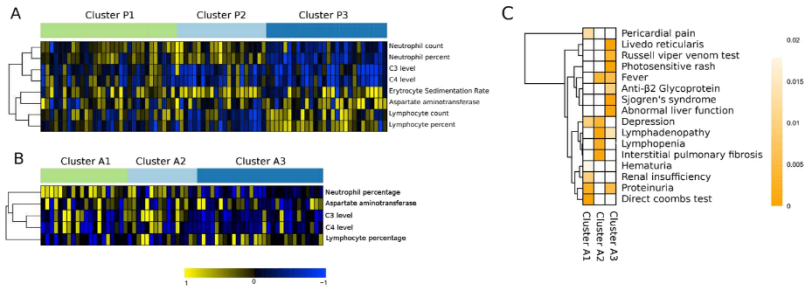


Figure 3. Clinical characterization of the clusters. A) Heatmap showing continuous significant clinical variables found in the pediatric set. Columns and rows represent the different patients and clinical variables, respectively. Summarized by the color scale is the correlation between SLEDAI and each clinical variable. The significance was obtained using ANOVA when comparing the correlation values of each cluster with respect to the others. B) Continuous significant variables found in adult set. C) Categorical clinical variables significantly enriched in adult set. The enrichment was calculated by Exact Fisher's test. Color represents the p-value of the enrichment.

185x68mm (300 x 300 DPI)

Arthritis & Rheumatology

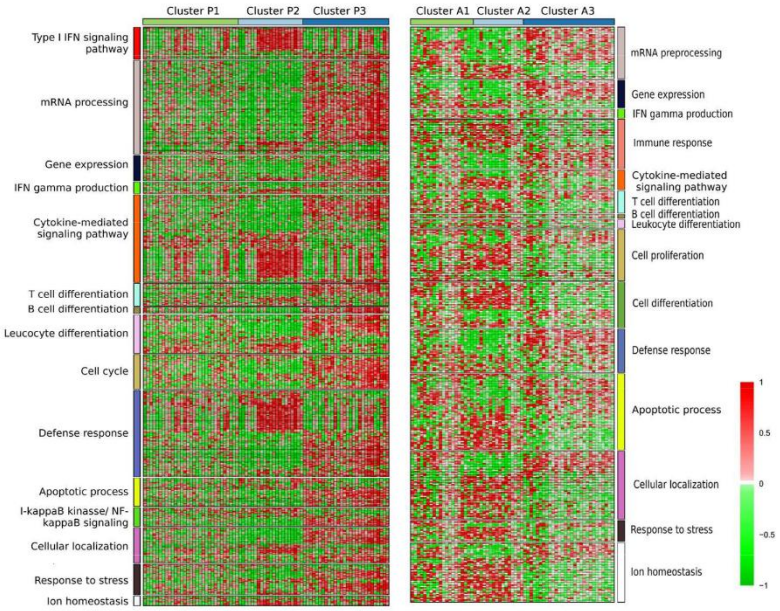


Figure 4. Functional analysis of the genes selected to stratify. The color scale represents the correlation between gene expression and SLEDAI across visits. Significant Gene Ontology pathways were defined by EnrichR web tool and GO pathways were grouped according tothe highest common hierarchy level (see methods).

185x169mm (300 x 300 DPI)

Arthritis & Rheumatology

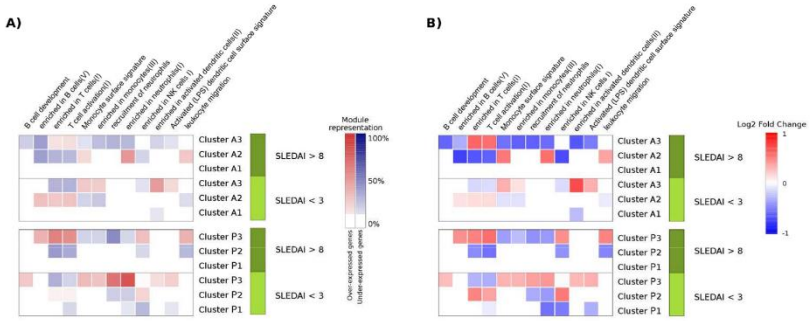


Figure 5.Modular functional analysis according to disease activity.Summary of the significant modular pathways related with cluster's progression across sets. We selected patients from each set with SLEDAI higher than 8 and performed a differential gene expression analysis between clusters. The same analysis was performed selecting patients with SLEDAI less than 3. With the lists of significant genes resulting in each differential gene expression comparison, we performed functional modular analyses. A)Red and blue color intensity represents the percentage of the genes from each modular pathway that is significantly over or under-expressed in a cluster respect to the other clusters at the same SLEDAI range, respectively. All significant pathways are shown in Figure S5. B) Red and blue color intensity represents the mean of the log2 fold change ofall significant genes that appears in each module for each subset.

254x104mm (300 x 300 DPI)

Arthritis & Rheumatology

Supplementary Materials

Toro, et al.: "Longitudinal Stratification of Gene Expression Reveals Three SLE Groups of Disease Activity Progression"

Table S1. Bi-dimensional correlation matrices. Sheet 1 of the Excel file contains the genes selected to stratify and their correlation values across patients of pediatric patients. Sheet 2 contains the genes selected to stratify and their correlation values across patients of adult patients.

Table S2. Classification of patients into the clusters. The Table contains the patient identifiers and the cluster to which they belong.

Cluster	Patients
Cluster P1	SLE-281, SLE-144, SLE-192, SLE-163, SLE-202, SLE-182, SLE-178, SLE-189, SLE-55, SLE-79, SLE-60, SLE-183, SLE-180, SLE-123, SLE-188, SLE-179, SLE-242, SLE-170, SLE-224, SLE-327, SLE-225, SLE-110, SLE-169, SLE-172, SLE-133, SLE-34, SLE-90, SLE-129, SLE-99, SLE-125, SLE-241
Cluster P2	SLE-197, SLE-168, SLE-308, SLE-201, SLE-200, SLE-155, SLE-143, SLE-325, SLE-212, SLE-175, SLE-256, SLE-254, SLE-176, SLE-40, SLE-331, SLE-264, SLE-78, SLE-304, SLE-152, SLE-199, SLE-234
Cluster P3	SLE-31, SLE-128, SLE-210, SLE-271, SLE-138, SLE-171, SLE-321, SLE-211, SLE-184, SLE-141, SLE-181, SLE-177, SLE-80, SLE-213, SLE-244, SLE-260, SLE-231, SLE-166, SLE-324, SLE-21, SLE-121, SLE-150, SLE-252, SLE-233, SLE-313, SLE-218, SLE-95, SLE-65
Cluster A1	1206, 1409, 1463, 1679, 1764, 1792, 1869, 1871, 2020, 2122, 2129, 24, 371, 966, 981, 1480, 1335, 1520, 1944, 704
Cluster A2	1436, 1842, 1041, 1227, 1702, 1913, 2016, 345, 699, 759, 1176, 113, 1620, 1705, 1924, 2067
Cluster A3	1001, 1174, 1179, 1182, 1269, 1807, 458, 911, 1699, 2103, 1052, 1178, 1263, 1424, 1478, 1537, 1927, 1938, 2104, 2119, 2128, 2132, 244, 317, 365, 453, 46, 582, 725

Table S3.Cluster stability results. The table contains the different cluster validation approaches tested and the mean Jaccard coefficient obtained for each cluster and test. Number 1 represents complete or total stability. *[1] In parenthesis, the number of visits; *[2]In parenthesis, the percentage of patients. *[3] The analysis was not performed because patients had less than three visits or time points. Global stability measures overall stability and it is calculated comparing the three clusters as one.

Validation Approach	Cluster P1 stability	Cluster P2 stability	Cluster P3 stability	Global pediatric stability	Cluster A1 stability	Cluster A2 stability	Cluster A3 stability	Global adult stability
Number of visits (5)*[1]	1	1	1	1	*[3]	*[3]	*[3]	*[3]
Number of visits (4) *[1]	1	1	1	1	*[3]	*[3]	*[3]	*[3]
Number of visits (3) *[1]	1	1	1	1	*[3]	*[3]	*[3]	*[3]
Bootstrapping (75%) *[2]	0.89	0.9	0.95	0.95	0.947	0.813	0.971	0.8741
Bootstrapping (62.5%) *[2]	0.83	0.9	0.94	0.86	0.905	0.767	0.865	0.827

Bootstrapping (50%) *[2]	0.77	0.89	0.87	0.81	0.879	0.7	0.957	0.779

Table S4. Association among demographic variable and medication variables with the clusters. The table contains the demographic variables with their p-values obtained by Fisher exact test comparing each cluster versus the rest. Each patient was considered once in any visit having a positive manifestation * Marks no available information to perform the analysis.

Categorical Variables							
Clinical variable	**Category**	**ClusterP1**	**ClusterP2**	**ClusterP3**	**Cluster A1**	**Cluster A2**	**Cluster A3**
Gender	Female/ Male	1	0.42	0.706	1	1	1
Race	Caucasian, White	0.184	0.124	1	0.289	0.559	0.8035
	Hispanic	0.061	0.601	0.158	*	*	*
	African-American, Black	0.429	0.766	0.175	0.259	1	0.4319
Treatment	Steroids	0.71	0.287	0.17	*	*	*
	Cyclophosphamide	1	1	1	*	*	*
	Oral steroids	0.508	0.881	0.582	*	*	*
	Mycophenolate	1	0.631	0.654	*	*	*
	Hydrochloroquine	0.694	1	0.587	*	*	*
	Methotrexate	0.652	1	0.338	*	*	*
	NSAIDs	0.831	0.342	0.823	0.816	1	1
	Acetylsalicylicacid	0.103	0.269	0.61	1	0.506	0.724
	Prednisone	*	*	*	0.695	1	0.858
	Immunosupressors	*	*	*	0.856	1	0.738
	Cytotoxic medicine	*	*	*	0.856	1	0.738
	Plaquenil	*	*	*	0.879	0.87	0.4
	Anti-HTN	*	*	*	0.746	1	0.765
	Triam	*	*	*	1	0.506	0.724
	Statin	*	*	*	0.825	0.1928	0.4106

Table S5. Detailed clinical information. This table shows the number of patients for each clinical variable.

		Cluster A1	Cluster A2	Cluster A3	Cluster P1	Cluster P2	Cluster P3
Gender	Male	0	0	0	3	3	2
	Female	20	16	28	28	18	26
Race	Asian	1	0	1			
	White	10	11	17			
	Black	9	5	8			
	Other	0	0	2			
	African-American				6	3	11
	Hispanic				23	10	13
	Caucasian				2	5	4
Age Onset	Mean	28.72	26.91	27.15	13.87	12.59	12.94
	Standard deviation	11.42	12.13	11.97	2.1	2.63	2.77
Clinical variables	FEVER	3	8	12			
	LYMPHAD	6	10	13			
	PHOTO	10	7	18			
	LIVEDO	7	8	19			
	PERICA	7	2	7			
	PROTEINU	11	5	14			
	HEMATURI	8	5	8			
	INSUFF	6	3	3			
	DEPRESS	10	9	10			
	COOMBS	9	2	5			
	LYMPHCT	7	9	11			
	RVVT	5	7	13			

Arthritis & Rheumatology

	AntyB2Gly	5	5	11			
	SJOGREN	2	2	10			
	FIBROSIS	2	5	1			
	LFT_CUM	7	7	15			

Table S6. Mean values across clusters. The table shows the mean values of SLEDAI and cell proportions across patients in each cluster, standard deviations are shown in parentheses. * Marks no available information to perform the analysis.

	Cluster P1	**Cluster P2**	**Cluster P3**	**Cluster A1**	**Cluster A2**	**Cluster A3**
SLEDAI	6.45 (5.02)	6.44 (5.29)	6.65 (6.45)	3.29 (3.46)	2.30 (2.81)	2.80 (3.13)
Neutrophil percentage	64.06 (13.44)	65.62 (15.19)	65.13 (14.61)	49.41 (8.55)	48.16 (8.73)	47.71 (9.29)
Lymphocyte percentage	25.21 (11.42)	24.31 (12.26)	23.83 (11.62)	*	*	*

Supplementary figure legends

Figure S1. Cell proportion and SLEDAI correlation overview. A) and B) show the mean and standard deviations of neutrophil and lymphocyte percentages, respectively, for each SLEDAI value across the three clusters.

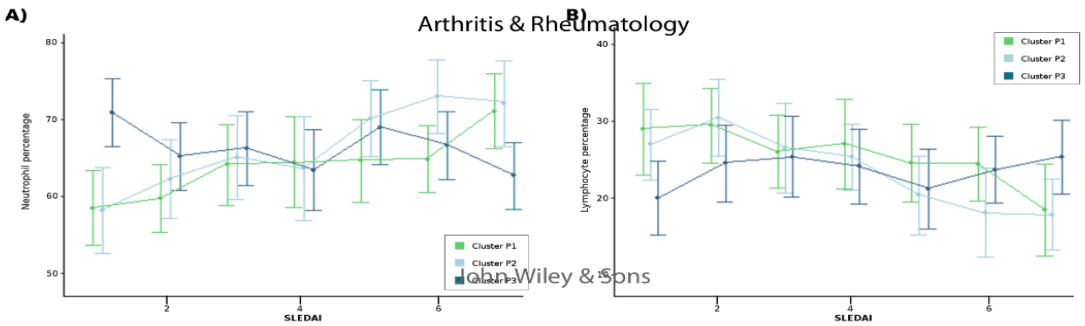


Figure S2. Quality control of cell proportion imputation. A) The picture shows the correlation between the real cellular proportions of pediatric patients and imputed cellular proportions in the same patients. B) Comparison of correlations between SLEDAI and real cell proportions in each patient across visits and correlations between SLEDAI and imputed cell proportions.

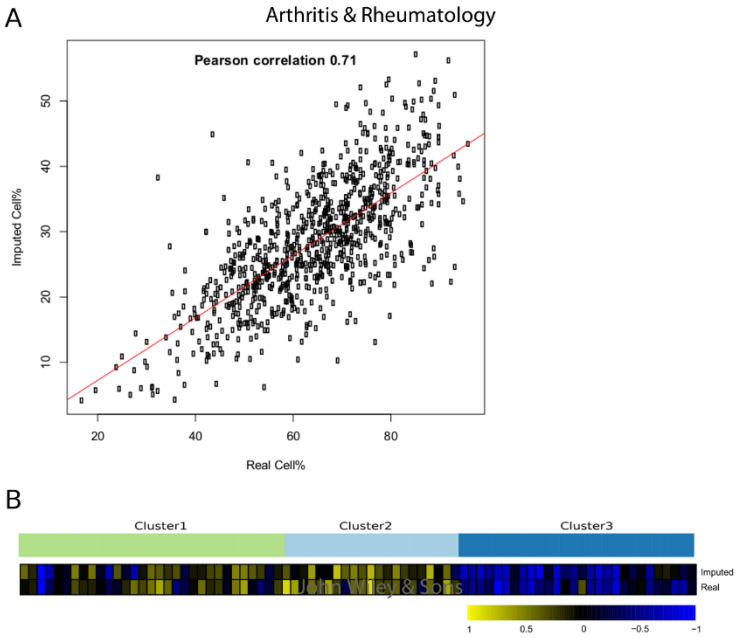
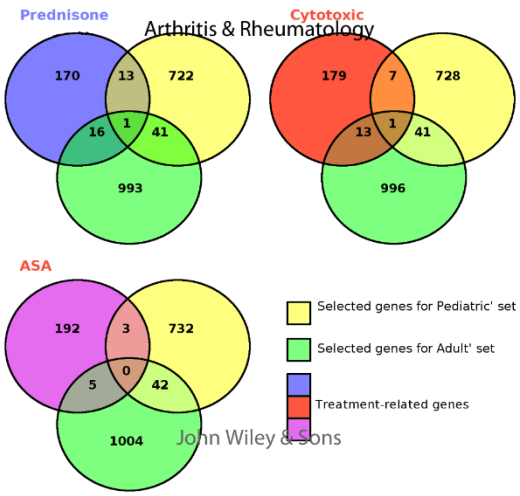
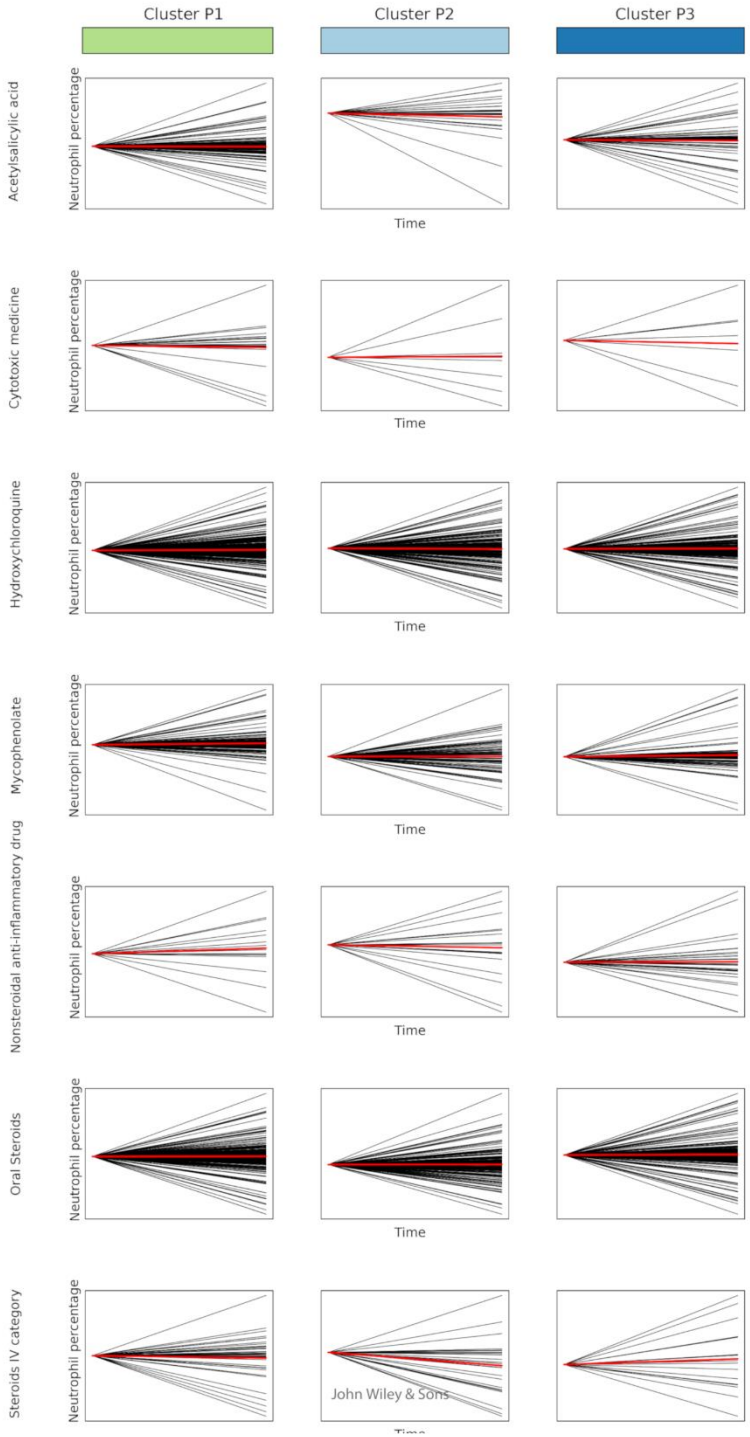


Figure S3. Treatment related genes. The picture shows the number of shared genes between genes selected to stratify in the two sets and the genes that best correlate with treatment doses. The treatment related genes were obtained in the same way as genes selected to stratify (see methods) but substituting the SLEDAI for the dose of each treatment. Only 2% of the genes were shared between the selected genes forming the clusters and those genes obtained with treatment doses.



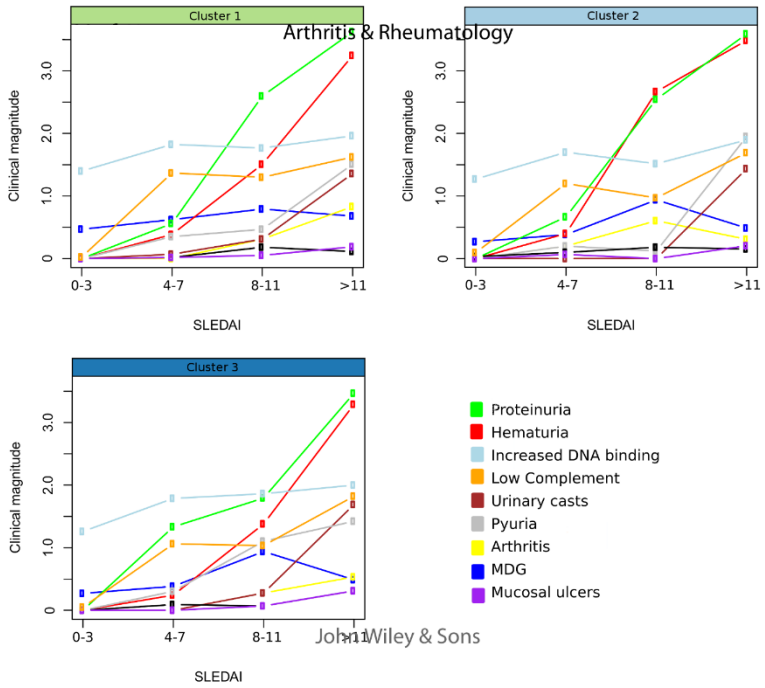
Arthritis & Rheumatology

Figure S4.Cell proportion modulation by treatments. The plot shows the tendencies of neutrophil proportions between each time point when a determinate treatment is applied and the next time point. The tendencies were calculated as the angle between percentage between the two points and the time difference. We normalized the time points in the plot to make comparable the tendencies of the different treatment applications.



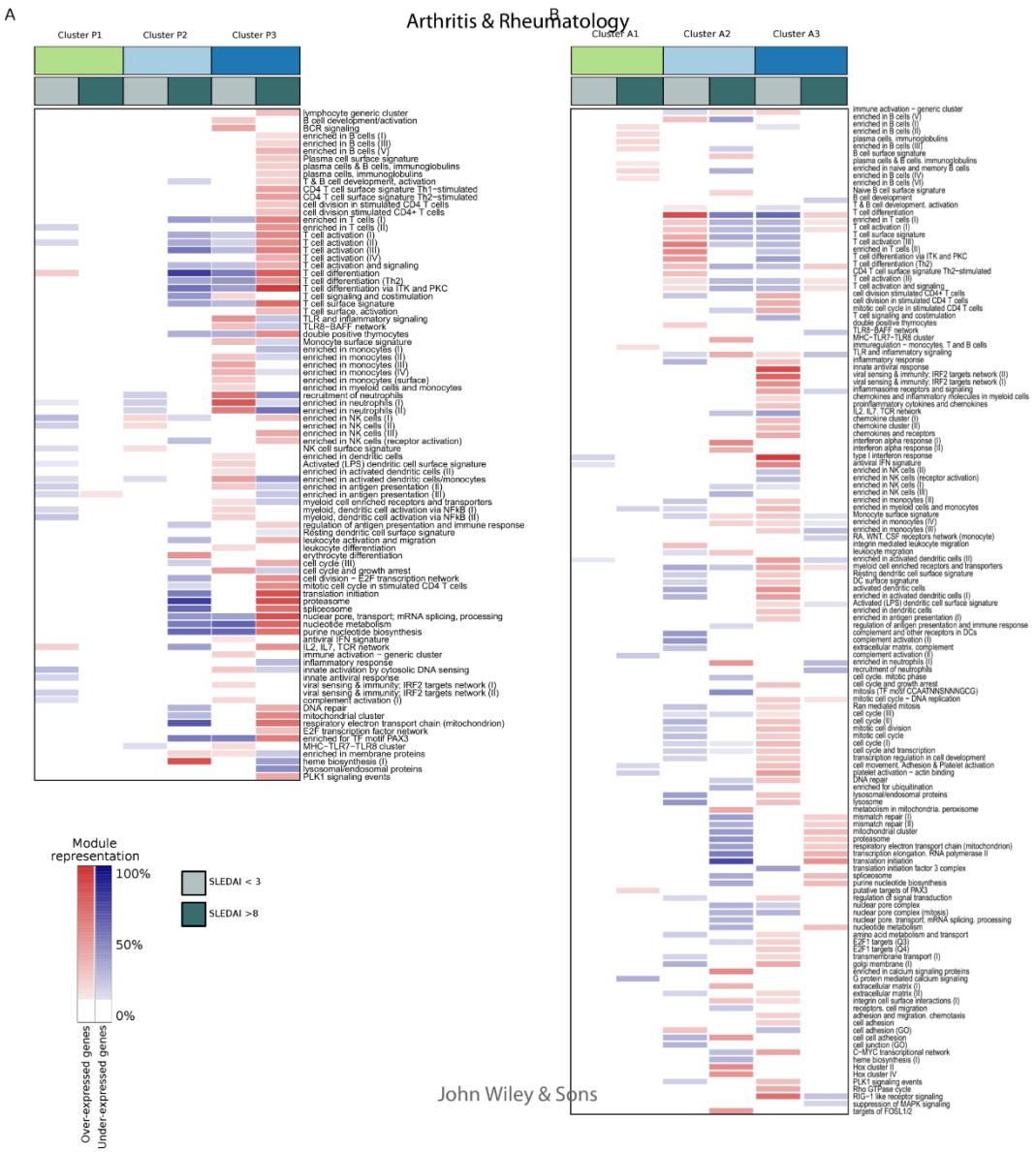
Arthritis & Rheumatology

Figure S5. SLEDAI decomposition. The plot represents the weight with which each component contributes to the SLEDAI score in each cluster. The pediatric clusters were divided into disease activity ranges.



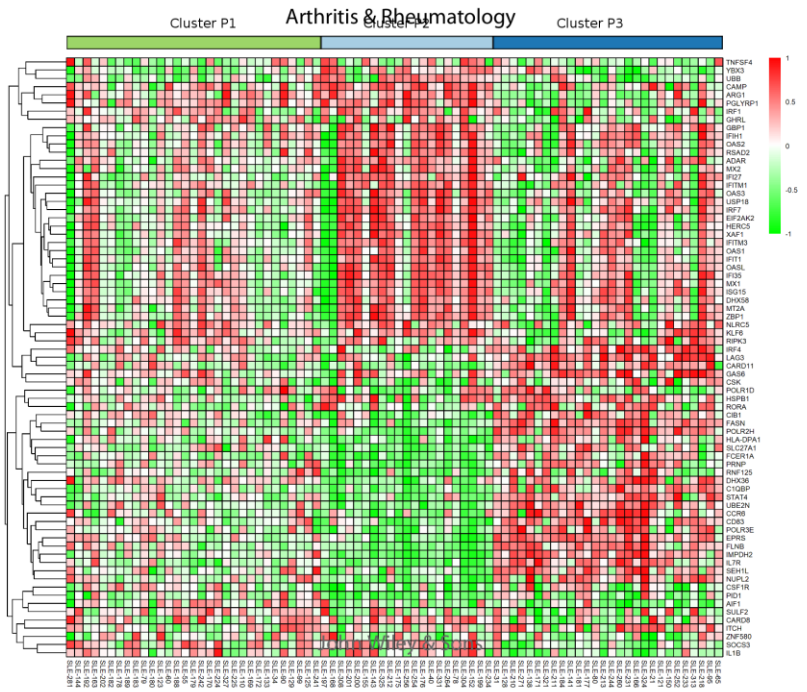
Arthritis & Rheumatology

Figure S6. All significant modular functions in the two sets. A) Significant modular functions for pediatric patient set. B) Significant modular functions for adult set.



Arthritis & Rheumatology

Figure S7. Heat map showing the genes from the functional cytokine signaling group negatively or positively correlated in each of the clusters.



R Packages Used in our analyses

mclust: (https://cran.r-project.org/web/packages/mclust/mclust.pdf).
ConsensusClusterPlus: (https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html).
clusteval: (https://cran.r-project.org/web/packages/clusteval/clusteval.pdf).
limma: (https://bioconductor.org/packages/release/bioc/html/limma.html).
tmod: (https://cran.r-project.org/web/packages/tmod/index.html).

6. CONCLUSIONES

En esta tesis, se ha abordado el problema de la autoinmunidad desde **dos perspectivas** diferentes. La primera se basa en la búsqueda de lo común, lo homogéneo, y consiste en identificar aquellos genes y rutas biológicas alteradas de forma consistente y conservadamente con respecto a individuos sanos, primero a través de diferentes enfermedades autoinmunes, mediante un meta-análisis integrando 3 enfermedades autoinmunes, y luego en LES específicamente, mediante la identificación de rutas patogénicas terapéuticas usando diferentes estudios. Una vez asumimos que hay mecanismos moleculares que se comportan de manera similarmente errónea dentro de estas patologías, la segunda perspectiva de abordaje se enfoca en la heterogeneidad, centrándonos exclusivamente en LES, pero siendo el trabajo extrapolable y aplicable al resto de enfermedades autoinmunes. Anteriormente se ha descrito el problema de la heterogeneidad multifactorial del LES y de cómo ésta puede reflejarse desde una sintomatología clínica diferente hasta una respuesta diferencial al tratamiento. Con este segundo enfoque, buscamos la identificación de mecanismos biológicos que permitan explicar la heterogeneidad dentro del LES, con el objetivo de agrupar a los pacientes en subgrupos más homogéneos desde el punto de vista molecular y funcional.

Estas dos perspectivas de trabajo nos han permitido tanto generar un punto de vista global de este tipo de patologías, como entrar en detalle en ciertos aspectos tanto moleculares como clínicos, generando un conocimiento amplio y variado en el campo, el cual nos brinda la oportunidad de entender y trabajar con los problemas aún persistentes en la autoinmunidad.

CONCLUSIONES

A continuación, se destacan de manera resumida las conclusiones más relevantes obtenidas en cada uno de los estudios publicados y recogidos en esta tesis.

1. El meta-análisis basado en datos de expresión génica nos reveló una firma de 371 genes diferencialmente expresados de forma consistente en LES, AR y SjS respecto a controles sanos, en muestras de sangre. 187 y 184 genes estaban sobre e infra expresados en las patologías, respectivamente. Estos genes trascienden más allá de una sola enfermedad, por lo que pueden jugar un papel fundamental en el desarrollo, mantenimiento o potenciación de la autoinmunidad.

2. Entre las rutas biológicas más sobre-expresadas, encontramos procesos de ciclo celular, apoptosis, señalización del sistema inmune mediado por citoquinas y la ruta del IFN I, mientras que la translación proteica y procesos metabólicos celulares son los dos mecanismos biológicos más infra-expresados.

3. Encontramos un total de 21 nuevos fármacos candidatos para tratar el LES, cuyos mecanismos de acción se asocian principalmente a inhibición de las vías conocidas como PI3K, mTOR, CDK e IGF. Observamos una estrecha relación entre todos los mecanismos de acción, los cuales actúan en la misma vía de señalización o en una vía de señalización dependiente, cuyo principal objetivo es la regulación de la apoptosis y la proliferación celular.

4. Los análisis de firmas derivadas del silenciamiento o estimulación de genes únicos resultaron en un conjunto de 140 genes como posibles dianas terapéuticas, entre los que encontramos genes relacionados con la ruta del PI3K, el IGF y la apoptosis. Los genes obtenidos concuerdan con los resultados derivados de fármacos.

5. Definimos 3 subgrupos claros de LES en base a cómo se comporta el perfil transcriptómico en relación con la actividad de la enfermedad. Estos 3 subgrupos fueron replicados en 2 cohortes independientes de LES.

6. Cada subgrupo muestra una serie de rutas biológicas las cuales se activan con los brotes de la enfermedad, siendo los subgrupos 2 y 3 prácticamente opuestos. Cuando aumenta la actividad de la enfermedad, las rutas de señalización del IFN I, así como la producción de citoquinas pro-inflamatorias se activa en el subgrupo 2 de LES, así como el estrés celular y algunos procesos de regulación positiva de la apoptosis, mientras que el subgrupo 3 sufre un incremento en procesos de ciclo y proliferación celular y una sobre activación de funciones relacionadas con linfocitos T. El grupo 1 es más heterogéneo desde el punto de vista celular, compartiendo gran parte de las funcionalidades con el grupo 2, pero también algunas con el grupo 3. Se podría decir que es un grupo de pacientes indefinidos.

7. Los subgrupos 1 y 2 están caracterizados por un aumento en los porcentajes de neutrófilos cuando la actividad de la enfermedad aumenta, mientras que en el grupo 3 aumentan los linfocitos.

8. Los subgrupos muestran fuertes diferencias desde el punto de vista clínico. Los grupos 1 y 2, los cuales están relacionados o dirigidos por neutrófilos, muestran un mayor número de casos de desarrollo de nefritis severa, mientras que el grupo 3, o grupo relacionado con linfocitos, muestra una mayor comorbidad con otras enfermedades autoinmunes como SjS, así como una asociación con daño hepático.

9. Por tanto, definimos 3 tipos de comportamientos de progresión del LES diferentes molecular, celular y clínicamente, los cuales reducen la heterogeneidad global de la

enfermedad desde el punto de vista transcriptómico, y que abren las puertas a una medicina más personalizada dentro de la enfermedad. Estos grupos, pueden ser usados para estudiar no sólo la etiología de la enfermedad, sino además para estudiar posibles tratamientos diferenciales más eficientes dentro de cada subgrupo.

7. NUEVAS PERSPECTIVAS

Los resultados de estos estudios nos abren las puertas a nuevas preguntas y nuevos trabajos con los que profundizar más tanto en la etiología de la autoinmunidad como en la medicina personalizada dentro de estas enfermedades. Una de las principales preguntas que surgen a partir del trabajo donde se estratifican a los pacientes en función de los patrones que dirigen la actividad de la enfermedad es: ¿Pueden estos grupos ser tratados terapéuticamente de forma diferente para obtener una mayor eficiencia y una mejor respuesta a los tratamientos?

Con el objetivo de responder a esta pregunta, o de dar unas bases que ayuden a responderla, desarrollamos el siguiente trabajo (actualmente en vías de publicación) titulado como "*Differential Treatments and Risk for Severe Nephritis Based on a Longitudinal Systemic Lupus Erythematosus Stratification*". En este trabajo, realizamos un análisis de reutilización de fármacos *in silico* usando CLUE para cada subgrupo de LES de forma específica, pero con el que extraemos no sólo los fármacos, sino además los mejores mecanismos de acción tratables en cada uno de los grupos. Entre los resultados más destacables observamos fuertes diferencias en los valores de similitud de los fármacos comúnmente usados en LES entre los subgrupos dirigidos por neutrófilos y el subgrupo dirigido por linfocitos, lo que nos hace hipotetizar que, de acuerdo a la capacidad de los fármacos para revertir la firma de expresión génica de cada grupo, podríamos obtener diferente respuesta a fármacos dependiendo del grupo al que pertenezca el paciente. De igual modo, al analizar los mejores fármacos candidatos, obtuvimos claras diferencias tanto a nivel de fármacos como de rutas biológicas terapéuticas. Tras este análisis, construimos un modelo logístico de clasificación basado en la variación del porcentaje

de neutrófilos con respecto al SLEDAI, el cual nos permite clasificar a un paciente dentro de LES dirigido por linfocitos o por neutrófilos, con el objetivo de poder incluir esta información para futuros trabajos donde se pruebe la eficiencia de los fármacos y se pueda verificar esta hipótesis más allá de los resultados obtenidos en el análisis *in silico*.

Uno de los problemas que encontramos en el estudio de la autoinmunidad es la falta o poca conexión entre la investigación científica y el entorno médico, algo que se ve reflejado en el uso de marcadores de actividad de la enfermedad como el SLEDAI. Este índice, sirve como referencia dentro de un paciente para medir de algún modo si el paciente presenta una sintomatología más o menos grave, pero no es explicativo de la sintomatología. Es decir, el SLEDAI no indica que tipo de afección clínica que padece el paciente, ni mucho menos cuales son los mecanismos moleculares alterados. Este hecho hace que un valor igual de SLEDAI no signifique lo mismo entre diferentes pacientes. Por tanto, la generación de un nuevo medidor de actividad de la enfermedad que refleje las afecciones que sufre el paciente desde el punto de vista molecular y clínico podría ser un avance importante en el campo, ya que daría más información al médico y permitiría una mejor decisión terapéutica.

Hasta ahora, nos hemos centrado tanto en la búsqueda de patrones comunes entre grupos de enfermedades como en la homogeneidad y la heterogeneidad dentro del LES. Otro trabajo futuro, además de estudiar la heterogeneidad y la estratificación de pacientes en el resto de patologías autoinmunes, podría centrarse en buscar las diferencias moleculares entre enfermedades, con el objetivo de encontrar las causas que provocan la aparición de un cierto fenotipo u otro.

8. PRODUCCIÓN CIENTÍFICA

En esta sección se hace un pequeño resumen de los artículos y otros trabajos publicados durante esta tesis, indicando el posicionamiento de las revistas en los que han sido publicados (Q: cuartil)

Artículos principales de la tesis:

- *Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis* (68). (Q2, factor de impacto: 4,269).
- *Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing* (69). (Q2, factor de impacto: 4,269).
- *Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression* (70). (Q1, factor de impacto: 7,871).

Otros trabajos originales o en colaboración:

- *ImaGEO: integrative gene expression meta-analysis from GEO database* (71). (Q1, factor de impacto: 5,481).
- *MetaGenyo: a web tool for meta-analysis of genetic association studies* (72). (Q1, factor de impacto: 2,213).

- *Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models* (73). (Q1, factor de impacto: 5,481).
- *In Silico Drug Design: Repurposing Techniques and Methodologies* (59) (Capitulo de libro).
- *Integrative analysis reveals a molecular stratification of systemic autoinmune diseases* (actualmente en proceso de publicación).
- *Analysis of transcription factor activity patterns in systemic lupus erythematosus* (actualmente en proceso de publicación).
- *Differential Treatments and Risk for Severe Nephritis Based on a Longitudinal Systemic Lupus Erythematosus Stratification (*actualmente en proceso de publicación*).*
- *A comprehensive and centralized database for exploring omics data in Systemic Autoimmune Diseases* (actualmente en proceso de publicación).
- *Methods and applications for gene expression meta-analysis* (artículo de revisión, actualmente en proceso de publicación)

9. INDICE DE FIGURAS DE LA TESIS

10. REFERENCIAS

1. McCullough KC, Summerfield A. Basic concepts of immune response and defense development. *ILAR J* 2005;46:230–240.

2. Belkaid Y, Hand T. Role of the Microbiota in Immunity and inflammation. *Cell* 2014;157:121–141.

3. Nicholson LB. The immune system. *Essays Biochem* 2016;60:275–301.

4. Smith DA, Germolec DR. Introduction to immunology and autoimmunity. *Environ Health Perspect* 1999;107:661–665.

5. Cooper GS, Stroehla BC. The epidemiology of autoimmune diseases. *Autoimmun Rev* 2003;2:119–125.

6. Di Battista M, Marcucci E, Elefante E, Tripoli A, Governato G, Zucchi D, et al. One year in review 2018: systemic lupus erythematosus. *Clin Exp Rheumatol* 2018;36:763–777.

7. Croia C, Bursi R, Sutera D, Petrelli F, Alunno A, Puxeddu I. One year in review 2019: pathogenesis of rheumatoid arthritis. *Clin Exp Rheumatol* 2019;37:347–357.

8. Romão VC, Talarico R, Scirè CA, Vieira A, Alexander T, Baldini C, et al. Sjögren's syndrome: state of the art on clinical practice guidelines. *RMD Open* 2018;4.

9. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinformatics* 2018;19:286–302.

10. Bumgarner R. DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol* 2013;0 22:Unit-22.1.

11. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63.

12. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* 2018;50:96.

13. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* 2013;41:D991-995.

14. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017;171:1437-1452.e17.

15. Waldron L, Riester M. Meta-Analysis in Gene Expression Studies. *Methods Mol Biol* 2016;1418:161–176.

16. Huan T, Esko T, Peters MJ, Pilling LC, Schramm K, Schurmann C, et al. A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. *PLOS Genetics* 2015;11:e1005035.

17. Magalhães JP de, Curado J, Church GM. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* 2009;25:875–881.

18. Chen R, Khatri P, Mazur PK, Polin M, Zheng Y, Vaka D, et al. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer Res* 2014;74:2892–2902.

19. Piras IS, Manchia M, Huentelman MJ, Pinna F, Zai CC, Kennedy JL, et al. Peripheral Biomarkers in Schizophrenia: A Meta-Analysis of Microarray Gene Expression Datasets. *Int J Neuropsychopharmacol* 2019;22:186–193.

20. Winkler JM, Fox HS. Transcriptome meta-analysis reveals a central role for sex steroids in the degeneration of hippocampal neurons in Alzheimer's disease. *BMC Syst Biol* 2013;7:51.

21. Song GG, Kim J-H, Seo YH, Choi SJ, Ji JD, Lee YH. Meta-analysis of differentially expressed genes in primary Sjogren's syndrome by using microarray. *Hum Immunol* 2014;75:98–104.

22. Arasappan D, Tong W, Mummaneni P, Fang H, Amur S. Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. *BMC Med* 2011;9:65.

23. Olsen N, Sokka T, Seehorn CL, Kraft B, Maas K, Moore J, et al. A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. *Ann Rheum Dis* 2004;63:1387–1392.

24. Tuller T, Atar S, Ruppin E, Gurevich M, Achiron A. Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. *Genes Immun* 2013;14:67–82.

25. Silva GL, Junta CM, Mello SS, Garcia PS, Rassi DM, Sakamoto-Hojo ET, et al. Profiling meta-analysis reveals primarily gene coexpression concordance between systemic lupus erythematosus and rheumatoid arthritis. *Ann N Y Acad Sci* 2007;1110:33–46.

26. Higgs BW, Liu Z, White B, Zhu W, White WI, Morehouse C, et al. Patients with systemic lupus erythematosus, myositis, rheumatoid arthritis and scleroderma share activation of a common type I interferon pathway. *Ann Rheum Dis* 2011;70:2029–2036.

27. Teruel M, Alarcón-Riquelme ME. Genetics of systemic lupus erythematosus and Sjögren's syndrome: an update. *Curr Opin Rheumatol* 2016;28:506–514.

28. Ibáñez K, Boullosa C, Tabarés-Seisdedos R, Baudot A, Valencia A. Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS Genet* 2014;10:e1004173.

29. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 2004;573:83–92.

30. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol* 2005;3:1171–1189.

31. Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature Protocols* 2015;10:823–844.

32. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? *Drug Discov Today* 2013;18:350–357.

33. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;3:673–683.

34. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Brief Bioinformatics* 2016;17:2–12.

35. Morales A, Gingell C, Collins M, Wicker PA, Osterloh IH. Clinical safety of oral sildenafil citrate (VIAGRA) in the treatment of erectile dysfunction. *Int J Impot Res* 1998;10:69–73; discussion 73-74.

36. Chong CR, Sullivan DJ. New uses for old drugs. *Nature* 2007;448:645–646.

37. King MD, Long T, Pfalmer DL, Andersen TL, McDougal OM. SPIDR: small-molecule peptide-influenced drug repurposing. *BMC Bioinformatics* 2018;19:138.

38. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science* 2008;321:263–266.

39. Duran-Frigola M, Aloy P. Recycling side-effects into clinical markers for drug repositioning. *Genome Med* 2012;4:3.

40. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;107:14621–14626.

41. Pan Y, Cheng T, Wang Y, Bryant SH. Pathway analysis for drug repositioning based on public database mining. *J Chem Inf Model* 2014;54:407–418.

42. Itadani H, Mizuarai S, Kotani H. Can Systems Biology Understand Pathway Activation? Gene Expression Signatures as Surrogate Markers for Understanding the Complexity of Pathway Activation. *Curr Genomics* 2008;9:349–360.

43. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* 2011;3:96ra77.

44. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. *Cell* 2000;102:109–126.

45. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313:1929–1935.

46. Jin G, Wong STC. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today* 2014;19:637–644.

47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–15550.

48. Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005;6:144.

49. Jia Z, Liu Y, Guan N, Bo X, Luo Z, Barnes MR. Cogena, a novel tool for co-expressed gene-set enrichment analysis, applied to drug repositioning and drug mode of action discovery. *BMC Genomics* 2016;17:414.

50. Vanhaelen Q, Mamoshina P, Aliper AM, Artemov A, Lezhnina K, Ozerov I, et al. Design of efficient computational workflows for in silico drug repurposing. *Drug Discov Today* 2017;22:210–222.

51. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci* 2010;31:115–123.

52. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. *Brief Bioinformatics* 2018;19:878–892.

53. Liu H, Song Y, Guan J, Luo L, Zhuang Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* 2016;17:539.

54. Chen X, Liu M-X, Yan G-Y. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;8:1970–1978.

55. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;8:573.

56. Nascimento ACA, Prudêncio RBC, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics* 2016;17:46.

57. Yang J, Li Z, Fan X, Cheng Y. Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization. *J Chem Inf Model* 2014;54:2562–2569.

58. Dai W, Liu X, Gao Y, Chen L, Song J, Chen D, et al. Matrix Factorization-Based Prediction of Novel Drug Indications by Integrating Genomic Space. *Comput Math Methods Med* 2015;2015:275045.

59. Anon. In Silico Drug Design - 1st Edition. Available at: https://www.elsevier.com/books/in-silico-drug-design/roy/978-0-12-816125-8. Accessed May 22, 2019.

60. Mohan C, Putterman C. Genetics and pathogenesis of systemic lupus erythematosus and lupus nephritis. *Nat Rev Nephrol* 2015;11:329–341.

61. Chambers SA, Rahman A, Isenberg DA. Treatment adherence and clinical outcome in systemic lupus erythematosus. *Rheumatology (Oxford)* 2007;46:895–898.

62. Sørlie T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur J Cancer* 2004;40:2667–2675.

63. Wang C, Machiraju R, Huang K. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods* 2014;67:304–312.

64. Pons-Estel GJ, Alarcón GS, Scofield L, Reinlib L, Cooper GS. Understanding the Epidemiology and Progression of Systemic Lupus Erythematosus. *Seminars in Arthritis and Rheumatism* 2010;39:257–268.

65. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis Rheum* 1992;35:630–640.

66. Manso PM, Vilar JA. *TSclust: Time Series Clustering Utilities.*; 2017. Available at: https://CRAN.R-project.org/package=TSclust. Accessed May 28, 2019.

67. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M, et al. Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* 2016;165:551–565.

68. 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;16:489.

69. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. *Arthritis Res Ther* 2017;19.

70. 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 & Rheumatology* 2018;70:2025–2035.

71. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, et al. ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics* 2019;35:880–882.

72. Martorell-Marugan J, Toro-Dominguez D, Alarcon-Riquelme ME, Carmona-Saez P. MetaGenyo: a web tool for meta-analysis of genetic association studies. *BMC Bioinformatics* 2017;18:563.

73. Carmona-Sáez P, Varela N, Luque MJ, Toro-Domínguez D, Martorell-Marugan J, Alarcón-Riquelme ME, et al. Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. *Bioinformatics* 2017;33:3691–3695.