CARACTERIZACIÓN DE LAS DIFERENCIAS DE SEXO EN ARTRITIS REUMATOIDE MEDIANTE EL METAANÁLISIS DE ESTUDIOS TRANSCRIPTÓMICOS

TRABAJO FINAL DE GRADO

Grado de Biotecnología

Curso académico 2020/2021

Universitat Politècnica de València

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

Centro de Investigación Príncipe Felipe

Autor: Edurne Urrutia Lafuente

Tutor académico: Dr. José Gadea Vacas

Cotutores externos: · Dr. Francisco García García

· Dña. Irene Pérez Díez

Unidad de Bioinformática y Bioestadística

PRINCIPE FELIPE CENTRO DE INVESTIGACION

València, julio 2021

Caracterización de las diferencias de sexo en artritis reumatoide mediante el metaanálisis de estudios transcriptómicos

Resumen

Las enfermedades autoinmunes, entre ellas la artritis reumatoide, presentan una mayor prevalencia en mujeres que en hombres. Según la literatura, el sistema inmunitario actúa de forma diferencial entre sexos. Aunque las hormonas sexuales desempeñan un papel fundamental, no se conocen por completo las diferencias en los mecanismos moleculares. Una mejor comprensión de estos mecanismos proporcionaría una mayor precisión en el diagnóstico y la personalización de tratamientos en función del sexo.

El objetivo de este Trabajo Final de Grado es, por lo tanto, la identificación y comprensión de los mecanismos moleculares diferenciales que subyacen a la artritis reumatoide por sexos, con el fin de mejorar el diagnóstico y la selección de los tratamientos para cada grupo de pacientes.

Para ello, se realizó, en primer lugar, una revisión sistemática de estudios de artritis reumatoide con datos de microarrays de expresión génica o RNA-seq de individuos sanos y pacientes. Se recopilaron 5 *datasets* con 570 muestras del repositorio GEO. A continuación, se analizaron individualmente los cinco estudios seleccionados. Este paso incluyó un análisis exploratorio, un análisis de expresión diferencial y un análisis de enriquecimiento funcional de procesos biológicos de la Gene Ontology (GO) y de rutas KEGG. Finalmente, se llevó a cabo un metaanálisis de genes mediante un modelo de efectos aleatorios, seguido de un enriquecimiento funcional de los resultados globales. De este modo, se identificaron 10 biomarcadores, 90 procesos biológicos y 11 rutas KEGG específicos de sexo estadísticamente significativos. Además, los resultados del enriquecimiento se sintetizaron empleando la herramienta REVIGO y los términos GO slim. Entre otras funciones, los resultados sugieren una actividad aumentada del sistema inmune en la mujer enferma, ligada a una mayor presencia de precursores energéticos; así como una sobrerrepresentación de términos relacionados con la síntesis proteica y la regulación negativa de los osteoblastos en hombres con artritis reumatoide.

En definitiva, este estudio ha permitido caracterizar las diferencias de sexo en pacientes con artritis reumatoide, identificando biomarcadores y funciones sobrerrepresentados en cada sexo. Estos hallazgos podrían conducir al descubrimiento de nuevas dianas terapéuticas y proporcionar la base para futuras investigaciones dirigidas a la medicina de precisión en la artritis reumatoide.

Palabras clave: Enfermedades autoinmunes; Artritis reumatoide; Diferencias de sexo; Expresión génica; Metaanálisis; Medicina personalizada.

Characterising sex differences in rheumatoid arthritis by meta-analysis of transcriptomic studies

Abstract

Autoimmune diseases, including rheumatoid arthritis (RA), are more prevalent in women than in men. According to the literature, the immune system acts differently between the sexes. Although sex hormones play a key role, the differences in the molecular mechanisms are not fully understood. A better understanding of these mechanisms could provide greater diagnostic precision and personalised treatments by sex. Therefore, the aim of this Final Degree Project was to identify and understand the molecular mechanisms underlying sex-based differences in rheumatoid arthritis, to improve the diagnosis and the selection of treatment for each group of patients.

This goal was achieved in three steps. First, we conducted a systematic review of rheumatoid arthritis studies with gene expression microarray or RNA-seq data from healthy individuals and patients. Five datasets that included 570 samples were retrieved from the Gene Expression Omnibus (GEO) repository. Second, the selected studies were individually analysed. This step included an exploratory analysis, a differential expression analysis, and a functional enrichment analysis of biological processes from the Gene Ontology (GO) and KEGG pathways. Finally, we performed a gene meta-analysis using a random effects model, followed by a functional enrichment of the overall results. Thus, our analysis revealed 10 biomarkers, 90 biological processes and 11 KEGG pathways as significantly sex-dependent. In addition, the results were synthesized using the REVIGO tool and GO slim terms. Among other functions, the outcomes from the enrichment suggest an increased activity of the immune system and a greater presence of energy precursors in female RA patients. Male RA patients, instead, showed an overrepresentation of terms related to protein biosynthesis and negative regulation of osteoblasts.

In conclusion, we were able to characterise sex differences in patients with rheumatoid arthritis, identifying biomarkers and functions overrepresented in each sex. Our findings could lead to the discovery of new therapeutic targets and be the basis of future research devoted to sex-specific precision medicine in rheumatoid arthritis.

Key words: Autoimmune diseases; Rheumatoid arthritis; Sex differences; Gene expression; Meta-analysis; Personalized medicine.

– Alumna: Dña. Edurne Urrutia Lafuente
– Tutor Académico: Prof. Dr. José Gadea Vacas
– Cotutores externos: Dr. Francisco García García y Dña. Irene Pérez Díez

Valencia, julio de 2021

– Tipo de licencia de autorización de acceso y difusión del TFG:

AGRADECIMIENTOS

En primer lugar, quisiera agradecer a Pepe Gadea, mi tutor UPV, por aconsejarme y ayudarme a orientarme en mi TFG.

A Paco García e Irene Pérez, por haberme hecho un hueco en su grupo y dedicarme su tiempo, por todo lo que me han enseñado, pero sobre todo por su buen humor y alegría contagiosa a lo largo de todo el proceso. Muchas gracias por ayudarme a disfrutar de este trabajo.

A Javier Forment, por conseguir que me picara el gusanillo de la bioinformática.

A Ro, Ester, Andrea, Marina, Alicia y Elvira, porque estos cuatro años han sido mucho mejores gracias a vosotras, porque las tartas de la Tarongería sirven para celebrar y para consolar, y porque sé que los próximos años seguiremos alegrándonos juntas a pesar de que por ahora nos separemos.

A mis padres y a Begoña, por acompañarme y apoyarme siempre, tengo mucha suerte de teneros a mi lado.

Y en especial al Doc, *teacher* y *coach* a tiempo parcial, y mi mejor compañero a tiempo completo a pesar de la distancia.

Muchísimas gracias a todos por acompañarme en este camino de aprendizaje en tantos sentidos.

ÍNDICE GENERAL

Índice de figuras

Índice de tablas

Acrónimos/Abreviaturas

ACPA anticuerpos antiproteínas citrulinadas

ADN ácido desoxirribonucleico

AINE antiinflamatorio no esteroideo

AR artritis reumatoide

ARN ácido ribonucleico

BH Benjamini-Hochberg

BP procesos biológicos

C1orf162 *Chromosome 1 open reading frame 162*

DL DerSimonian y Laird

EXOC3L2 *Exocyst complex component 3 like 2*

FAIR *Findability, Accessibility, Interoperability, and Reuse*

FAME fármaco antirreumático modificador de la enfermedad

FDR *False Discovery Rate*

FR factor reumatoide

GEO *Gene Expression Omnibus*

GNAI1 *G protein subunit alpha i1*

GO *Gene Ontology*

GSEA *Gene Set Enrichment Analysis*

HLA antígeno leucocitario humano

KEGG *Kyoto Encyclopedia of Genes and Genomes*

LINC00482 *Long intergenic non-protein coding RNA 482*

logFC logaritmo de la magnitud de cambio

LOR logaritmo de los *odds ratio*

NA no disponible

PCA análisis de componentes principales

RNA-seq secuenciación de ARN

SLE lupus eritematoso sistémico

TMM *Trimmed Mean of M-values*

TPRG1 *Tumor protein p63 regulated 1*

1. INTRODUCCIÓN

1.1. ENFERMEDADES AUTOINMUNES

Las enfermedades autoinmunes son patologías debidas a una excesiva actividad del sistema inmune que, en respuesta a un desencadenante, comienza a producir anticuerpos que dañan los tejidos sanos del propio organismo. Aunque históricamente se consideraron enfermedades raras, se ha demostrado la existencia de en torno a 100 enfermedades autoinmunes que afectan al 3-5% de la población, con una mayor prevalencia en mujeres. Esta susceptibilidad a las enfermedades autoinmunes en mujeres se explica, al menos en parte, por una respuesta inmune más fuerte. A pesar de la evidencia de esta diferencia de sexos, la investigación biomédica y la práctica clínica no consideran la variable sexo en el manejo de estas enfermedades, y los mecanismos que subyacen a estas diferencias continúan sin conocerse por completo (Moulton, 2018; Klein y Flanagan, 2016; Wang et al., 2015; Fish, 2008).

Las enfermedades autoinmunes son enfermedades complejas, resultado de interacciones entre factores genéticos y ambientales que llevan a una desregulación del sistema inmune y finalmente a la pérdida de tolerancia a los antígenos propios. Esta pérdida de tolerancia puede ser específica de un órgano, como la diabetes tipo I, o sistémica, como la artritis reumatoide, y conlleva la aparición de autoanticuerpos y/o células autorreactivas (Stafford et al., 2020; Banchereau et al., 2017).

La tolerancia, por lo tanto, es un proceso clave en el mantenimiento de la homeostasis del sistema inmune. Consta de una tolerancia central, que tiene lugar en el timo y en la médula ósea, y una tolerancia periférica o selección secundaria, que permite eliminar las células T y B autorreactivas. A pesar de ello, un cierto número de linfocitos autorreactivos puede filtrarse a la periferia. No obstante, es importante señalar que la existencia de estos linfocitos T y B autorreactivos y su capacidad para producir autoanticuerpos no conduce necesariamente al desarrollo de la patología. Esta condición transitoria se denomina autoinmunidad fisiológica (Avrameas y Selmi, 2013; Salinas et al., 2013). Por el contrario, cuando se pierde la tolerancia inmunitaria y los autoanticuerpos y los linfocitos autorreactivos intervienen en el proceso inflamatorio, se desarrolla la autoinmunidad clásica o patológica, que provoca finalmente daños en los tejidos (Wang et al., 2015). Ejemplo de esta autoinmunidad patológica es la artritis reumatoide (AR).

1.2. ARTRITIS REUMATOIDE

La AR es una enfermedad autoinmune crónica que se caracteriza por la inflamación de las articulaciones y la destrucción ósea. Además, en la mayoría de los casos se detectan autoanticuerpos frente a la inmunoglobulina G (factor reumatoide, FR) y frente a proteínas citrulinadas (ACPAs). Estos anticuerpos se adhieren al revestimiento de las articulaciones, de modo que las células del sistema inmunitario atacan dichas articulaciones, causando inflamación, hinchazón y dolor. Además, tiende a ser simétrica y afectar específicamente a determinadas articulaciones de las manos y los pies. En ausencia de tratamiento, puede derivar en daño permanente de las articulaciones y discapacidad irreversible (Smolen et al., 2018; Frank-Bertoncelj et al., 2017).

Se trata de una enfermedad compleja que puede desencadenarse por factores ambientales en individuos genéticamente susceptibles. Asimismo, es heterogénea, con presentación clínica y mecanismos patogénicos variables entre pacientes con el mismo diagnóstico. De hecho, aunque los

autoanticuerpos son una característica importante de la AR, los pacientes se pueden clasificar en dos grupos principales en función de la presencia (AR seropositiva) o ausencia de dichos autoanticuerpos (AR seronegativa). La AR seropositiva constituye dos tercios de los casos de AR y suele tener una progresión más severa (Malmström et al., 2016).

1.2.1. Epidemiología

La AR afecta a la población a nivel global y es tres veces más frecuente en mujeres que en hombres. La mayoría de los estudios epidemiológicos se han realizado en países occidentales, y muestran una prevalencia del 0,5-1,1% y una incidencia de entre 20 y 50 casos por cada 100.000 habitantes. Sin embargo, parecen diferir entre etnias y regiones geográficas, y además, la falta de uniformidad en la clasificación e inicio de la enfermedad dificultan su estimación (Tobón et al., 2010). Respecto a la mortalidad, ésta es mayor entre los pacientes de AR que en la población general, y la esperanza de vida, de 3 a 10 años menor. Además, los datos de supervivencia no han mejorado en las últimas décadas, por lo que la diferencia respecto a la población general ha aumentado (Gonzalez et al., 2007).

Por lo tanto, se trata de una enfermedad con un elevado impacto sobre el paciente y el sistema sanitario, debido a su elevada prevalencia, cronicidad con episodios de brotes y remisiones y su potencial para generar discapacidad. El dolor crónico y la discapacidad repercuten negativamente en la calidad de vida del paciente y de su entorno, afectando a su bienestar físico y emocional, a sus relaciones y a su vida laboral. Además, genera un importante consumo de recursos, debido al uso de fármacos de elevado coste (García De Yébenes y Loza, 2018). Por ello, disponer de biomarcadores que permitan estratificar a los pacientes y optar por el tratamiento idóneo sería clave para minimizar su impacto.

1.2.1.1. FACTORES DE RIESGO

La AR es una enfermedad multifactorial. Aunque tiene un fuerte componente genético, se ha demostrado que existen otros factores de riesgo con un papel importante en la susceptibilidad (Smolen et al., 2018):

Genética. La heredabilidad de la AR se estima en torno al 60% (MacGregor et al., 2000). En total, se han identificado unos 100 loci asociados con susceptibilidad a la AR. Entre los loci que muestran una mayor asociación, destacan los específicos del antígeno leucocitario humano (HLA) de clase II. En concreto, los alelos asociados con la AR comparten una secuencia de aminoácidos denominada 'epítopo compartido' (Gregersen et al., 1987). Sin embargo, también se han identificado otros loci de riesgo con asociaciones más débiles, la mayoría relacionados con rutas inmunitarias e inflamatorias (Okada et al., 2014). Estas variantes suelen situarse en regiones *enhancer* que permiten la regulación de genes distantes (Martin et al., 2015). Además, se ha observado cierto efecto acumulativo cuando están presentes simultáneamente varios alelos de riesgo. La comprensión de esta compleja regulación contribuirá a la identificación de las vías clave que conducen a la enfermedad y permitirá la estratificación de la población con AR en grupos basados en la ruta causal (Smolen et al., 2018).

Sexo. La mayoría de enfermedades autoinmunes muestran un desequilibrio en la prevalencia entre hombres y mujeres. En el caso de la AR, las mujeres son 2-3 veces más propensas a desarrollarla que los hombres (Platzer et al., 2019). Esta ratio disminuye con la edad, lo que sugiere la influencia de las hormonas femeninas (Kim y Kim, 2020). Además, en las mujeres la enfermedad suele presentarse hacia la mediana edad o en el momento de la menopausia. Los hombres, por el

contrario, presentan un inicio más tardío de la enfermedad, son más propensos a ser positivos al FR y tienen títulos más altos de ACPAs. También tienen una mayor probabilidad de lograr la remisión. Estas diferencias se atribuyen, en parte, a los efectos estimulantes de los estrógenos sobre el sistema inmunitario; sin embargo, el papel de los factores hormonales en el desarrollo de la AR continúa sin ser completamente conocido (Maynard et al., 2020).

Epigenética. Tanto la metilación del ácido desoxirribonucleico (ADN) en la región del HLA como la acetilación de las histonas podrían tener un papel en el desarrollo de la AR, ya que proporcionan un mecanismo a través del cual los factores ambientales pueden inducir cambios en la actividad celular. Un ejemplo de ello sería la metilación diferencial en la región del HLA entre fumadores y no fumadores (Meng et al., 2017; Liu et al., 2013).

Tabaquismo. Este hábito aumenta el riesgo de AR, duplicándose el riesgo entre fumadores con más de 20 años de consumo en comparación con los no fumadores. El aumento del riesgo podría estar mediado por modificaciones epigenéticas, ya que el tabaquismo se ha asociado con la hipometilación de ciertas regiones del ADN, mientras que el tratamiento con fármacos antirreumáticos modificadores de la enfermedad (FAMEs) induce la hipermetilación de dichas regiones (Smolen et al., 2018; Svendsen et al., 2016).

Inhalación de polvo. La exposición a la sílice y al polvo textil parecen ser factores de riesgo ambiental para la artritis reumatoide (Stolt et al., 2005).

Microbiota. La enfermedad periodontal, mediada por la microbiota oral, también se asocia a un mayor riesgo de desarrollar artritis reumatoide (Hajishengallis, 2015). Asimismo, la microbiota intestinal puede desempeñar un papel importante, ya que los individuos con AR muestran una diversidad disminuida en comparación con la población general, junto con una abundancia de ciertos taxones bacterianos poco frecuentes (Chen et al., 2016).

Otros. Factores modificables del estilo de vida como la obesidad también parecen estar relacionados con un moderado aumento del riesgo de padecer artritis reumatoide, especialmente en hombres (Ljung y Rantapää-Dahlqvist, 2016).

En definitiva, ninguno de estos factores se considera como el responsable único de la patología. Posiblemente no exista una única causa, vía de progresión, ni aproximación terapéutica. Esto podría implicar la existencia de subclases de AR que dependan de los factores de riesgo anteriormente citados, por lo que esta enfermedad se beneficiaría ampliamente de la medicina de precisión (Platzer et al., 2019).

1.2.2. Terapia

La estrategia actual de tratamiento de la AR consiste en el seguimiento estricto de la actividad de la enfermedad y el cambio de terapia si no se alcanza un objetivo determinado, siendo la remisión el objetivo final. Así, en fases iniciales se logra normalizar la función física, mientras que en la fase establecida se aspira a maximizarla. Además, este enfoque evita la aparición de daños o, si ya existe destrucción articular, su progresión (Smolen et al., 2018).

Los principales tipos de medicamentos para la artritis reumatoide incluyen los fármacos antirreumáticos modificadores de la enfermedad (FAMEs), los glucocorticoides, los antiinflamatorios no esteroideos (AINEs) y los analgésicos.

1. INTRODUCCIÓN

Los FAMEs interfieren en el proceso inflamatorio, y se puede distinguir entre los sintéticos (pequeños fármacos químicos) y los biológicos (anticuerpos monoclonales o construcciones de receptores). Estos últimos, como por ejemplo los inhibidores del factor de necrosis tumoral (TNF), tienen como diana proteínas solubles y asociadas a la membrana celular. Los FAME sintéticos, a su vez, pueden dividirse en convencionales (como el metotrexato, fundamental en el tratamiento actual), y dirigidos, que tienen como diana moléculas específicas intracelulares, como los inhibidores de la janus kinasa (Burmester y Pope, 2017).

Los glucocorticoides también poseen actividad modificadora de la enfermedad, pero sus efectos adversos impiden su uso a largo plazo. Sin embargo, dada su rápida actividad antiinflamatoria, pueden administrarse durante un periodo limitado junto con los FAMEs convencionales al inicio del tratamiento (Burmester y Pope, 2017).

Los agentes sintomáticos, como los analgésicos o los AINEs, mejoran los signos y síntomas, pero no modifican el proceso subyacente, por lo que no interfieren en los mecanismos que conducen al daño articular. Su función consiste en aliviar el dolor y la hinchazón, mediante la inhibición de la síntesis de prostaglandinas (Smolen et al., 2018).

En definitiva, en las dos últimas décadas ha habido una gran mejora de la prognosis de la artritis reumatoide. Esto se debe a diversos motivos: por un lado, la optimización del uso del inmunosupresor metotrexato, principal fármaco en el tratamiento de la AR, en combinación con las nuevas terapias biológicas. Además, la mejor comprensión de la patogénesis de la AR mediante el reconocimiento de citoquinas clave ha permitido el desarrollo de fármacos dirigidos. También el control estricto de los síntomas clínicos y una rápida adaptación del tratamiento han contribuido a esta mejora. Por otro lado, la identificación de nuevos autoanticuerpos ha mejorado la precisión del diagnóstico, y los nuevos criterios de clasificación desarrollados facilitan el reconocimiento y el estudio de la enfermedad en su fase inicial, facilitando el diagnóstico precoz y el inicio temprano del tratamiento (Smolen et al., 2018).

La combinación de estos avances ha mejorado considerablemente los resultados del tratamiento en la mayoría de los pacientes. Aun así, no todos logran la remisión de la enfermedad, y el coste del tratamiento ha aumentado enormemente. Por ello, la identificación de biomarcadores que permitan no sólo caracterizar la severidad de la enfermedad, sino también personalizar el tratamiento, véase en función del sexo, es fundamental de cara a una medicina de precisión que posibilite que todos los pacientes alcancen la remisión (Smolen et al., 2018; Burmester y Pope, 2017).

1.3. TECNOLOGÍAS DE ALTO RENDIMIENTO Y DATOS ÓMICOS

El desarrollo exponencial de las tecnologías de alto rendimiento en los últimos años ha permitido el estudio de los procesos que se enmarcan en el dogma central de la biología molecular (Figura 1), dando lugar a las denominadas ómicas: genómica, transcriptómica o proteómica, entre otras. Esto, a su vez, ha generado una cantidad ingente de datos con un gran potencial en ámbitos como la medicina de precisión; aunque también con limitaciones, como su manejo bioinformático y almacenamiento. Por ello, ha sido necesario generar repositorios de datos, como *Gene Expression Omnibus* (GEO) (Barrett et al., 2013) y ArrayExpress (Athar et al., 2019), que permitan su gestión. Estos recursos son actualizados y corregidos periódicamente (Lightbody et al., 2019; Manzoni et al., 2018; Reuter et al., 2015). En concreto, este estudio se centrará en la transcriptómica.

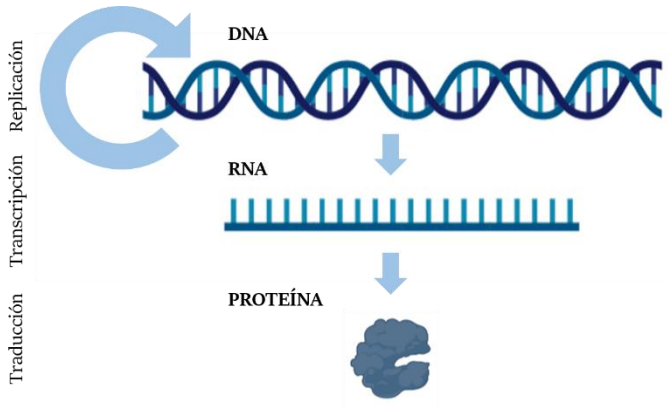

Figura 1 | Dogma central de la biología molecular. Muestra el flujo de la información genética: replicación del ácido desoxirribonucleico (DNA), transcripción a ácido ribonucleico (RNA) y traducción a proteína.

1.3.1. Transcriptómica

La transcriptómica es la tecnología ómica empleada para el estudio del transcriptoma de un organismo, es decir, la suma de la totalidad de sus transcritos, tanto codificantes como no codificantes, presentes en un momento dado en una célula. La cuantificación de la expresión génica en diferentes tejidos, condiciones o puntos temporales proporciona información sobre cómo se regulan los genes y ayuda a inferir las funciones de genes no anotados. Además, permite estudiar cómo cambia la expresión en distintos órganos y ha sido fundamental para comprender las enfermedades. Actualmente existen dos técnicas fundamentales para su análisis: los microarrays y la secuenciación de ácido ribonucleico o ARN (RNA-seq) (Lowe et al., 2017).

Los microarrays consisten en un conjunto de sondas (oligómeros de nucleótidos cortos) fijadas sobre un soporte sólido en una determinada posición. Tras el marcaje de los transcritos con fluorescencia, se hibridan con las sondas del microarray. Finalmente, la intensidad de la fluorescencia en cada *spot* permite cuantificar la abundancia del transcrito. Esta técnica, publicada por primera vez en 1995, permitió la detección de miles de transcritos simultáneamente a un coste muy reducido. Sin embargo, requiere el conocimiento previo del organismo de interés (Lowe et al., 2017).

El RNA-seq, por su parte, combina la secuenciación de alto rendimiento con métodos computacionales para cuantificar los transcritos a partir de sus conteos. Para ello, requiere la reconstrucción computacional del transcrito original alineando las lecturas con un genoma de referencia o entre sí (ensamblaje *de novo*). Al no partir de un grupo definido de sondas, permite la identificación de nuevos transcritos, la detección de isoformas y el análisis de otros tipos de ARN, como los microARN (miRNA). Por ello, desde las primeras descripciones en 2006, ha avanzado rápidamente, convirtiéndose en la técnica transcriptómica dominante. Sin embargo, requiere de un complejo procesamiento que incluye el enriquecimiento de los transcritos, la fragmentación, la amplificación y la elección del tipo de secuenciación (*single* o *paired-end*, específica de hebra o no específica) (Lowe et al., 2017; Conesa et al., 2016).

1.3.2. Análisis transcriptómicos

El progreso de las tecnologías de alto rendimiento ha traído a la par el desarrollo de numerosos métodos computacionales para el análisis de datos ómicos, que varían en función del objetivo del estudio en cuestión y del tipo de datos a analizar (Nekrutenko y Taylor, 2012; Allison et al., 2006; Cannata et al., 2005).

En el caso de la transcriptómica, el preprocesado de los datos es un paso fundamental en el análisis, y difiere según la plataforma empleada. Como resultado, se obtiene una matriz de expresión normalizada que será el punto de partida para las principales aplicaciones de la transcriptómica: la predicción de clases (clasificación de las muestras en función de su perfil de expresión en grupos ya conocidos mediante un modelo matemático), el descubrimiento de clases (identificación de nuevos subgrupos con perfiles de expresión similares) o la comparación de clases (identificación de genes diferencialmente expresados entre grupos) (Tarca et al., 2006). Este estudio se centrará en esta última.

1.3.3. Caracterización funcional

El análisis de expresión diferencial proporciona como resultado una lista de genes individuales ordenados en función de su expresión diferencial entre grupos. Extraer un significado de esta lista y detectar así procesos biológicos supone un desafío.

En este sentido, se sabe que las funciones celulares son llevadas a cabo por módulos integrados por moléculas que interaccionan entre sí. Las funciones que desempeñan dichos módulos se han representado de diferentes maneras, siendo la Ontología de Genes (GO) y la *Kyoto Encyclopedia of Genes and Genomes* (KEGG) las anotaciones funcionales más populares. La Ontología de Genes proporciona información sobre la función de los genes y sus productos, y se divide en tres ontologías: procesos biológicos, funciones moleculares y componentes celulares (Carbon et al., 2021; Ashburner et al., 2000). La KEGG, por su parte, está compuesta por las bases de datos PATHWAY (información funcional de moléculas que interactúan), GENES (información genómica) y LIGAND (compuestos químicos y enzimas) (Kanehisa et al., 2021; Kanehisa y Goto, 2000). A efectos prácticos, los módulos funcionales se definen como conjuntos de genes que comparten anotaciones funcionales extraídas de estos repositorios.

El método Gene Set Enrichment Analysis (GSEA) permite enfocarse en dichos grupos de genes que comparten una función biológica, regulación o locus común, y determinar si tienden a sobreexpresarse o subexpresarse coordinadamente (Subramanian et al., 2005). Este método puede detectar los módulos incluso si la expresión de sus componentes génicos no difiere significativamente al analizarlos de forma individual (Montaner y Dopazo, 2010). En este estudio se realizó el enriquecimiento funcional tanto con términos GO como con rutas KEGG.

1.3.4. Métodos de metaanálisis

El metaanálisis consiste en una síntesis sistemática y cuantitativa de la literatura apoyada por métodos estadísticos cuyo objetivo consiste en agregar y contrastar los hallazgos de múltiples estudios relacionados, y analizar las causas de la variación entre estudios (Gurevitch et al., 2018). De este modo, permite mejorar el poder estadístico de los resultados, y aporta una visión más general y completa que los estudios individuales, ayudando así a resolver resultados aparentemente contradictorios al estimar un efecto medio ponderado en todo el grupo de estudios analizados.

Para ello, los resultados relevantes de cada estudio se cuantifican y expresan en una escala común, de modo que los valores resultantes pueden compararse. Se pueden emplear diferentes medidas de resultados, como las *odds* y *risk ratios*, diferencias de media estandarizadas, o ratios logarítmicas de respuesta, por lo que se utiliza genéricamente el término "*effect size*" o tamaño del efecto (Gurevitch et al., 2018; Hornik, 2009).

A continuación, estas medidas del efecto se introducen en un modelo estadístico, con el objetivo de evaluar los efectos globales y la heterogeneidad de los resultados. Estos modelos se

basan en la asunción de un efecto común ("efectos fijos") o de "efectos aleatorios". El modelo de efectos fijos asume que la variación entre estudios se debe a la varianza intra-estudio (muestreo). Sin embargo, los estudios combinados en un metaanálisis suelen tener diferencias en su diseño y realización que pueden dar lugar a resultados heterogéneos, más variables de lo que se supone en el modelo de efectos fijos. Los modelos de efectos aleatorios tienen en cuenta estas diferencias, por lo que incluyen un parámetro de varianza de la heterogeneidad entre estudios (Langan et al., 2019; Gurevitch et al., 2018).

Entre los diferentes métodos empleados para estimar la varianza de la heterogeneidad en los modelos aleatorios, el método DerSimonian-Laird se considera la aproximación estándar en investigación clínica y médica, y es especialmente útil para proporcionar una estimación del efecto global y caracterizar la heterogeneidad de los efectos en una serie de estudios (DerSimonian y Laird, 2015). Por ello, éste será el método empleado en este análisis.

2. OBJETIVOS

La AR es una enfermedad que muestra diferencias clínicas y epidemiológicas entre sexos. Sin embargo, no se conocen los mecanismos que subyacen a estas diferencias. Por ello, el objetivo general de este trabajo es la identificación y comprensión de los mecanismos moleculares que afectan de forma diferencial entre sexos en la artritis reumatoide. Esta información permitiría una mayor precisión en el diagnóstico y selección de los tratamientos para cada grupo de pacientes.

Este objetivo general se alcanzará mediante el cumplimiento de los siguientes objetivos específicos:

1. Revisión sistemática y selección de estudios de artritis reumatoide con datos de expresión de microarrays o RNA-seq en los repositorios Gene Expression Omnibus (GEO), ArrayExpress, u otras bases de datos.
2. Análisis individual de cada estudio seleccionado, incluyendo un análisis de expresión diferencial y un análisis de enriquecimiento funcional de procesos biológicos de la GO y de rutas KEGG.
3. Metaanálisis de genes y enriquecimiento funcional de los resultados globales, que posibilitará la identificación de biomarcadores específicos por sexo.

3. MATERIAL Y MÉTODOS

El análisis bioinformático descrito en esta sección se realizó en el lenguaje de programación R (versión 4.0.3) (R Core Team, 2020). El código está disponible en github (https://github.com/edurlaf/TFG_meta-analysis_RA). Las librerías y paquetes empleados se recogen en el Anexo B (Tabla B.1).

3.1. Revisión sistemática y selección de estudios

La primera parte del estudio consistió en una revisión sistemática. Este proceso permite comparar toda la evidencia disponible para responder a una pregunta concreta, mediante la recopilación y síntesis de los resultados de estudios individuales (Higgins, 2019). Además, al combinarse con un metaanálisis cuantitativo, es posible evaluar la magnitud del resultado en todos los estudios relevantes y analizar las causas de la variación entre los resultados. Por ello, han adquirido una especial importancia en los últimos años en el ámbito de la salud, y se emplean a menudo como punto de partida para el desarrollo de guías de práctica clínica (Gurevitch et al., 2018; Moher et al., 2009).

Las revisiones sistemáticas siguen un proceso estructurado y predefinido que requiere métodos rigurosos y reproducibles para garantizar que los resultados sean fiables y significativos. En primer lugar, es necesario establecer un objetivo o pregunta concreta, y especificar unos criterios de elegibilidad. A continuación, se identifican todos los registros relevantes, se seleccionan los estudios que cumplan los criterios establecidos y se evalúa el riesgo de sesgo. Finalmente, se extraen los datos y, en la mayoría de los casos, se realiza un metaanálisis, que mediante métodos estadísticos permite resumir y combinar los resultados de estudios independientes y extraer así conclusiones (Liberati et al., 2009; Moher et al., 2009).

Por lo tanto, en este primer paso se revisó exhaustivamente la investigación existente relacionada con la artritis reumatoide, y se seleccionaron los estudios que permitirían establecer las diferencias entre sexos en esta enfermedad. Todo ello conforme a las directrices estipuladas en la declaración PRISMA, resultado de una colaboración internacional que tiene como objetivo estandarizar la metodología de las revisiones sistemáticas (Urrutia y Bonfill, 2010; Liberati et al., 2009).

En primer lugar, se identificaron los estudios de expresión génica disponibles en las bases de datos públicas GEO (Barrett et al., 2013) y ArrayExpress (Athar et al., 2019) que cumplían los criterios de inclusión especificados en la Tabla 1. Para ello, se emplearon las palabras clave '*rheumatoid arthritis*', '*expression profiling by array*' y '*expression profiling by high throughput sequencing*'. Además, se incluyeron sólo los estudios en '*Homo sapiens*', y se excluyeron los basados en '*cell lines*', '*organoid*' o '*xenograft*'. Dado que las muestras se clasificarían posteriormente en función del sexo y el diagnóstico en cuatro grupos experimentales (mujeres AR, hombres AR, mujeres control y hombres control), y que el objetivo era que cada grupo estuviera compuesto por al menos 3 individuos, se excluyeron también los registros con un número de muestras inferior a 12. También se consultaron los registros sobre artritis reumatoide disponibles en la base de datos ADEx de Genyo (Martorell-Marugán et al., 2020).

Tabla 1 | Criterios de inclusión y exclusión empleados en la identificación de estudios candidatos. n, tamaño muestral.

Criterios de inclusión	Criterios de exclusión
"*Rheumatoid arthritis*"	Otras enfermedades
n ≥ 12	n < 12
RNA-seq o array de expresión	Otros tipos de experimentos (*single-cell*, microRNAs, tratamientos *ex vivo*)
Homo sapiens	Cultivos celulares u otros organismos
Información de sexo	Sin información de sexo
Al menos 3 individuos en cada grupo	Sin controles o sin suficientes individuos por grupo

A continuación, durante la fase de cribado, se descartaron los estudios atendiendo a los criterios de exclusión descritos en la Tabla 1. Finalmente, se evaluaron a texto completo los artículos asociados a cada estudio con el fin de seleccionar los que cumplían todos los requisitos de elegibilidad, como contener información sobre el sexo (Urrutia y Bonfill, 2010). El flujo de este proceso se recoge en un diagrama PRISMA (Figura 2).

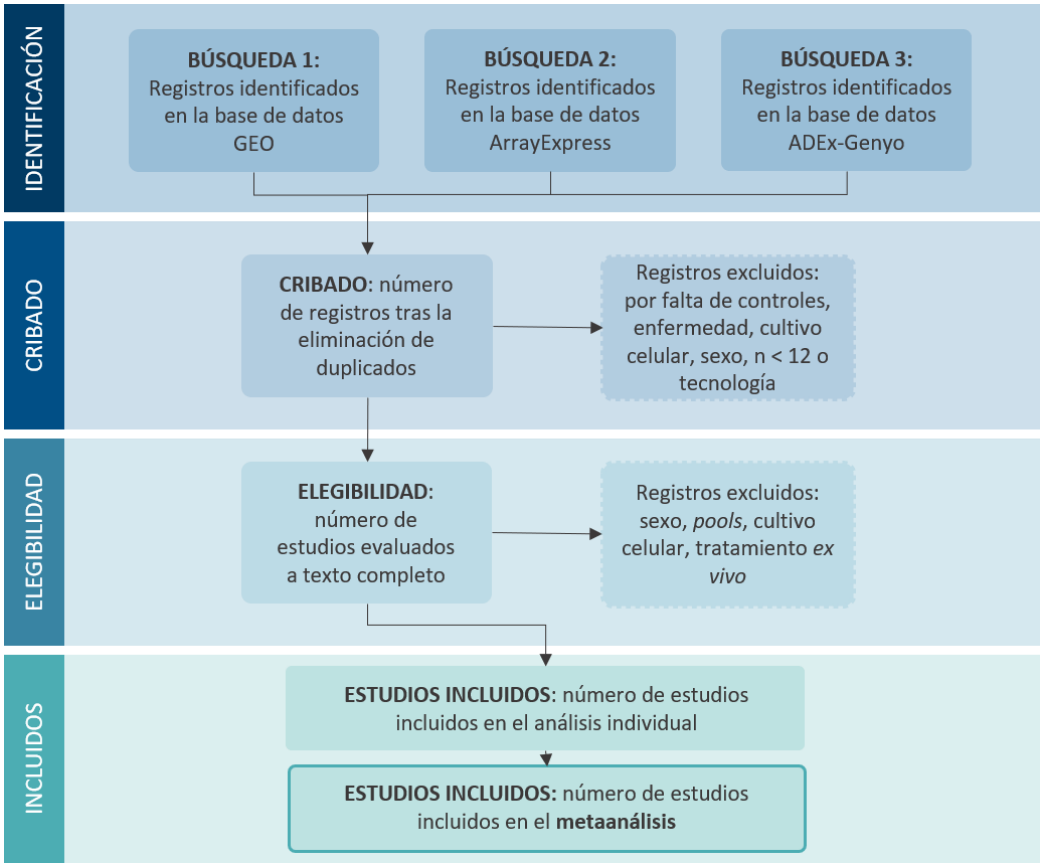

Figura 2 | Flujo de información a través de las diferentes fases de una revisión sistemática. Esquema adaptado de Cochrane Handbook (Higgins, 2019). n, tamaño muestral.

3.2. Análisis individual de los estudios

Para llevar a cabo el metaanálisis, se analizaron los estudios individualmente y, finalmente, se combinaron los resultados.

3.2.1. Procesamiento de los datos

Todos los estudios seleccionados para el análisis pertenecen a la base de datos GEO (Barrett et al., 2013), por lo que los datos se obtuvieron mediante la función `getGEO()` del paquete **GEOquery** (Sean

y Meltzer, 2007). Este paquete sirve como intermediario entre GEO, un repositorio con miles de experimentos de expresión génica, y BioConductor (Huber et al., 2015), un software de código abierto construido en R para el análisis de datos genómicos.

Cada grupo de datos asociado a un tipo de tecnología ómica requiere un método específico de normalización. Por ello, en los estudios de microarrays se partió de los datos ya normalizados. En el caso de los estudios de RNA-seq, se descargaron manualmente las matrices de conteo crudas y, mediante la función **filterByExpres()**, se filtraron los genes con un mínimo de conteos para el análisis estadístico. A continuación, se normalizaron por el método Trimmed Mean of M-values (TMM), con la función **calcNormFactors()** del paquete **edgeR** (Robinson y Oshlack, 2010; Robinson et al., 2009). Este método se basa en la asunción de que la mayoría de los genes no están diferencialmente expresados, y realiza una normalización del tamaño de librerías entre muestras. Para ello toma una muestra como referencia, y calcula los *fold change* y los niveles de expresión absolutos con relación a ella. Se excluyen o "recortan" los genes más expresados y los que poseen las mayores *log ratios*, y a continuación se halla la media de los *fold change* para cada muestra. Los recuentos de lecturas se escalan por esta media recortada y el recuento total de su muestra (Abbas-Aghababazadeh et al., 2018; Evans et al., 2018).

Además, se cribaron todos los estudios con el fin de eliminar muestras duplicadas o que no se correspondieran con controles o pacientes de artritis reumatoide, y se comprobó la ausencia de datos no disponibles (NAs) y el rango de los datos.

3.2.1.1 Análisis exploratorio

Se realizó un análisis exploratorio de cada uno de los estudios, con el objetivo de comprobar que presentaban la misma escala y asegurar la ausencia de valores anómalos, y de esta forma garantizar su comparación. También se comprobó la distribución de las muestras en los diferentes grupos experimentales y la presencia de efecto *batch* (variabilidad técnica sistemática que aparece cuando las muestras se procesan y miden en diferentes lotes, no relacionada con la variación biológica) (Lazar et al., 2013).

Para ello, el primer paso consistió en describir los niveles de expresión por muestra en cada estudio mediante diagramas de caja, construidos con el paquete **plotly** (Sievert, 2020). Además, en el caso de que no se hubiera hecho, se aplicó la transformación logarítmica en base 2, asegurando así la misma escala y facilitando la interpretación de los datos.

A continuación, se realizó un *clustering* jerárquico en base a la distancia de correlación, con el fin de identificar patrones de expresión génica en las muestras (Altman y Krzywinski, 2017; Zhang et al., 2009). A modo complementario, se realizó también un análisis de componentes principales (PCA). El PCA permite obtener una imagen completa de los datos, ya que captura la varianza del conjunto de datos, y ayuda a determinar su origen y a identificar valores atípicos, reduciendo la dimensionalidad de los datos (Zhang et al., 2009). Ambos son métodos no supervisados que buscan patrones sin requerir información previa al respecto (Lever et al., 2017). Tanto el *clustering* como el PCA se representaron mediante el paquete **ggplot2** (Wickham, 2016).

3.2.1.2 Estandarización de los datos clínicos

Para facilitar la integración de los datos en el metaanálisis posterior, se renombraron los datos clínicos incluidos dentro de las variables factoriales diagnóstico y sexo. De este modo, las muestras

se clasificaron en función del diagnóstico como "*Control*" (individuos sanos) o "*RA*" (pacientes diagnosticados con artritis reumatoide); y como "*Female*" o "*Male*" en función del sexo.

Además, se generó una nueva variable que incluye el grupo experimental al que pertenece cada muestra mediante la combinación de las variables diagnóstico y sexo. Así, las muestras se clasificaron en cuatro posibles grupos experimentales: *Control_Male*, *Control_Female*, *RA_Male* y *RA_Female*.

3.2.1.3 Anotación

Las diferencias en la anotación de las sondas complican el metaanálisis de estudios pertenecientes a diferentes plataformas (Arloth et al., 2015). Por ello, es fundamental transformar el código de las sondas de cada microarray en un identificador común que permita compararlas entre sí. En este caso, se optó por el identificador único de la base de datos Entrez Gene del *National Center for Biotechnology Information* (NCBI) (Maglott et al., 2011).

Para anotar los microarrays, se descargaron los paquetes de R asociados a cada plataforma y, empleando la función **select()** y el identificador de sonda como argumento **keytype**, se extrajeron los identificadores Entrez, así como el nombre de gen y su símbolo correspondiente. La función **match()** permitió relacionar las sondas de las matrices de expresión con la información extraída del paquete asociado. Los paquetes empleados fueron **illuminaHumanv2.db** (Dunning et al., 2015), **hgu133plus2.db** (Carlson, 2016a) y **hgu219.db** (Carlson, 2016b). Dado que los microarrays suelen contener diversas sondas que hibridan con el mismo gen, la anotación con el identificador Entrez genera duplicados en la matriz de expresión. Esta redundancia es una fuente de ruidos y errores en análisis posteriores (Zhang et al., 2009). Por ello, se obtuvo un valor único de expresión para cada gen mediante el cálculo de la mediana de los niveles de expresión de todas las sondas que hibridaban con el mismo gen.

Los identificadores de gen de los estudios de RNA-seq también se sustituyeron por los identificadores Entrez. En este caso, sin embargo, se empleó **BiomaRt** (Durinck et al., 2009). Este paquete de Bioconductor proporciona una Interfaz de Programación de Aplicaciones (API) para los servicios web de *BioMart* desde R. *BioMart*, a su vez, facilita el acceso a mapeos, relaciones y anotaciones de diversas bases de datos biológicas como Ensembl y Reactome. Para ello, se conectó con la base de datos Ensembl (Howe et al., 2021) y, a partir del dataset hsapiens_gene_ensembl, se extrajeron los atributos especificados (identificador Entrez y nombre y símbolo del gen) mediante las funciones **useMart()**, **useDataset()** y **getBM()** respectivamente.

3.2.2. Análisis de expresión diferencial

Uno de los objetivos más comunes en investigación genómica consiste en identificar genes expresados diferencialmente entre condiciones experimentales (Phipson et al., 2016). En este estudio, se trató de identificar los genes diferencialmente expresados entre mujeres y hombres con artritis reumatoide, según el siguiente contraste de interés:

Ecuación 1

$$(Mujeres_{AR} - Mujeres_{Control}) - (Hombres_{AR} - Hombres_{Control})$$

También se estudiaron los siguientes contrastes:

Ecuación 2

$$Mujeres_{AR} - Mujeres_{Control}$$

Ecuación 3

$$Hombres_{AR} - Hombres_{Control}$$

Con este fin se empleó el paquete de software **limma** (componente de Bioconductor), uno de los paquetes más populares para el análisis de expresión diferencial por su rapidez y buen rendimiento. Se emplea tanto en el análisis de microarrays como de RNA-seq y, tras un

preprocesado y normalización iniciales, permite emplear *pipelines* muy similares para ambos (Ritchie et al., 2015; Seyednasrollah et al., 2013).

Para ello, toma como *input* una matriz con valores de expresión, donde cada fila representa un gen, y cada columna una muestra. En primer lugar, se construyó una matriz de diseño con las variables (el grupo experimental al que pertenece cada muestra y el lote, en los estudios que corresponda) y otra matriz con los contrastes de interés (Ecuación 1-3) mediante las funciones **model.matrix()** del paquete **stats** y **makeContrasts()** del paquete **limma** respectivamente. A continuación, la función **lmFit()** del paquete **limma** ajustó un modelo lineal a cada gen a partir de las matrices de expresión y de diseño mediante el método de mínimos cuadrados. Seguidamente, la función **contrasts.fit()** calculó los coeficientes estimados y los errores estándar para cada contraste (Smyth et al., 2011).

Los experimentos de expresión génica suponen un reto a nivel estadístico, ya que suelen contener un gran número de datos, pero un tamaño muestral pequeño. Por ello, el empleo de métodos estadísticos univariantes sobre cada gen puede producir resultados imprecisos. Una posibilidad consiste en emplear la información del *dataset* completo al realizar la inferencia sobre cada gen individual. Por ello, se aplicó finalmente la función **eBayes()**, que calcula los estadísticos *t* y *F* moderados y las probabilidades logarítmicas de expresión diferencial mediante la moderación empírica de Bayes de los errores estándar. Como resultado, los grados de libertad de las varianzas individuales aumentaron, reflejando así la información adicional obtenida y aumentando el poder estadístico. En definitiva, el paquete **limma**, al analizar los experimentos de forma integrada, permitió modelar las correlaciones entre las muestras y ajustar los efectos de lote (Phipson et al., 2016).

En el caso de los estudios de RNA-seq, la relación media-varianza se modeló con pesos de precisión (enfoque "voom") o con una tendencia previa a Bayes (enfoque "limma-trend"). Si la profundidad de secuenciación es consistente en todas las muestras, limma-trend (el uso de **eBayes()** con **trend=TRUE**) es el enfoque más simple y robusto. Para ello, la relación entre el tamaño más grande de la librería y el más pequeño no debe ser superior a 3 (Law et al., 2014; Smyth et al., 2011). Por el contrario, cuando el tamaño de las librerías es variable entre muestras, se aplica la función **voom()** antes del ajuste lineal (Costa-Silva et al., 2017). El método voom incorpora la tendencia de la varianza media en las ponderaciones de precisión, mientras que limma-trend la incorpora en la moderación empírica de Bayes.

Finalmente, la función **topTable()** mostró los parámetros estadísticos de cada gen, entre ellos el P-valor ajustado mediante el método de Benjamini y Hochberg (BH o False Discovery Rate, FDR) (Benjamini y Hochberg, 1995). También se calculó el error estándar.

3.2.3. Análisis de enriquecimiento funcional

El enriquecimiento funcional permite detectar conjuntos de genes que se sobreexpresan o subexpresan coordinadamente. Para llevarlo a cabo, se empleó el paquete **mdgsa** (Montaner y Dopazo, 2010) y se partió de los datos de expresión diferencial de los genes. Además, en el caso de los términos GO, y tal y como se ha descrito en la anotación, se emplearon las funciones **useMart()**, **useDataset()** y **getBM()** para extraer de la base de datos Ensembl de BioMart los identificadores Entrez y los términos GOs asociados en una matriz de anotación.

Se emplearon, respectivamente, las funciones **annotMat2list()**, **splitOntologies()** y **propagateGO()** del paquete **mdgsa** para convertir la matriz de anotación en lista, dividirla en función de la ontología a la que pertenece cada término, y propagar la anotación. De esta forma los genes asociados a un término GO heredan la anotación de sus ancestros. A continuación, se filtraron los bloques funcionales demasiado específicos (menos de 10 genes asociados) o genéricos (más de 500) mediante la función **annotFilter()**. A partir de los estadísticos obtenidos en el análisis de expresión diferencial, la función **pval2index()** transformó los P-valores y el signo de su estadístico asociado en

un índice de clasificación. Después, dicho índice se transformó en cuartiles de una distribución normal (**method = "normalize"**) mediante la función **indexTransform(),** para que su distribución fuera adecuada como variable independiente del modelo de regresión logística univariante empleado en el GSEA. Finalmente, el enriquecimiento se realizó mediante la función **uvGsa()**. Se consideraron significativos los términos con un P-valor ajustado por el método de Benjamini-Yekutieli (Benjamini y Yekutieli, 2001) inferior a 0.05.

En el caso de las rutas KEGG, se empleó el paquete **org.HS.eg.db** (Carlson, 2019) para extraer los términos KEGG (*path*) y obtener la matriz de anotación. El resto del procedimiento se realizó de modo similar, aunque sin dividir ni propagar la anotación, ya que no se aplica a estructuras no jerárquicas.

3.3. Metaanálisis

En este estudio se realizó el metaanálisis a nivel génico. Para ello, se empleó el paquete **metafor**, que contiene funciones para el ajuste de modelos de metaanálisis de efectos fijos y aleatorios, y permite incluir variables moderadoras. También provee funciones para la representación de los resultados (como los gráficos de bosque, de embudo o radiales) y la evaluación del ajuste del modelo (Viechtbauer, 2020).

Este paquete requiere que el tamaño del efecto se calcule antes del proceso de ajuste del modelo (Lortie y Filazzola, 2020). Para ello, se tomaron como *input* los resultados del análisis diferencial y se construyeron dos matrices: una con los logaritmos de la magnitud de cambio de cada gen (logFC) y otra con los errores estándar. Después, se filtraron y descartaron los genes que no estuvieran presentes en al menos dos estudios.

3.3.1. Determinación de la medida combinada del efecto

Para el metaanálisis se empleó la función **rma()**. Esta función ajusta modelos meta-analíticos de efectos fijos, aleatorios o mixtos con o sin moderadores mediante modelos lineales. Se emplea con las medidas de efecto habituales, como *log odds ratios* o diferencias de medias. Los resultados observados se especifican mediante el argumento **yi** (es decir, las matrices logFC), y los errores estándar mediante el argumento **sei**. Se optó por un modelo de efectos aleatorios, que considera la heterogeneidad de los estudios, empleando el método DerSimonian-Laird (DL) para estimar su varianza (Langan et al., 2019, Viechtbauer, 2020).

El metaanálisis se realizó para cada uno de los genes, por lo que también se calculó el P-valor ajustado mediante el método FDR, y se consideraron significativos los resultados con un P-valor ajustado inferior a 0.05.

3.3.2. Evaluación de la heterogeneidad

La función **rma()** realiza automáticamente una prueba de heterogeneidad: la prueba Q de Cochran, que comprueba si la variabilidad en los tamaños de los efectos o resultados observados es mayor de lo que cabría esperar basándose únicamente en la variabilidad del muestreo. Una prueba significativa sugiere que los verdaderos efectos o resultados son heterogéneos (Viechtbauer, 2020). Sin embargo, este test presenta una baja potencia al trabajar con un número reducido de estudios, por lo que el análisis de heterogeneidad se complementó con *boxplots* de las magnitudes de cambio y los errores estándar, así como con gráficos de embudo.

Un gráfico de embudo es un gráfico de dispersión de las estimaciones del efecto de los estudios individuales frente a una medida del tamaño o precisión de cada estudio, y permite evaluar la variabilidad y la presencia de sesgos. Frecuentemente se elige el error estándar de la estimación del efecto como medida del tamaño del estudio y se representa en el eje vertical con una escala invertida. De esta manera los estudios más grandes y potentes se sitúan en la parte superior, con una dispersión más reducida. Las estimaciones del efecto de los estudios más pequeños, por el

contrario, se dispersan más en la parte inferior. En ausencia de sesgo y de heterogeneidad entre estudios, la dispersión se deberá únicamente a la variación del muestreo y el gráfico se asemejará a un embudo invertido simétrico (Sterne et al., 2011).

3.3.3. Representación de resultados

Posteriormente se construyeron, además de los mencionados gráficos de embudo, gráficos de bosque para cada uno de los genes significativos hallados en el metaanálisis.

Los gráficos de bosque son una representación gráfica del metaanálisis, y muestran de un vistazo la información de los estudios individuales que conforman el metaanálisis, la cantidad de variación entre estudios y una estimación del resultado global. El eje horizontal representa el estadístico de estudio (en este caso, el logFC), mientras que la línea vertical punteada se conoce como "línea de efecto nulo", y se sitúa en el valor en el que no hay diferencias entre grupos. En el caso de los estadísticos absolutos, este valor se corresponde con el 0. En la parte superior aparece cada estudio individual, con dos componentes: una caja y una línea horizontal. La caja negra representa la estimación puntual de cada estudio, así como su tamaño (a mayor tamaño, mayor número de muestras en el estudio). La línea o bigote representa el intervalo de confianza del estudio (generalmente 95%). La parte inferior, en cambio, muestra un diamante que representa la estimación puntual (vértices verticales del diamante) y el intervalo de confianza (vértices horizontales) del metaanálisis global. Si los vértices horizontales del diamante cruzan la línea de efecto nulo, el resultado combinado no es estadísticamente significativo (Sedgwick, 2015; Lewis y Clarke, 2001).

3.3.4. Enriquecimiento funcional de los resultados del metaanálisis

Finalmente, se realizó un enriquecimiento funcional a partir de los resultados del metaanálisis, similar al descrito para los estudios individuales (apartado 3.2.3), con el objetivo de enriquecer la lista de genes consenso obtenida e identificar funciones. Se llevó a cabo tanto a nivel de procesos biológicos de la GO, como de rutas KEGG.

4. RESULTADOS

A continuación, se muestran los principales resultados obtenidos en el estudio. En los Anexos A y B se recogen el resto de figuras y tablas complementarias.

4.1. Revisión sistemática y selección de estudios

En la revisión sistemática de las bases de datos GEO, ArrayExpress y ADEx realizada en febrero de 2021 se identificaron, tras la eliminación de duplicados, 200 estudios candidatos. Durante el cribado, se eliminaron 159 estudios por no reunir los criterios necesarios para su inclusión en la revisión: falta de controles, enfermedad diferente a la AR, sexo (sin información al respecto o falta de pacientes masculinos), tecnología empleada o uso de cultivos celulares (por interferir en la expresión génica). De los 41 estudios restantes, se descartaron otros 35 tras la evaluación de los artículos, debido a falta de información de sexo, al empleo de *pools* de varias muestras como control o al tratamiento *ex vivo* de las muestras con estimulantes.

Finalmente, se seleccionaron 6 estudios, 5 de ellos identificados en la base de datos GEO, y uno localizado en ADEx. Cabe señalar que este último no había sido detectado en la búsqueda en GEO ya que entre sus identificadores no constaba el término artritis reumatoide. Los estudios seleccionados se muestran en la Tabla 2, y en la Figura 3 el diagrama PRISMA representa el esquema de flujo del proceso.

Tabla 2 | Estudios seleccionados tras la revisión sistemática. Se indica el identificador del estudio, la plataforma empleada y las publicaciones asociadas.

Estudio	Plataforma	Artículo
GSE117769	Illumina HiSeq 2500	Sin publicación
GSE93272	Affymetrix Human Genome U133 Plus 2.0 Array	Tasaki et al., 2018
GSE97165, GSE89408	Illumina HiSeq 2000	Walsh et al., 2017
GSE17755	Hitachisoft AceGene Human Oligo Chip 30K 1 Chip Version	Lee et al., 2011
GSE15573	Illumina human-6 v2.0 expression beadchip	Teixeira et al., 2009
GSE110169	Affymetrix Human Genome U219 Array	Hu et al., 2018

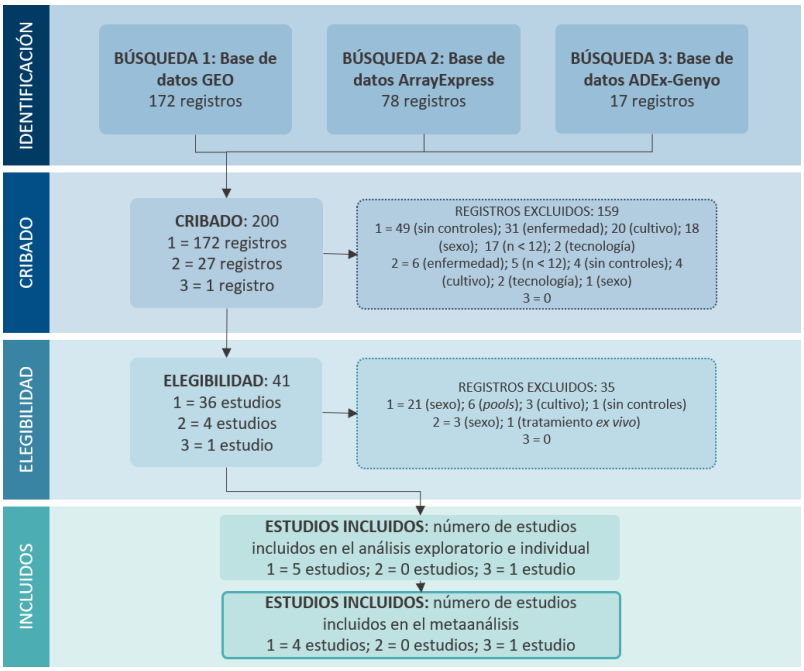

Figura 3 | Diagrama de flujo con los resultados de la revisión sistemática, según la declaración PRISMA. *Abreviaturas: GEO, Gene Expression Omnibus; n, tamaño muestral.*

4. RESULTADOS

El principal factor limitante en la selección de estudios fue la falta de controles, seguido del sexo, la enfermedad (los resultados de la búsqueda incluían otras enfermedades autoinmunes) y el empleo de cultivos celulares. Otros factores de exclusión fueron el tamaño muestral (inferior a 12, debido a la presencia de réplicas), el empleo de *pools*, la tecnología utilizada (*single-cell* o microRNAs) o el tratamiento *ex vivo* de las muestras.

4.2. Análisis individual de los estudios

4.2.1. Procesamiento de los datos

Una vez descargados los datos de los estudios y normalizadas las matrices de conteos de los estudios de RNA-seq (GSE117769 y GSE89408), se cribaron las muestras, seleccionando las de interés. Tras este procesado, se evaluó la normalización (listado de las normalizaciones empleadas en cada *dataset* en el Anexo B, Tabla B.2) y la distribución de las muestras en los diferentes grupos experimentales mediante diagramas de cajas (*boxplots*), *clusterings* jerárquicos y PCAs. Además, se describió el número de individuos en cada grupo y en total (Figura 4). Como se observa, aunque a nivel de estudios individuales la distribución de sexos no es equilibrada, en conjunto sí se encuentra balanceada. Sin embargo, existe cierto desequilibrio entre los grupos de controles y pacientes.

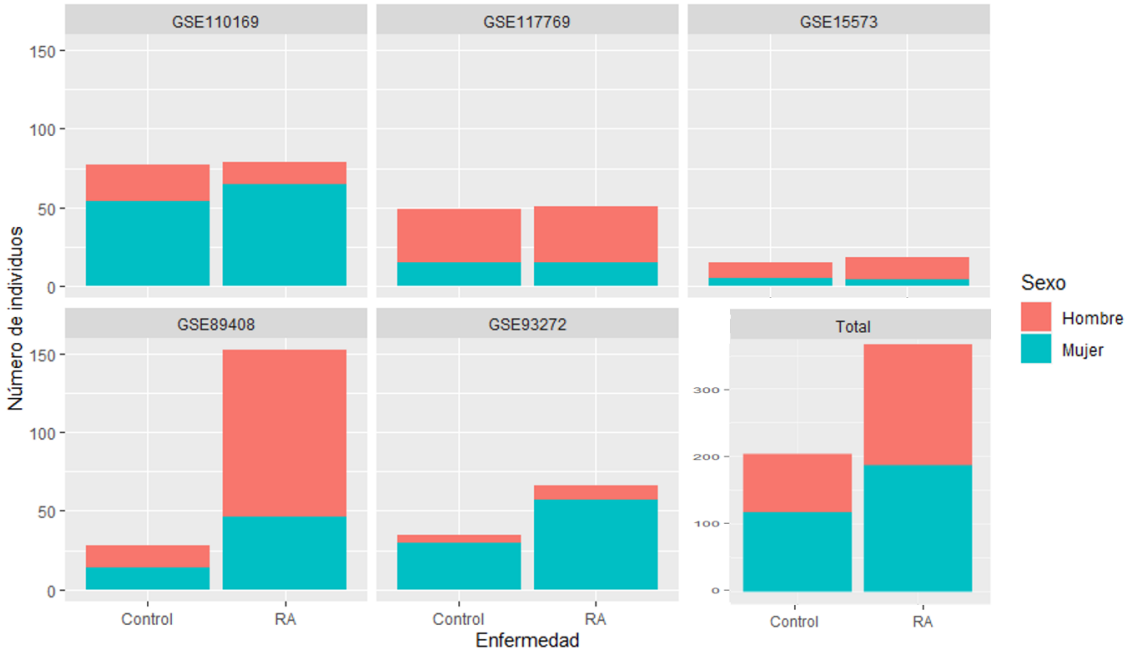

Figura 4 | Número de individuos en cada estudio y en total. Se encuentran distribuidos por grupo experimental (sexo y diagnóstico). *Abreviaturas: RA, Artritis reumatoide.*

GSE110169

En este estudio de microarrays se desarrolló una firma genética a partir de sangre periférica. El objetivo era monitorizar la respuesta farmacodinámica a la administración de glucocorticoides en pacientes con artritis reumatoide y lupus eritematoso sistémico (SLE) (Hu et al., 2018). La **¡Error! No se encuentra el origen de la referencia.** muestra la distribución de individuos en cada grupo experimental, tras eliminar las muestras de pacientes de SLE y de 5 individuos sin sexo asignado.

El *boxplot,* el análisis de componentes principales y el *clustering* se realizaron en función de los grupos experimentales y del lote, con el fin de comprobar si existía un efecto *batch*. En el *boxplot* en función de los grupos experimentales (Figura 5) se observa que los datos se encuentran

normalizados, y no parece haber valores atípicos. Sin embargo, se aprecian 4 grupos que no coinciden con los grupos experimentales. La Figura 6, coloreada en función de los lotes, permite confirmar que se trata de un claro efecto *batch*. Este efecto es una de las principales fuentes de variación no deseada y obstaculiza la integración de datos, por lo que fue necesario considerarlo en la siguiente etapa del análisis de expresión diferencial (Lazar et al., 2013). Cabe señalar que, comparando la Figura 5 y la Figura 6, se observa que todos los lotes contienen muestras de cada grupo experimental.

Tanto el análisis de *clustering* como el PCA (Anexo A, Figuras A.1-3) agruparon las muestras en función del lote al que pertenecían. Dado que la mayor variabilidad está asociada al efecto lote, el posible agrupamiento de las muestras en función del grupo experimental quedó enmascarado.

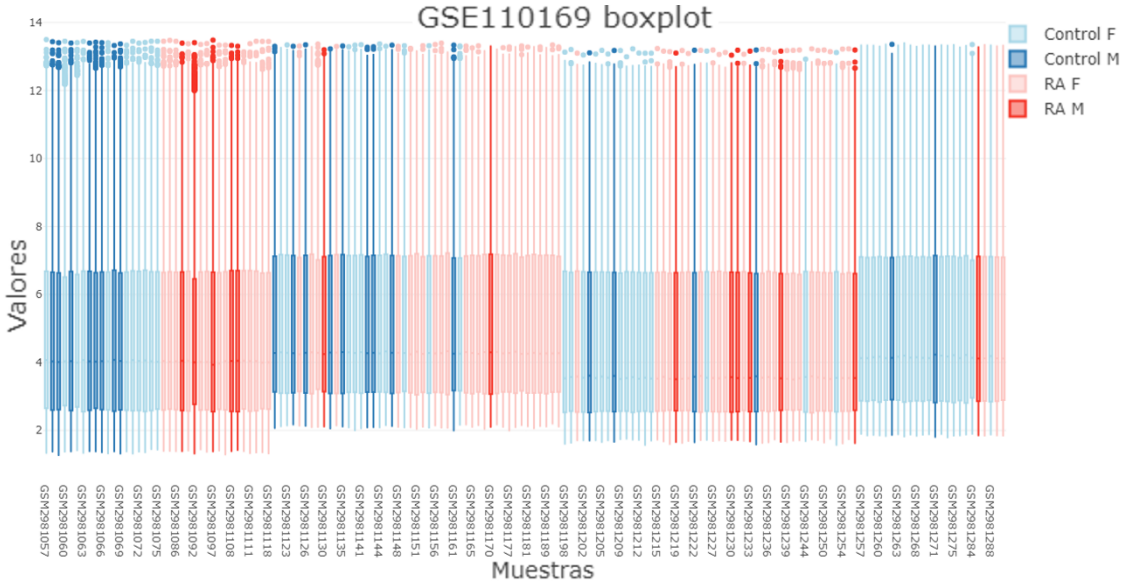

Figura 5 | Diagrama de cajas del estudio GSE110169. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres enfermas (RA F) y hombres enfermos (RA M). Se observan 4 grupos de muestras que no coinciden con los experimentales.

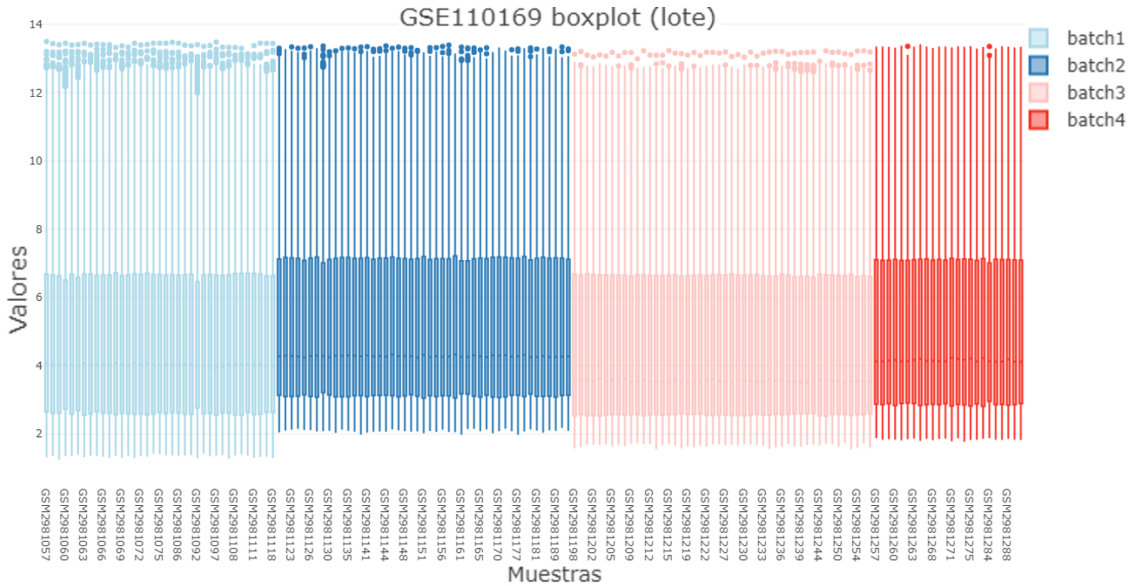

Figura 6 | Diagrama de cajas del estudio GSE110169 en función de la variable lote (*batch*). Representa los valores de expresión (eje Y) de cada muestra (eje X). Se observan 4 grupos de muestras que coinciden con los lotes.

4. RESULTADOS

GSE117769

En este estudio se evaluaron las diferencias de expresión génica entre pacientes de artritis reumatoide y donantes sanos. Para ello se comparó mediante RNA-seq el transcriptoma de sangre completa de 50 pacientes de AR con donantes sanos, emparejados por edad, género y etnia. La Figura 4 muestra la distribución de individuos en cada grupo experimental tras eliminar las muestras de psoriasis artrítica y espondilitis anquilosante.

Tras realizar el análisis exploratorio (*boxplot*, PCA y *clustering*), se decidió eliminar la muestra GSM3308533, por tratarse de un valor anómalo, como se observa en la Figura 7. Sin embargo, a pesar de su eliminación, ni el análisis de *clustering* ni el PCA mostraron agrupamiento en función de los grupos experimentales (Anexo A, Figuras A.4-8).

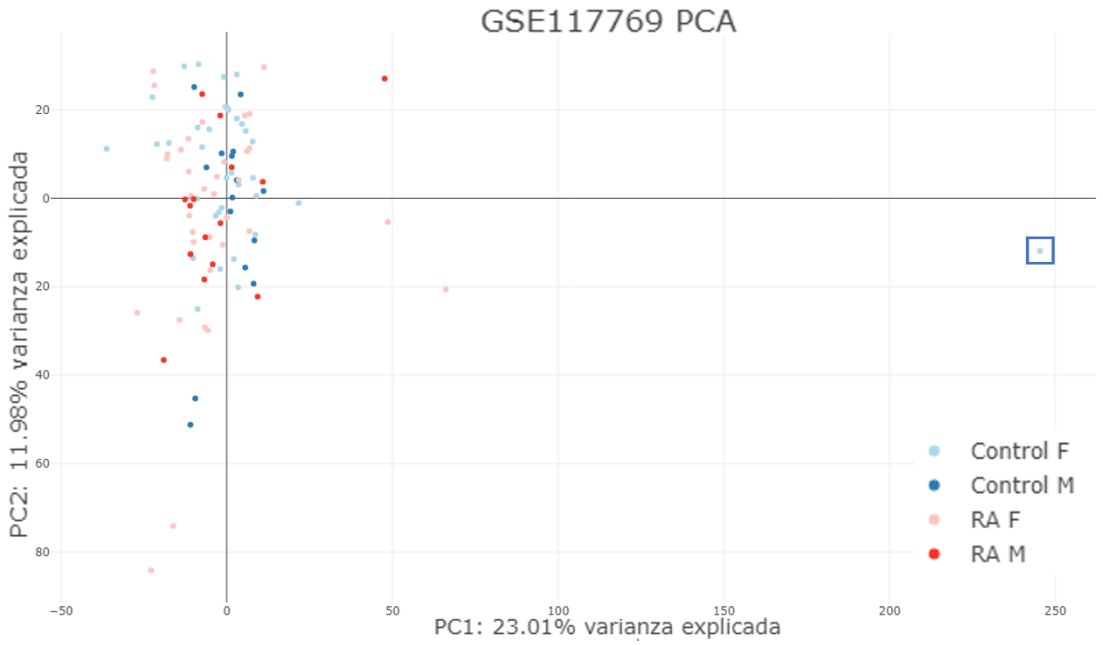

Figura 7 | Análisis de componentes principales (PCA) del estudio GSE117769 antes de la eliminación de la muestra GSM3308533. Se muestran las componentes principales (PC) 1 y 2, y las muestras se han coloreado en función del grupo experimental. Encuadrada aparece la muestra GSM3308533. *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

GSE89408 y GSE97165

Este estudio, formado por dos *datasets*, consistió en comparar mediante RNA-seq tejidos sinoviales pre- y postratamiento de pacientes con AR temprana, con el objetivo de observar los efectos de una combinación de fármacos anti-reumatoides (Walsh et al., 2017). La Figura 4 muestra la distribución de individuos en cada grupo experimental tras eliminar las muestras de osteoartritis, artralgia y artritis indiferenciada, y las muestras duplicadas postratamiento del *dataset* GSE89408.

Tras el análisis exploratorio inicial, se decidió eliminar el *dataset* GSE97165 del metaanálisis, con el fin de evitar posibles problemas de efecto lote y por no contener todos los grupos experimentales (Anexo A, Figuras A.9-14). Tanto el PCA como el *clustering* del GSE89408 mostraron cierto agrupamiento de las muestras en función del diagnóstico (Figura 8 y Figura 9).

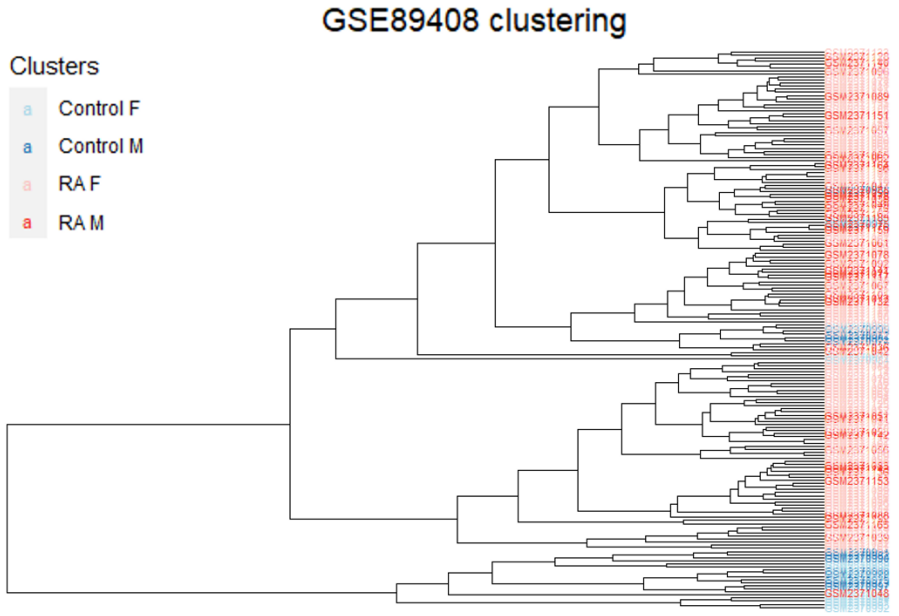

Figura 8 | *Clustering* jerárquico del GSE89408 en base a la distancia de correlación. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (AR) (RA F) y hombre con AR (RA M).

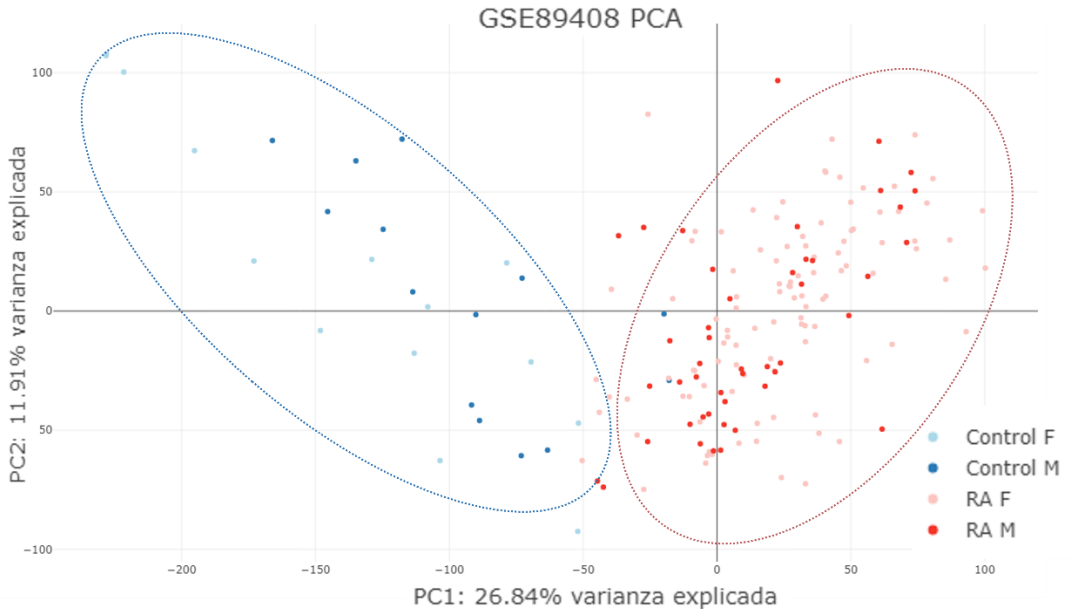

Figura 9 | Análisis de componentes principales (PCA) del estudio GSE89408. Se muestran las componentes principales (PC) 1 y 2, y las muestras se han coloreado en función del grupo experimental. La línea punteada azul engloba los controles, y la roja los enfermos. *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

GSE93272

Este estudio consiste en una monitorización de la respuesta a fármacos de pacientes con AR mediante una aproximación multi-ómica en sangre periférica. El fin es conocer las diferencias subyacentes a la remisión clínica sostenida de la enfermedad y el estado sano (Tasaki et al., 2018). La Figura 4 muestra la distribución de individuos en cada grupo experimental, tras eliminar las muestras duplicadas de pacientes de AR y seleccionar las previas al inicio del tratamiento y los controles.

El *boxplot*, el PCA y el *clustering* se realizaron también tomando como variable la información de lote, con el fin de comprobar si existía efecto *batch* (Anexo A, Figuras A.15-19). El *boxplot* en función de los grupos experimentales (Figura 10) permite comprobar la normalización

de los datos, y no parece haber valores atípicos. La comparación de los PCAs coloreados en función de los grupos (Figura 11) y de los lotes (Anexo A, Figura A.19), por su parte, parece indicar la ausencia de efecto *batch*. Aun así, la variable lote se tendrá en cuenta en las siguientes etapas de análisis de expresión diferencial. Tanto en el PCA como en el *clustering* parece observarse cierta agrupación en función del diagnóstico.

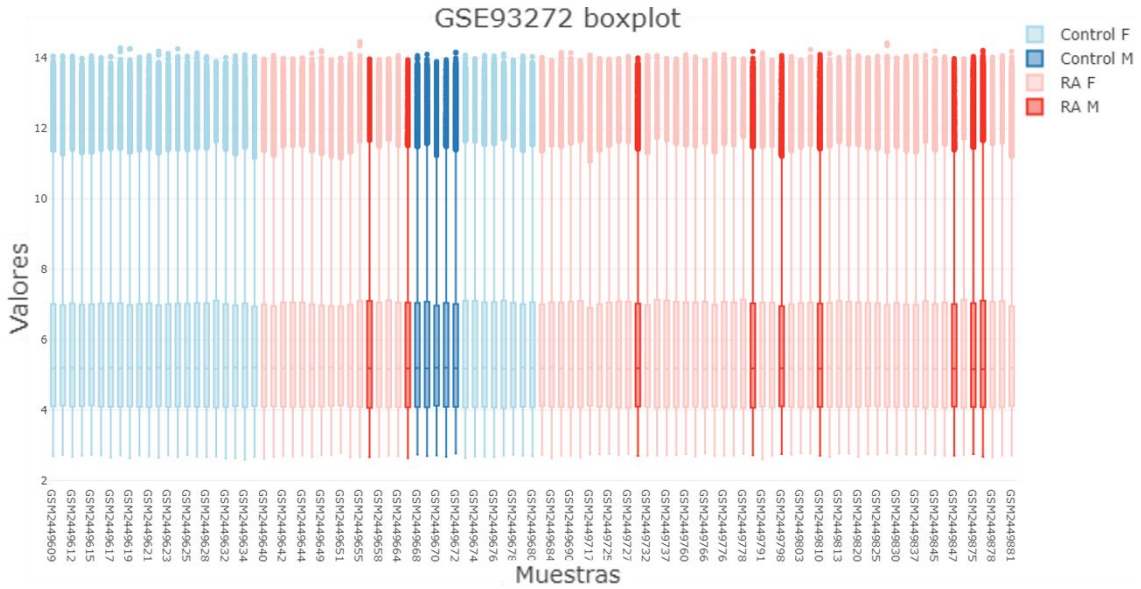

Figura 10 | Diagrama de cajas del estudio GSE93272. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales. *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

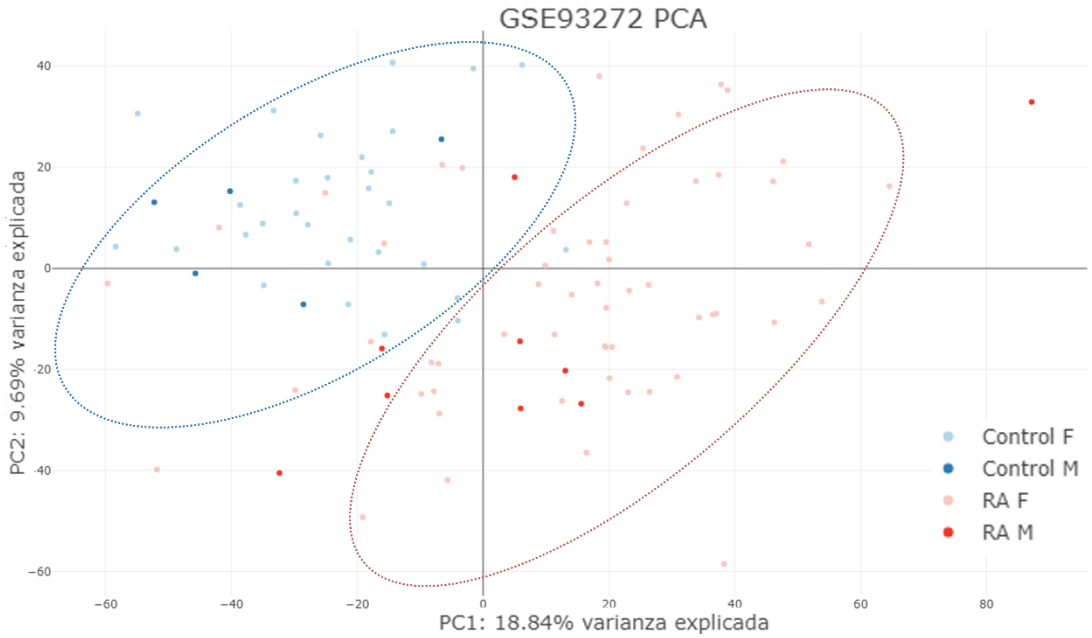

Figura 11 | Análisis de componentes principales (PCA) de las muestras del estudio GSE93272. Se muestran las componentes principales (PC) 1 y 2. Coloreado en función de los grupos experimentales. Se observa cierta agrupación en función del diagnóstico: controles (línea de puntos azul) y RA (línea de puntos roja). *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

GSE17755

En este estudio se caracterizaron las diferencias biológicas y funcionales del SLE comparando los perfiles de expresión de pacientes con controles sanos y muestras de otras enfermedades autoinmunes, como la AR (Lee et al., 2011). El estudio se realizó mediante microarrays de dos canales y se normalizó a la mediana. Además, más del 50% de los valores de la matriz de expresión no estaban disponibles, por lo que, tras el análisis exploratorio, se decidió eliminar del metaanálisis.

4. RESULTADOS

GSE15573

El objetivo de este estudio consistió en identificar perfiles de expresión génica de pacientes con AR, con el fin de conocer sus mecanismos moleculares. De este modo, se detectaron 339 genes expresados diferencialmente, entre los que destaca la sobreexpresión de un grupo de genes relacionados con la inmunidad y la defensa. Estos genes podrían estar sistemáticamente activados en la AR (Teixeira et al., 2009). La Figura 4 muestra la distribución de individuos en cada grupo experimental.

En este caso, el *boxplot* inicial permitió comprobar que no se había aplicado la transformación logarítmica en base 2 durante la normalización, ya que el rango de los valores de expresión oscilaba entre 85 y 60600. La Figura 12 muestra el *boxplot* tras la transformación.

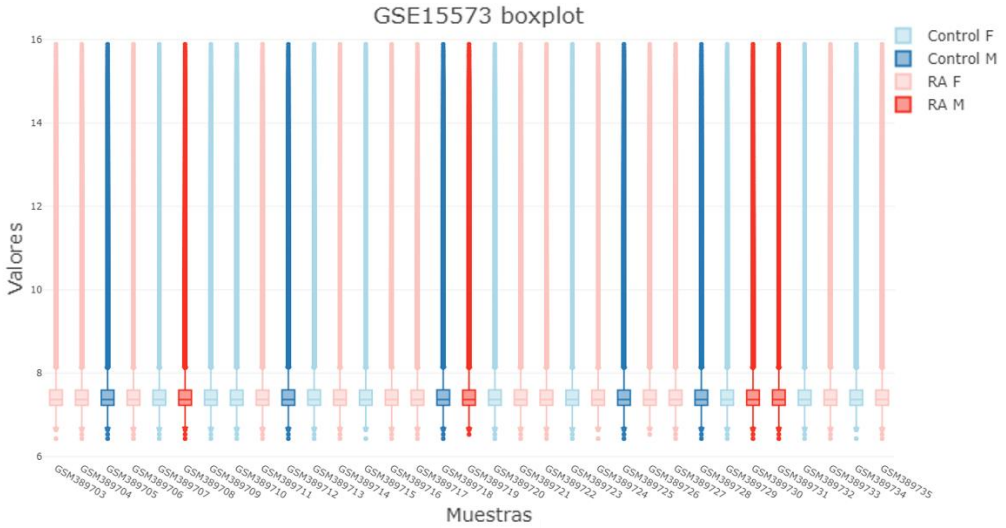

Figura 12 | Boxplot del GSE15573. Se muestran los valores de expresión normalizados (eje vertical) de cada una de las muestras (eje horizontal). Coloreado en función de los cuatro grupos experimentales: mujeres y hombres control, y mujeres y hombres con artritis reumatoide. *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

Tanto en el *clustering* (Figura 13) como en el PCA (Anexo A, Figura A.20) se observa cierta tendencia a agrupar las muestras en función de su condición (artritis reumatoide o control). En este sentido, los controles de sexo femenino parecen tener un comportamiento más consistente.

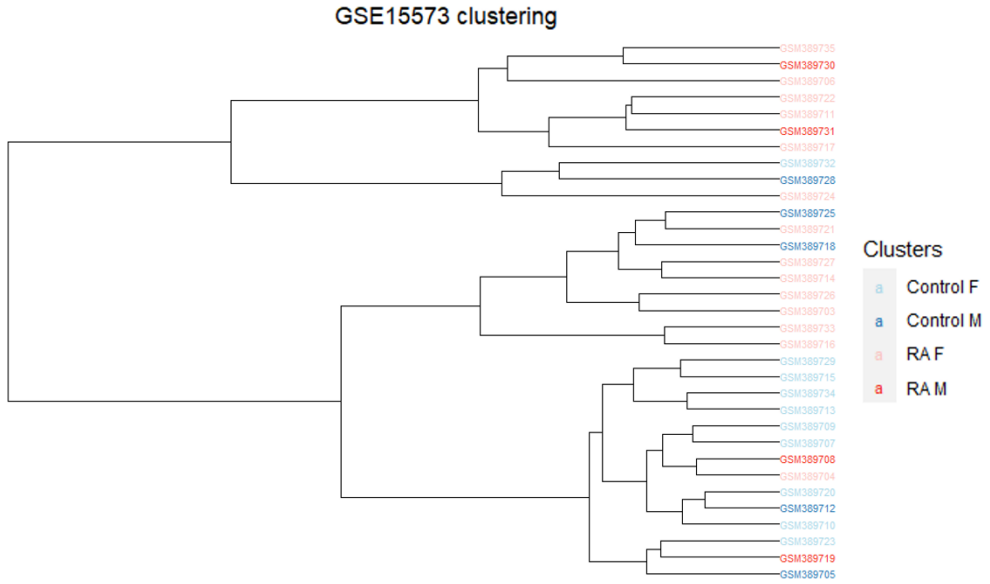

Figura 13 | *Clustering* jerárquico de las muestras del estudio GSE15573 en base a la distancia de correlación. Coloreado en función de los grupos experimentales control F y M y RA F y M. *Abreviaturas: F, mujer; M, hombre; RA, artritis reumatoide.*

En general, en los diagramas de cajas se observa que la mediana y dispersión de los valores de expresión intra-estudio son similares. Sin embargo, algunos de los PCAs y *clusterings* no muestran agrupación en función de los grupos experimentales. Esto podría deberse a la heterogeneidad de las muestras de artritis reumatoide, ya que pertenecen a pacientes con y sin tratamiento, de diferentes serotipos y estadios de la enfermedad, e incluso de distinto origen (sangre periférica y tejido sinovial).

4.2.2. Análisis de expresión diferencial

La Tabla 3 recoge el número de genes expresados diferencialmente en cada estudio a un nivel significativo, atendiendo a su P-valor ajustado por el método FDR (inferior a 0.05). Se muestra tanto para el contraste de interés (Ecuación 1), como para las comparaciones entre mujeres AR y mujeres control, y hombres AR y hombres control (Ecuación 2 y Ecuación 3, respectivamente). Los resultados se han agrupado en función del logaritmo de la magnitud de cambio (logFC) como sobreexpresados (*Up*, logFC superior a 0) o subexpresados (*Down*, logFC inferior a 0). No parece observarse ninguna tendencia clara de sobreexpresión o subexpresión en mujeres frente a hombres, ni en enfermos respecto a controles.

Tabla 3 | Resumen del análisis de expresión diferencial en cada estudio. Se muestran los resultados para las tres comparaciones: Mujer AR - Mujer Control, $♀_{AR} - ♀_{C}$; Hombre AR - Hombre Control, $♂_{AR} - ♂_{C}$; y $(♀_{AR} - ♀_{C}) - (♂_{AR} - ♂_{C})$. Los términos *Up* y *Down* reflejan el grupo experimental en el que los genes están sobrerrepresentados: *Up* en mujeres AR y *Down* en hombres AR.

	Total de genes		$(♀_{AR} - ♀_{C}) - (♂_{AR} - ♂_{C})$	$♀_{AR} - ♀_{C}$	$♂_{AR} - ♂_{C}$
GSE15573	19248	Up	0	149	0
		Down	0	267	0
GSE93272	20947	Up	0	2452	699
		Down	4	2339	408
GSE110169	19445	Up	1290	1027	3272
		Down	1235	1504	4117
GSE89408	13104	Up	1020	5476	4197
		Down	951	4496	3841
GSE117769	11464	Up	0	26	4
		Down	0	39	3

Abreviaturas: ♀, mujer; ♂, hombre; AR, artritis reumatoide; C, control; Up, sobreexpresado; Down, subexpresado.

Los principales genes expresados diferencialmente en un nivel significativo en cada uno de los estudios para el contraste de interés se muestran en las Tablas 4-6, ordenados por su logFC.

En el estudio GSE93272 (Tabla 4), los cuatro genes identificados se encuentran sobreexpresados en los hombres con AR respecto a las mujeres con AR.

Tabla 4 | Genes expresados diferencialmente en un nivel significativo en el estudio GSE93272. Se indica su identificador Entrez, nombre del gen, logaritmo de la magnitud de cambio (logFC) y su P-valor ajustado por el método Benjamini-Hochberg.

Entrez ID	Símbolo	Nombre del gen	logFC	P-valor ajustado
55320	MIS18BP1	MIS18 binding protein 1	-1.371	.025
79612	NAA16	N-alpha-acetyltransferase 16, NatA auxiliary subunit	-1.025	.049
60496	AASDHPPT	Aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase	-0.817	.025
9086	EIF1AY	Eukaryotic translation initiation factor 1A Y-linked	-0.678	.025

En los estudios GSE110169 (Tabla 5) y GSE89408 (Tabla 6) se identificaron genes sobreexpresados tanto en hombres como en mujeres. En este último, entre los genes

sobreexpresados en mujeres con AR se encontraron varios relacionados con el sistema inmune. En este sentido, cabe señalar que éste es el único estudio cuyas muestras son de tejido sinovial, en lugar de sangre periférica.

Tabla 5 | Genes expresados diferencialmente en un nivel significativo en el estudio GSE110169. Se muestra una selección de 10 de ellos, ordenados por su logaritmo de la magnitud de cambio (logFC). Se indica también su identificador Entrez, nombre del gen y su P-valor ajustado por el método Benjamini-Hochberg.

Entrez ID	Símbolo	Nombre del gen	logFC	P-valor ajustado
170482	CLEC4C	C-type lectin domain family 4 member C	1.350	.016
1675	CFD	Complement factor D	1.192	.025
4900	NRGN	Neurogranin	1.127	.018
6678	SPARC	Secreted protein acidic and cysteine rich	1.091	.045
55796	MBNL3	Muscleblind like splicing regulator 3	1.081	.029
286097	MICU3	Mitochondrial calcium uptake family member 3	-1.386	.014
339318	ZNF181	Zinc finger protein 181	-1.423	.009
51187	RSL24D1	Ribosomal L24 domain containing 1	-1.453	.009
7220	TRPC1	Transient receptor potential cation channel subfamily C member 1	-1.567	.007
55166	CENPQ	Centromere protein Q	-1.623	.005

Tabla 6 | Genes expresados diferencialmente en un nivel significativo en el estudio GSE89408. Se muestra una selección de 10, ordenados por su logaritmo de la magnitud de cambio (logFC). Se indica también su identificador Entrez, nombre del gen y su P-valor ajustado por el método Benjamini-Hochberg.

Entrez ID	Símbolo	Nombre del gen	logFC	P-valor ajustado
3820	KLRB1	Killer cell lectin like receptor B1	1.918	<.001
353345	GPR141	G protein-coupled receptor 141	1.726	<.001
151888	BTLA	B and T lymphocyte associated	1.688	0.009
100533467	BIVM-ERCC5	BIVM-ERCC5 readthrough	1.661	<.001
1233	CCR4	C-C motif chemokine receptor 4	1.598	.010
27129	HSPB7	Heat shock protein family B (small) member 7	-0.470	.024
1674	DES	Desmin	-0.498	.018
126	ADH1C	Alcohol dehydrogenase 1C (class I), gamma polypeptide	-0.544	.009
387590	TPTEP1	TPTE pseudogene 1	-0.544	.040
28999	KLF15	Kruppel like factor 15	-0.564	.006

En los estudios GSE15573 y GSE117769 no se encontraron genes expresados diferencialmente a un nivel significativo. En este último, el análisis exploratorio no mostró agrupamiento de los grupos experimentales en el PCA ni en el *clustering*, lo que podría correlacionarse con la falta de resultados significativos en el análisis de expresión diferencial. Además, los resultados de este estudio no se publicaron, por lo que es posible que la calidad de los datos sea cuestionable. En cuanto al estudio GSE15573, aunque en el *clustering* sí que mostró cierta agrupación, las muestras de AR pertenecen a pacientes en tratamiento con fármacos modificadores de la enfermedad, y tanto seropositivos como seronegativos, lo que podría haber interferido en el análisis.

4.2.3. Análisis de enriquecimiento funcional

Tras el análisis de expresión diferencial se realizó el análisis de enriquecimiento funcional de cada estudio, tanto para procesos biológicos (BP) de la GO, como para rutas KEGGs. Los resultados principales se recogen en la Tabla 7. Esta tabla muestra las funciones significativas con un P-valor ajustado por el método FDR inferior a 0.05, y las agrupa según su logaritmo de los *odds ratio* (LOR):

Up, si se encuentran sobrerrepresentadas en mujeres con AR, y *Down* si son más activas en hombres. En este caso tampoco se observa ninguna tendencia de sobreexpresión en hombres ni mujeres con AR.

Tabla 7 | Resumen del análisis de enriquecimiento funcional en cada estudio. Los términos *Up* y *Down* reflejan el grupo experimental en el que las funciones o rutas están sobrerrepresentadas: *Up* en mujeres con artritis reumatoide (AR) y *Down* en hombres con AR. Se muestra el resumen tanto para los procesos biológicos (BP) de la Gene Ontology (GO) como para las rutas KEGG.

	Términos GO BP significativos			**Rutas KEGG significativas**		
	Up	Down	Total	Up	Down	Total
GSE15573	24	22	46	6	13	19
GSE89408	31	87	118	2	2	4
GSE93272	31	291	322	8	18	26
GSE110169	109	10	119	9	1	10
GSE117769	79	15	94	22	3	25

Además, se realizaron diagramas UpSet (Conway et al., 2017) con el fin de cuantificar las funciones compartidas por los diferentes estudios (Figura 14). Estos gráficos muestran las intersecciones entre grupos y su tamaño. El gráfico de barras horizontal muestra el tamaño de cada grupo, es decir, el número total de términos identificados en cada estudio. El gráfico vertical, por su parte, indica el tamaño de la intersección de los grupos señalados por puntos en la parte inferior. Como se observa en la figura, ninguno de los términos aparece sobrerrepresentado de manera significativa en más de tres estudios.

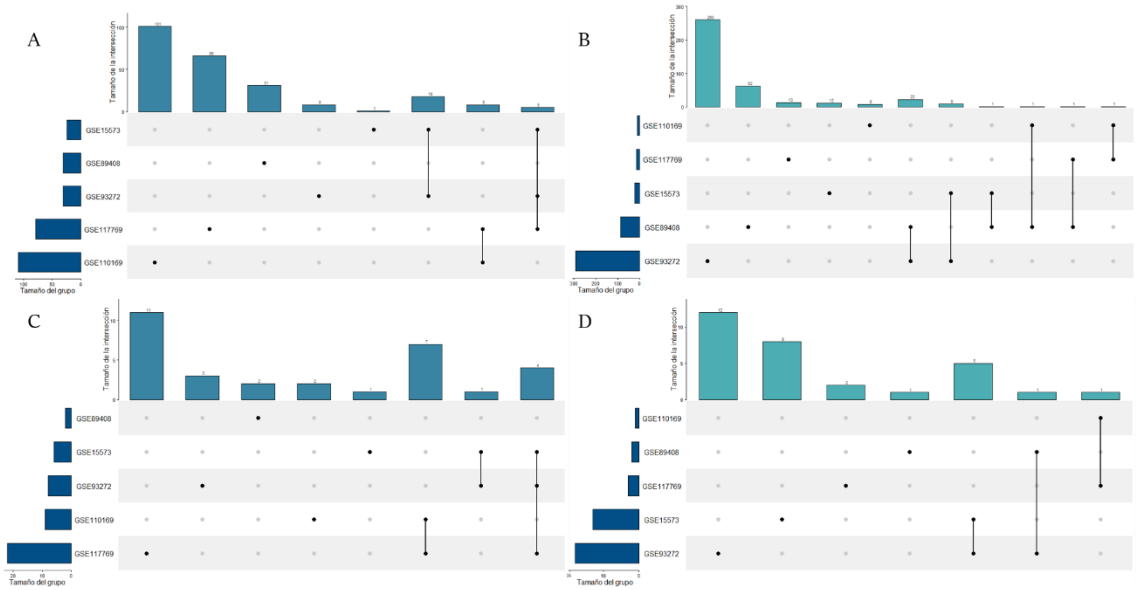

Figura 14 | Diagramas UpSet del enriquecimiento funcional. Se muestra el número de funciones compartidas por los 5 estudios (GSE15573, GSE89408, GSE93272, GSE117769 y GSE110169). El gráfico vertical indica el tamaño de la intersección, es decir, el número de funciones compartidas por los estudios indicados por puntos en la parte inferior. El gráfico horizontal representa el número total de funciones de cada estudio. A: Procesos biológicos (BP) de la Gene Ontology (GO) sobrerrepresentados en mujeres con artritis reumatoide (AR). B: Términos GO BP sobrerrepresentados en hombres con AR. C: Rutas KEGG sobrerrepresentados en mujeres con AR. D: Rutas KEGG sobrerrepresentados en hombres con AR.

4.3. Metaanálisis de genes

La Tabla 8 muestra los genes diferencialmente expresados en un nivel significativo (P-valor ajustado por el método BH inferior a 0.1) en el contraste de interés (Ecuación 1) como resultado del metaanálisis aplicando el método DerSimonian-Laird. En ella se observan 5 genes sobreexpresados en mujeres con AR y 5 en hombres. De Figura 15 a Figura 19 se muestran los gráficos de bosque y

de embudo de aquellos cuyo P-valor ajustado es inferior a 0.05. Las figuras del resto de genes se pueden encontrar en el Anexo A.

Tabla 8 | Genes diferencialmente expresados como resultado del metaanálisis. Se indica su identificador Entrez, símbolo, nombre, logaritmo de la magnitud de cambio, P-valor ajustado por el método Benjamini-Hochberg y N.

Entrez ID	Símbolo	Nombre del gen	logFC	P-valor ajustado	N
84518	CNFN	Cornifelin	0.329	.043	3
128346	C1orf162	Chromosome 1 open reading frame 162	0.241	.073	5
284185	LINC00482	Long intergenic non-protein coding RNA 482	0.206	.044	3
9986	RCE1	Ras converting CAAX endopeptidase 1	0.191	.073	5
90332	EXOC3L2	Exocyst complex component 3 like 2	0.180	.073	4
56897	WRNIP1	Werner helicase interacting protein 1	-0.082	.006	5
54946	SLC41A3	Solute carrier family 41 member 3	-0.093	.082	5
2770	GNAI1	G protein subunit alpha i1	-0.153	.094	4
728130	NUTM2D	NUT family member 2D	-0.187	.047	3
285386	TPRG1	Tumor protein p63 regulated 1	-0.197	.006	5

Abreviaturas: Entrez ID, identificador Entrez; logFC, logaritmo de la magnitud de cambio; N, número de estudios.

En las Figuras 15, 16 y 18 se puede observar que no todos los estudios han sido evaluados en el metaanálisis. Esto se debe a que la información de estos genes no estaba disponible en los estudios de microarrays. Por otra parte, los gráficos de embudo muestran que todos los estudios se encuentran dentro de la región de confianza delimitada por la zona blanca, lo que confirma la robustez del metaanálisis. Sin embargo, y aunque en ningún caso el resultado global cruza la línea de efecto nulo, en la mayoría de los genes el tamaño del efecto no es muy grande.

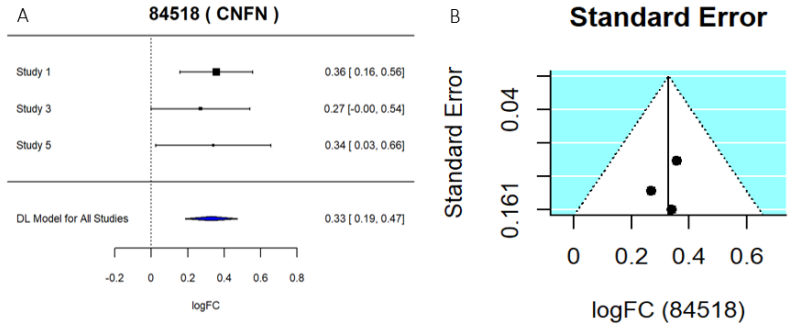

Figura 15 | Gen CNFN. A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (tres filas superiores, expresado como logFC) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos vertical marca la línea de efecto nulo. El tamaño de la caja en los estudios individuales es proporcional al tamaño del estudio. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. La zona blanca delimita la región de confianza. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

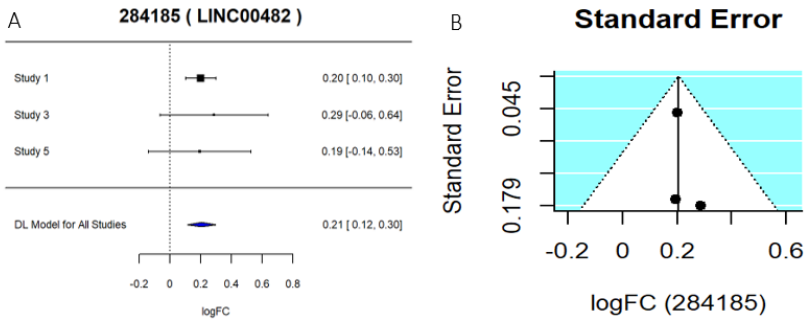

Figura 16 | Gen LINC00482: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. La zona blanca delimita la región de confianza. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

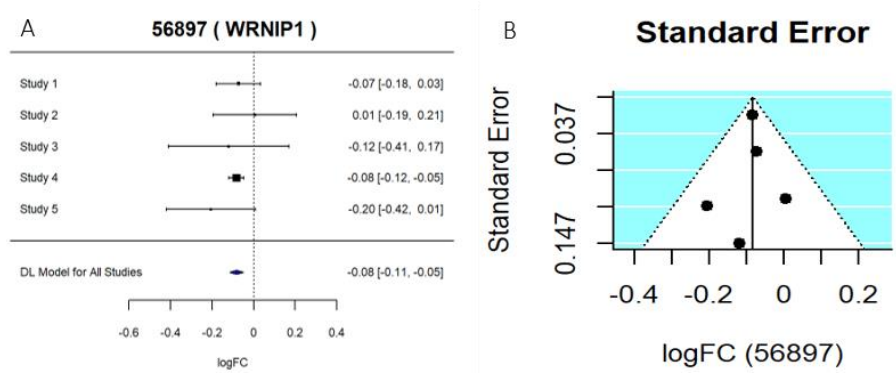

Figura 17 | Gen WRNIP1. A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. La zona blanca delimita la región de confianza. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

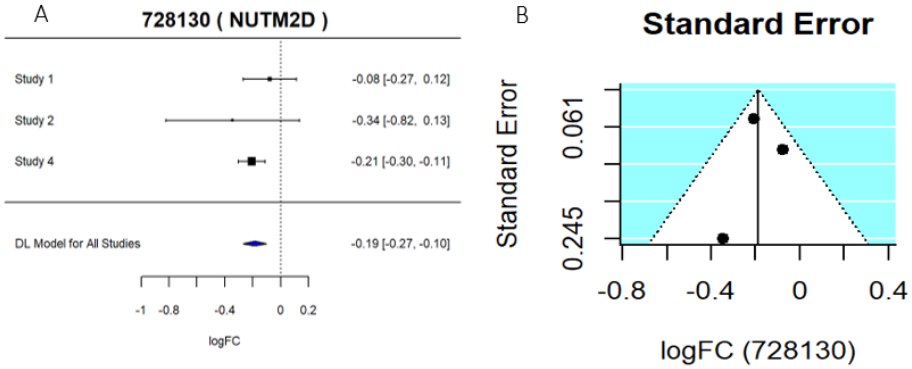

Figura 18 | Gen NUTM2D. A: Gráfico de bosque del metaanálisis. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. La zona blanca delimita la región de confianza. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

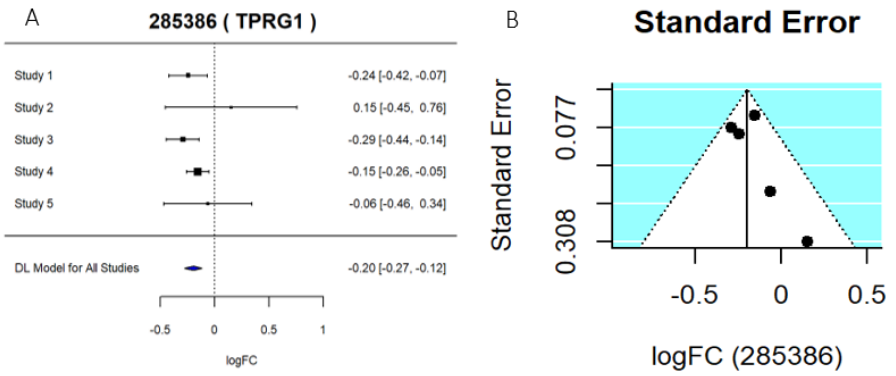

Figura 19 | Gen TPRG1. A: Gráfico de bosque del metaanálisis. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. La zona blanca delimita la región de confianza. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

En cuanto a los genes identificados, cabe señalar la sobreexpresión en hombres del *Tumor protein p63 regulated 1* (TPRG1), previamente descrito en AR, y de *G protein subunit alpha i1* (GNAI1), en enfermedades del sistema inmune. En mujeres destaca la sobreexpresión de *Exocyst complex component 3 like 2* (EXOC3L2), descrito en artritis, y la de *Chromosome 1 open reading*

frame 162 (C1orf162) y cornifelina, en enfermedades autoinmunes como el lupus eritematoso y el síndrome de Sjögren, o la psoriasis, respectivamente (Ghoussaini et al., 2021).

Finalmente, a partir de la lista de todos los genes ordenados en función de su logFC del metaanálisis, se realizó un GSEA, con el fin de enriquecer esa lista de genes consenso e identificar funciones. Al igual que en etapas previas, se realizó tanto a nivel de procesos biológicos de la GO como de rutas KEGG.

4.3.1. Términos GO para procesos biológicos (BP)

En el enriquecimiento de términos GO BP se detectaron 90 términos significativos (P-valor ajustado por el método Benjamini-Yekutieli inferior a 0.05), 52 de ellos enriquecidos en mujeres con AR y 38 en hombres. En el Anexo B se muestra una tabla con una selección de 10 de los términos enriquecidos con mayor LOR en mujeres y 10 en hombres.

Dado el elevado número de términos identificados, se emplearon dos estrategias para facilitar su interpretación. Por un lado, se determinaron los GO slims (Harris et al., 2004), conjuntos reducidos de términos GO que proporcionan una visión más amplia. Tras esta reducción, se representaron en la Figura 20. Y por otro lado, se empleó REVIGO (Supek et al., 2011), un servidor web que resume listas de términos GO mediante un algoritmo de *clustering* basado en medidas de similitud semántica (Figura 21). En ambas figuras se puede observar un gran número de procesos biológicos relacionados con el sistema inmune sobrerrepresentados en las mujeres, como el procesamiento y presentación de antígenos. También se encuentran sobrerrepresentados la generación de metabolitos precursores y energía, la respuesta al estrés y la transducción de señales. En hombres, por el contrario, se ha encontrado una mayor representación de términos relacionados con procesos metabólicos de compuestos de nitrógeno celular, el procesamiento del ARN mensajero, el transporte de proteínas y la regulación negativa de la osificación.

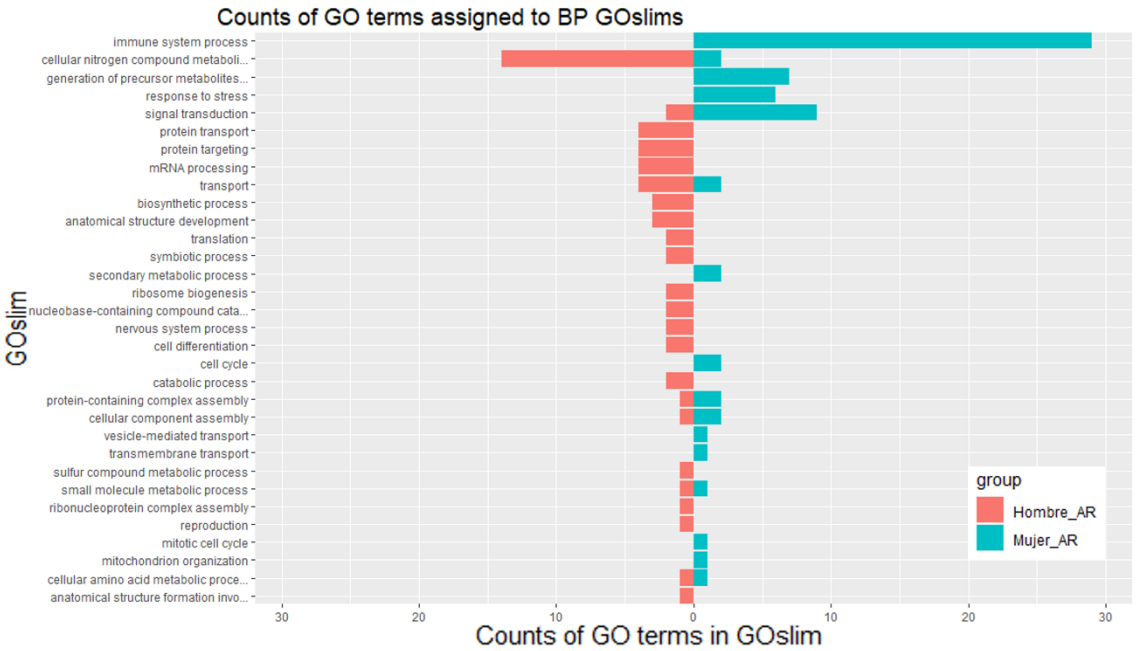

Figura 20 | Pirámide con los términos GOslim. El eje horizontal muestra el número de términos GO asociados a cada término GOslim. En rojo y azul se muestran, respectivamente, los términos sobrerrepresentados en hombres y mujeres con AR. *Abreviaturas: AR, artritis reumatoide.*

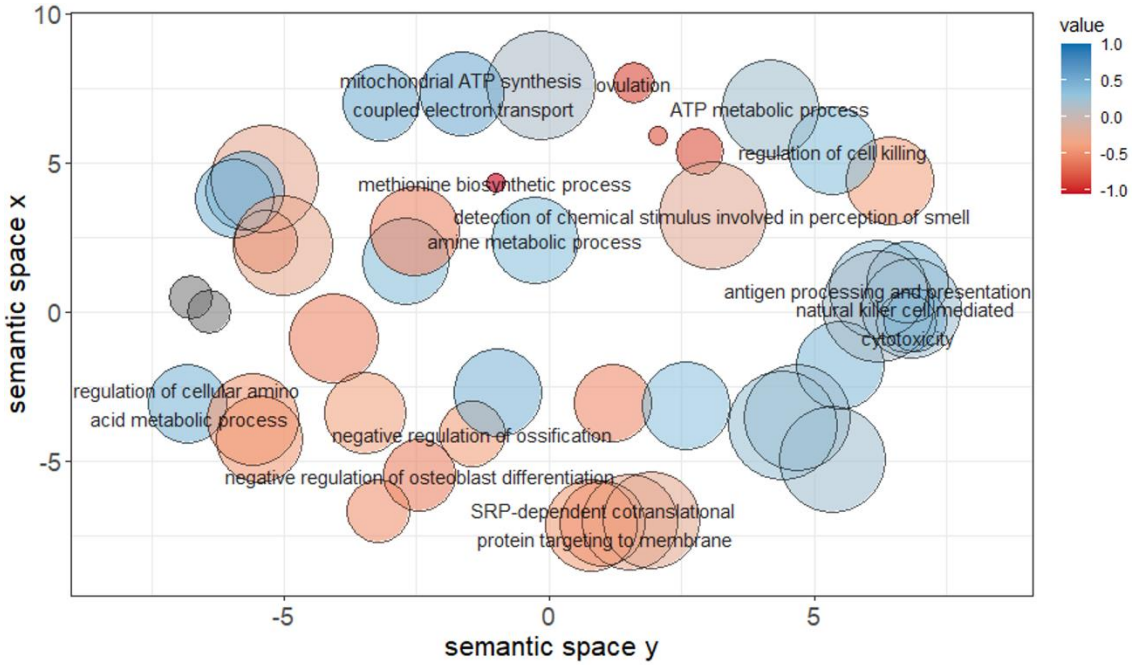

Figura 21 | Resumen de los términos GO de procesos biológicos significativos en el enriquecimiento del metaanálisis. Los ejes X e Y representan la similitud semántica de los términos GO. La escala muestra los valores de los logaritmos de los *odds ratio* (LOR): positivo/azules para los términos sobrerrepresentados en mujeres con artritis reumatoide (AR) y negativos/rojos en hombres con AR. El tamaño de la burbuja refleja la especificidad del término.

4.3.2. Rutas KEGG

En el enriquecimiento de rutas KEGG se identificaron 11 términos significativos (P-valor ajustado por el método Benjamini-Yekutieli inferior a 0.05), nueve de ellos en mujeres y dos en hombres (Tabla 9).

Tabla 9 | Términos KEGG significativos en el enriquecimiento del metaanálisis. Se indican el identificador hsa del término KEGG, su nombre, LOR (positivo en mujeres, negativo en hombres), P-valor ajustado por el método Benjamini-Yekutieli.

ID	Nombre	LOR	P-valor ajustado
5310	Asthma	0.616	.017
3050	Proteasome	0.606	.002
4672	Intestinal immune network for IgA production	0.468	.028
190	Oxidative phosphorylation	0.455	<.001
4142	Lysosome	0.423	<.001
5010	Alzheimer's disease	0.380	<.001
5012	Parkinson's disease	0.376	<.001
5016	Huntington's disease	0.333	<.001
4514	Cell adhesion molecules (CAMs)	0.289	.022
4740	Olfactory transduction	-0.166	.032
3010	Ribosome	-0.613	<.001

Abreviaturas: ID, identificador KEGG; LOR, logaritmos de los *odds ratio.*

5. DISCUSIÓN

La AR, al igual que otras enfermedades autoinmunes, presenta diferencias importantes en su prevalencia, severidad y respuesta al tratamiento entre hombres y mujeres (Platzer et al., 2019). Estas diferencias entre sexos no son únicas en este tipo de enfermedades. Sin embargo, el sexo como factor determinante en la salud está desatendido en la investigación y la literatura. Los datos de sujetos masculinos se han considerado durante mucho tiempo la norma médica y han dado lugar a una evidencia que no es totalmente replicable entre sexos (Merriman et al., 2021). Así, desigualdades de sexo en la investigación biomédica como la infrarrepresentación de las mujeres o la no inclusión de la variable sexo en los análisis están perjudicando a los pacientes, tanto a nivel de comprensión de la enfermedad como de su tratamiento (Kim et al., 2010; Fish, 2008). Por ello, considerar el sexo tanto en el proceso de investigación como en el diagnóstico, prevención y tratamiento de las enfermedades es un paso fundamental hacia la medicina de precisión, que beneficiaría la salud de hombres y mujeres (Mauvais-Jarvis et al., 2020). A pesar de ello, actualmente, gran parte de la investigación se sigue llevando a cabo sin tenerlo en cuenta, por lo que, como se ha podido comprobar en la búsqueda sistemática realizada en este trabajo, numerosos estudios carecen de dicha información o no disponen de una representación equitativa de ambos sexos. Esto, unido a la falta de estandarización de los datos públicos, dificulta su reutilización y reanálisis.

En este sentido, en 2016 se propusieron los principios FAIR, con el fin de promover la reutilización de los datos académicos haciéndolos más localizables, accesibles, interoperables y reutilizables por humanos y máquinas (Lamprecht et al., 2019). Para ello es necesario establecer ciertas normas relativas a los datos, metadatos y a las condiciones de acceso, facilitando la puesta en común de forma sistemática y transparente. Aunque los principios FAIR suponen un paso adelante necesario para impulsar la administración de los datos, requieren un trabajo continuo y han surgido problemas relacionados con la adhesión insuficiente por parte de la comunidad científica o su aplicación inadecuada (Boeckhout et al., 2018; Wilkinson et al., 2016). Así, en este trabajo, por ejemplo, ha sido necesario integrar estudios realizados en distintas plataformas que no seguían un estándar de nomenclatura en sus metadatos.

En cualquier caso, la disponibilidad de datos reutilizables es una oportunidad para extraer, integrar y analizar datos nuevos y existentes y avanzar así en los descubrimientos. El metaanálisis, en concreto, puede proporcionar una imagen más general y completa de las evidencias que cualquier estudio individual. De hecho, esta técnica proporciona un medio más potente y menos sesgado que los enfoques convencionales, incluidas las revisiones narrativas y los métodos cuantitativos como el "recuento de votos" (Gurevitch et al., 2018). En este estudio, el metaanálisis ha permitido, mediante un abordaje *in silico*, combinar los estudios individuales identificados en la búsqueda sistemática para responder a una pregunta que no había sido planteada en ellos: abordar las diferencias de sexo en la artritis reumatoide.

Aunque autores como Platzer (2019) ya han tratado este tema, la aproximación empleada en este estudio supone una mejora metodológica al respecto. A pesar de que los métodos integrativos aumentan, en principio, el poder estadístico y conllevan una mejor caracterización del sistema bajo estudio, es importante considerar la variación sistemática entre diferentes estudios. Sin embargo, el estudio de Platzer combinó los datos sin eliminar previamente estos posibles sesgos. Ignorar las variaciones sistemáticas que diferencian a los estudios heterogéneos puede producir resultados que no se distinguen estadísticamente de las conjeturas al azar (Lagani et al., 2016). Aquí se propone un método integrativo que ha demostrado ser eficaz para abordar esta cuestión: el metaanálisis

calcula los estadísticos oportunos de cada grupo de datos individualmente y a continuación combina y resume los resultados. Además, se han seleccionado rigurosamente los estudios, y la comparación entre mujeres y hombres con AR incluye todos los grupos experimentales.

Esto ha permitido identificar una sobreactivación de la respuesta inmune tanto innata como adaptativa en las mujeres, con gran número de funciones relacionadas con el procesamiento y presentación de antígenos (GO:0002478, GO:0002479, GO:0019882, GO:0019884, GO:0042590, GO:0048002 y GO:0050851). Es conocido que las mujeres muestran respuestas innatas y adaptativas más fuertes que los hombres, lo que contribuye a su susceptibilidad a enfermedades inflamatorias y autoinmunes como la artritis reumatoide (Klein y Flanagan, 2016). Tras estas diferencias parecen encontrarse las hormonas sexuales (estrógenos, progesterona y andrógenos), que intervendrían en la regulación del sistema inmune modulando la presentación de antígenos, la activación de linfocitos, la expresión de genes codificantes de citoquinas y el *homing* de células inmunes mediante receptores expresados en éstas (Whitacre, 2001).

Además, en las mujeres también se encuentra sobrerrepresentada la generación de metabolitos precursores y energía, en especial la fosforilación oxidativa (GO:0006119; GO:0006120; GO:0022900; GO:0022904; GO:0042773; GO:0042775; GO:0045333 y el término KEGG hsa:190). Esto podría correlacionarse con la mencionada activación de las respuestas inmunitarias, que requiere de una gran cantidad de energía y recursos metabólicos en las enfermedades crónicas inflamatorias como la artritis reumatoide (Klein y Flanagan, 2016). Así, la activación de las células inmunes podría estimular la glicólisis y otras rutas celulares metabólicas (Spies et al., 2012, Yang et al., 2015).

Otra ruta KEGG sobrerrepresentada en mujeres con AR es la correspondiente al proteasoma. En este sentido, se han asociado niveles elevados de inmunoproteasomas con la inflamación y el desarrollo de la autoinmunidad (Verbrugge et al., 2015). Los inmunoproteasomas facilitan la presentación de antígenos endógenos y exógenos, que se encuentran entre las funciones sobrerrepresentadas en mujeres. Además, estudios previos también encontraron proteasomas circulantes en suero de pacientes con diversas enfermedades autoinmunes, entre ellas la AR. Estos proteasomas circulantes podrían actuar como autoantígenos e inducir la respuesta autoinmune, ya que también se detectaron anticuerpos contra ellos (Sixt y Dahlmann, 2008; Feist et al., 2000). Dado que las distintas enfermedades autoinmunes parecen mostrar distintos patrones de proteasomas circulantes, podrían servir como potenciales biomarcadores. Por otra parte, a falta de más investigación, la sobrerrepresentación de proteasomas podría ser una diana terapéutica en mujeres con AR, ya que los inhibidores de proteasas han mostrado resultados prometedores en el tratamiento de enfermedades autoinmunes (Verbrugge et al., 2015). Cabe señalar que otras rutas KEGG identificadas en mujeres con AR como “Enfermedad de Alzheimer”, “Enfermedad de Parkinson” y “Enfermedad de Huntington”, aparecen relacionadas con los proteasomas (DisGeNET, 2010).

Por otra parte, en los hombres se ha identificado una sobrerrepresentación de la regulación negativa de la osificación, en concreto de la diferenciación de osteoblastos (GO:0045668, GO:0030279, GO:0030514). Frecuentemente los pacientes con AR se ven afectados por severas pérdidas óseas locales y sistémicas. Este proceso se debe tanto a un aumento de la erosión mediada por osteoclastos, como a una disminución de la formación ósea mediada por osteoblastos. Aunque en la AR ya se había descrito la inhibición por parte de las células B de la formación ósea mediante la supresión de la diferenciación de osteoblastos (Sun et al., 2018), no se conocía su mayor implicación en hombres. Sin embargo, sí que se había descrito en algunos estudios una mayor

frecuencia de enfermedad erosiva en hombres que en mujeres (Weyand et al., 1998). Por lo tanto, esta regulación negativa de los osteoblastos podría ser uno de los mecanismos subyacentes.

Entre los términos aparentemente sobrerrepresentados en hombres, llama la atención la ovulación. Sin embargo, considerando en la Ecuación 1 que este término sería igual a 0 en hombres, el signo negativo del LOR se correspondería con una mayor representación de la ovulación en mujeres control respecto a mujeres con AR. Esto podría relacionarse con la menor edad de los controles.

Otras funciones sobrerrepresentadas en hombres se relacionan con la biosíntesis de proteínas (compuestos nitrogenados celulares: GO:0000184, GO:000375, GO:0000377; traducción: GO:0002181, GO:0006413; proceso metabólico de aminoácidos: GO:0006521, GO:0009086 entre muchos otros, o la ruta KEGG hsa:3010). Estudios previos han descrito esta sobrerrepresentación en sangre periférica en pacientes cuyos tejidos mostraban un alto grado de inflamación, lo que parecía sugerir una relación con la inducción de la transcripción génica por parte de las citoquinas proinflamatorias (Van Baarsen et al., 2010). Además, la disminución de la respuesta inflamatoria por parte de los estrógenos podría explicar la diferencia entre sexos (Favalli et al., 2019). Sin embargo, otros estudios sugieren que la inflamación crónica en la AR podría favorecer el catabolismo proteico (Walsmith y Roubenoff, 2002). En cualquier caso, la inflamación desregula el metabolismo proteico, y aunque es necesaria más investigación al respecto, su expresión diferencial entre sexos podría ser una diana terapéutica.

Respecto a los genes identificados en el metaanálisis, en mujeres se ha detectado la sobreexpresión de EXOC3L2, previamente descrito en AR, y de C1orf162 y cornifelina, descritos en enfermedades autoinmunes (Ghoussaini et al., 2021). Otras secuencias sobreexpresadas en mujeres incluyen a *Long intergenic non-protein coding RNA 482* (LINC00482). Los RNAs largos no codificantes son claves en la regulación de múltiples procesos inmunológicos, como la regulación transcripcional de la inmunidad innata y adaptativa (Klein y Flanagan, 2016). En concreto, LINC00482 se ha relacionado con la inflamación en el cáncer de vejiga, aunque no se conocía su implicación en la AR (Wang et al., 2021). En hombres destaca la sobreexpresión de TPRG1, descrito en AR, y de GNAI1, en enfermedades del sistema inmune. También se ha detectado la sobreexpresión de *Werner helicase interacting protein 1* (WRNIP1), un gen relacionado con la respuesta inmune innata (Tan et al., 2017). Otros genes identificados, como *Solute carrier family 41 member 3* (SLC41A3) o *NUT family member 2D* (NUTM2D) requieren de mayor investigación para poder relacionarlos con la AR. En ninguno de estos genes se había descrito su expresión diferencial entre hombres y mujeres.

En definitiva, las funciones y genes identificados en este estudio pueden proporcionar la base para futuras investigaciones que permitirían comprender los mecanismos que subyacen a las diferencias de sexo en la AR. Así, este mayor conocimiento de la enfermedad contribuiría a un enfoque adaptado a la medicina personalizada y de precisión.

Por otra parte, dada la falta de información de la variable sexo en varios de los estudios identificados en la revisión sistemática, de cara al futuro sería interesante realizar una imputación del sexo para incorporar dichos estudios al metaanálisis. Así, con un número mayor de estudios disponible, se podría valorar la selección de estudios de un único tipo de muestra, y considerar variables como el subtipo de AR, la fase o el tratamiento.

6. CONCLUSIONES

1. La metodología empleada en este estudio es una aproximación robusta y eficaz que, tras la revisión sistemática de las bases de datos GEO, ArrayExpress y ADEx, permite evaluar e integrar los datos de diferentes estudios de transcriptómica, contribuyendo así a la reutilización de datos según dictan los principios FAIR.
2. La revisión sistemática realizada en este estudio demuestra la falta de estandarización en los estudios biomédicos y, en especial, la falta de inclusión de la variable sexo. Su consideración es, además, fundamental para obtener resultados replicables en ambos sexos.
3. Se han identificado 10 biomarcadores de AR específicos de sexo: 5 en mujeres y 5 en hombres.
4. Se han identificado 90 procesos biológicos GO específicos de sexo (52 en mujeres y 38 en hombres) y 11 rutas KEGG (9 en mujeres y 2 en hombres).
5. Entre las funciones sobrerrepresentadas en mujeres destaca la activación del sistema inmune y la generación de metabolitos precursores y energía. En hombres cabe señalar la regulación negativa de los osteoblastos y la biosíntesis de proteínas.
6. Este estudio ha permitido caracterizar los mecanismos que subyacen a las diferencias de sexo en la AR, contribuyendo a la aproximación de la investigación biomédica hacia una medicina personalizada y de precisión.

7. BIBLIOGRAFÍA

ABBAS-AGHABABAZADEH, F.; LI, Q. & FRIDLEY, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS ONE* 13:1–21.

ALLISON, D. B.; CUI, X.; PAGE, G. P. & SABRIPOUR, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics* 7:55–65.

ALTMAN, N. & KRZYWINSKI, M. (2017). Points of Significance: Clustering. *Nature Methods* 14:545–546.

ARLOTH, J.; BADER, D. M.; RÖH, S. & ALTMANN, A. (2015). Re-Annotator: Annotation pipeline for microarray probe sequences. *PLoS ONE* 10:1–13.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPI, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M. & SHERLOCK, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:25–29.

ATHAR, A.; FÜLLGRABE, A.; GEORGE, N.; IQBAL, H.; HUERTA, L.; ALI, A.; SNOW, C.; FONSECA, N. A.; PETRYSZAK, R.; PAPATHEODOROU, I.; SARKANS, U. & BRAZMA, A. (2019). ArrayExpress update - From bulk to single-cell expression data. *Nucleic Acids Research* 47:D711–D715.

AVRAMEAS, S. & SELMI, C. (2013). Natural autoantibodies in the physiology and pathophysiology of the immune system. *Journal of Autoimmunity* 41:46–49.

BANCHEREAU, R.; CEPIKA, A.-M.; BANCHEREAU, J. & PASCUAL, V. (2017). Understanding Human Autoimmunity and Autoinflammation Through Transcriptomics. *Annual review of immunology* 35:337–370.

BARRETT, T.; WILHITE, S. E.; LEDOUX, P.; EVANGELISTA, C.; KIM, I. F.; TOMASHEVSKY, M.; MARSHALL, K. A.; PHILLIPPY, K. H.; SHERMAN, P. M.; HOLKO, M.; YEFANOV, A.; LEE, H.; ZHANG, N.; ROBERTSON, C. L.; SEROVA, N.; DAVIS, S. & SOBOLEVA, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research* 41:D991.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57:289–300.

BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *The annals of Statistics* 29:1165–1188.

BOECKHOUT, M.; ZIELHUIS, G. A. & BREDENOORD, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? *European Journal of Human Genetics* 26:931–936.

BURMESTER, G. R. & POPE, J. E. (2017). Novel treatment strategies in rheumatoid arthritis. *The Lancet* 389:2338–2348.

CANNATA, N.; MERELLI, E. & ALTMARE, R. B. (2005). Time to organize the bioinformatics resourceome. *PLoS Computational Biology* 1:0531–0533.

CARBON, S.; DOUGLASS, E.; GOOD, B. M.; UNNI, D. R.; HARRIS, N. L.; MUNGALL, C. J.; BASU, S.; CHISHOLM, R. L.; DODSON, R. J.; HARTLINE, E.; FEY, P.; THOMAS, P. D.; ALBOU, L. P.; EBERT, D.; KESLING, M. J.; MI, H.; MURUGANUJAN, A.; HUANG, X.; MUSHAYAHAMA, T.; ... & ELSER, J. (2021). The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research* 49:D325–D334.

CARLSON, M. (2016a). hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation

data (chip hgu133plus2). *R package version 3.2.3*.

CARLSON, M. (2016b). hgu219.db: Affymetrix Human Genome 219 Plate annotation data (chip hgu219). *R package version 3.2.3*.

CARLSON, M. (2019). org.Hs.eg.db: Genome wide annotation for Human. *R package version 3.8.2*.

CHEN, J.; WRIGHT, K.; DAVIS, J. M.; JERALDO, P.; MARIETTA, E. V.; MURRAY, J.; NELSON, H.; MATTESON, E. L. & TANEJA, V. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Medicine* 8:43.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZEŚNIAK, M. W.; GAFFNEY, D. J.; ELO, L. L.; ZHANG, X. & MORTAZAVI, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology* 17:1–19.

CONWAY, J. R.; LEX, A. & GEHLENBORG, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33:2938–2940.

COSTA-SILVA, J.; DOMINGUES, D. & LOPES, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* 12:1–18.

DERSIMONIAN, R. & LAIRD, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* 45:139–145. Elsevier B.V.

DUNNING, M.; LYNCH, A. & ELDRIDGE, M. (2015). illuminaHumanv2.db: Illumina HumanWG6v2 annotation data (chip illuminaHumanv2). *R package version 1.26.0*.

DURINCK, S.; SPELLMAN, P. T.; BIRNEY, E. & HUBER, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* 4:1184–11191.

EVANS, C.; HARDIN, J. & STOEBEL, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics* 19:776–792.

FAVALLI, E. G.; BIGGIOGGERO, M.; CROTTI, C.; BECCIOLINI, A.; RAIMONDO, M. G. & MERONI, P. L. (2019). Sex and Management of Rheumatoid Arthritis. *Clinical Reviews in Allergy and Immunology* 56:333–345.

FEIST, E.; KUCKELKORN, U.; BURMESTER, G. & KLOETZEL, P. (2000). Diagnostic Importance of Anti-Proteasome Antibodies. *International archives of allergy and immunology* 123:92–97.

FISH, E. N. (2008). The X-files in immunity: sex-based differences predispose immune responses. *Nature Reviews Immunology* 8:737–744.

FRANK-BERTONCELJ, M.; TRENKMANN, M.; KLEIN, K.; KAROUZAKIS, E.; REHRAUER, H.; BRATUS, A.; KOLLING, C.; ARMAKA, M.; FILER, A.; MICHEL, B. A.; GAY, R. E.; BUCKLEY, C. D.; KOLLIAS, G.; GAY, S. & OSPELT, C. (2017). Epigenetically-driven anatomical diversity of synovial fibroblasts guides joint-specific fibroblast functions. *Nature Communications* 8.

GARCÍA DE YÉBENES, M. y LOZA, E. (2018). Artritis reumatoide: epidemiología e impacto sociosanitario. *Reumatología Clínica* 14:3–6.

GHOUSSAINI, M.; MOUNTJOY, E.; CARMONA, M.; PEAT, G.; SCHMIDT, E. M.; HERCULES, A.; FUMIS, L.; MIRANDA, A.; CARVALHO-SILVA, D.; BUNIELLO, A.; BURDETT, T.; HAYHURST, J.; BAKER, J.; FERRER, J.; GONZALEZ-URIARTE, A.; JUPP, S.; KARIM, M. A.; KOSCIELNY, G.; MACHLITT-NORTHEN, S.; … & DUNHAM, I. (2021). Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research* 49:D1311–D1320.

GONZALEZ, A.; KREMERS, H. M.; CROWSON, C. S.; NICOLA, P. J.; DAVIS, J. M.; THERNEAU, T. M.; ROGER, V. L. & GABRIEL, S. E. (2007). The widening mortality gap between rheumatoid arthritis

patients and the general population. *Arthritis and Rheumatism* 56:3583–3587.

GREGERSEN, P. K.; SILVER, J. & WINCHESTER, R. J. (1987). The Shared Epitope Hypothesis. *Arthritis and Rheumatism* 30:1205–1212.

GUREVITCH, J.; KORICHEVA, J.; NAKAGAWA, S. & STEWART, G. (2018). Meta-analysis and the science of research synthesis. *Nature* 555:175–182.

HAJISHENGALLIS, G. (2015). Periodontitis: From microbial immune subversion to systemic inflammation.

HANG, L. & NAKAMURA, R. M. (1997). Current concepts and advances in clinical laboratory testing for autoimmune diseases. *Critical Reviews in Clinical Laboratory Sciences* 34:275–311.

HARRIS, M. A.; CLARK, J.; IRELAND, A.; LOMAX, J.; ASHBURNER, M.; FOULGER, R.; EILBECK, K.; LEWIS, S.; MARSHALL, B.; MUNGALL, C.; RICHTER, J.; RUBIN, G. M.; BLAKE, J. A.; BULT, C.; DOLAN, M.; DRABKIN, H.; EPPIG, J. T.; HILL, D. P.; NI, L.; … & WHITE, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32:258–261.

HENRY, V. J.; BANDROWSKI, A. E.; PEPIN, A. S.; GONZALEZ, B. J. & DESFEUX, A. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database : the journal of biological databases and curation* 2014:1–5.

HIGGINS, J. P. T. (2019). *Cochrane handbook for systematic reviews of interventions*. P. (J. P. T. Higgins, Ed.) (Second edition.).

HORNIK, K. (2009). Conducting Meta-Analyses in R with the metafor Package. *Journal Of Statistical Software* 31.

HOWE, K. L.; ACHUTHAN, P.; ALLEN, J.; ALLEN, J.; ALVAREZ-JARRETA, J.; RIDWAN AMODE, M.; ARMEAN, I. M.; AZOV, A. G.; BENNETT, R.; BHAI, J.; BILLIS, K.; BODDU, S.; CHARKHCHI, M.; CUMMINS, C.; DA RIN FIORETTO, L.; DAVIDSON, C.; DODIYA, K.; EL HOUDAIGUI, B.; FATIMA, R.; … & FLICEK, P. (2021). Ensembl 2021. *Nucleic Acids Research* 49:D884–D891.

HU, Y.; CARMAN, J. A.; HOLLOWAY, D.; KANSAL, S.; FAN, L.; GOLDSTINE, C.; LEE, D.; SOMERVILLE, J. E.; LATEK, R.; TOWNSEND, R.; JOHNSEN, A.; CONNOLLY, S.; BANDYOPADHYAY, S.; SHADICK, N.; WEINBLATT, M. E.; FURIE, R. & NADLER, S. G. (2018). Development of a Molecular Signature to Monitor Pharmacodynamic Responses Mediated by In Vivo Administration of Glucocorticoids. *Arthritis and Rheumatology* 70:1331–1342.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S.; BRAVO, H. C.; DAVIS, S.; GATTO, L.; GIRKE, T.; GOTTARDO, R.; HAHNE, F.; HANSEN, K. D.; IRIZARRY, R. A.; LAWRENCE, M.; LOVE, M. I.; MACDONALD, J.; OBENCHAIN, V.; OLEŠ, A. K.; … & MORGAN, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 12:115–121.

INTEGRATIVE BIOMEDICAL INFORMATICS GROUP GRIB/IMIM/UPF. (2010). DisGeNET v7.0, visto el 6 de junio de 2021, https://www.disgenet.org/.

KANEHISA, M.; FURUMICHI, M.; SATO, Y.; ISHIGURO-WATANABE, M. & TANABE, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research* 49:D545–D551.

KANEHISA, M. & GOTO, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28:27–30.

KIM, A. M.; TINGEN, C. M. & WOODRUFF, T. K. (2010). Sex bias in trials and treatment must end. *Nature* 465:688–689.

KIM, J. R. & KIM, H. A. (2020). Molecular mechanisms of sex-related differences in arthritis and associated pain. *International Journal of Molecular Sciences* 21:1–21.

7. BIBLIOGRAFÍA

KLEIN, S. L. & FLANAGAN, K. L. (2016). Sex differences in immune responses. *Nature Reviews Immunology* 16:626–638.

LAGANI, V.; KAROZOU, A. D.; GOMEZ-CABRERO, D.; SILBERBERG, G. & TSAMARDINOS, I. (2016). A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions. *BMC Bioinformatics* 17.

LAMPRECHT, A.-L.; GARCIA, L.; KUZAK, M.; MARTINEZ, C.; ARCILA, R.; MARTIN DEL PICO, E.; DOMINGUEZ DEL ANGEL, V.; VAN DE SANDT, S.; ISON, J.; MARTINEZ, P. A.; MCQUILTON, P.; VALENCIA, A.; HARROW, J.; PSOMOPOULOS, F.; GELPI, J. L.; CHUE HONG, N.; GOBLE, C. & CAPELLA-GUTIERREZ, S. (2019). Towards FAIR principles for research software. *Data Science* 3:37–59.

LANGAN, D.; HIGGINS, J. P. T.; JACKSON, D.; BOWDEN, J.; VERONIKI, A. A.; KONTOPANTELIS, E.; VIECHTBAUER, W. & SIMMONDS, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods* 10:83–98.

LAW, C. W.; CHEN, Y.; SHI, W. & SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15:1–17.

LAZAR, C.; MEGANCK, S.; STEENHOFF, D.; COLETTA, A.; MOLTER, C.; WEISS-SOL, D. Y.; DUQUE, R.; BERSINI, H. & NOWE, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics* 14:469–490.

LEE, H. M.; SUGINO, H.; AOKI, C. & NISHIMOTO, N. (2011). Underexpression of mitochondrial-DNA encoded ATP synthesis-related genes and DNA repair genes in systemic lupus erythematosus. *Arthritis Research and Therapy* 13.

LEVER, J.; KRZYWINSKI, M. & ALTMAN, N. (2017). Points of Significance: Principal component analysis. *Nature Methods* 14:641–642.

LEWIS, S. & CLARKE, M. (2001). Forest plots: trying to see the wood and the trees. *BMJ* 322:1479–1480.

LIBERATI, A.; ALTMAN, D. G.; TETZLAFF, J.; MULROW, C.; GØTZSCHE, P. C.; IOANNIDIS, J. P. A.; CLARKE, M.; DEVEREAUX, P. J.; KLEIJNEN, J. & MOHER, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Pp. e1–e34 *Journal of clinical epidemiology*.

LIGHTBODY, G.; HABERLAND, V.; BROWNE, F.; TAGGART, L.; ZHENG, H.; PARKES, E. & BLAYNEY, J. K. (2019). Review of applications of high-throughput sequencing in personalized medicine: Barriers and facilitators of future progress in research and clinical application. *Briefings in Bioinformatics* 20:1795–1811.

LIU, Y.; ARYEE, M. J.; PADYUKOV, L.; DANIELE FALLIN, M.; HESSELBERG, E.; RUNARSSON, A.; REINIUS, L.; ACEVEDO, N.; TAUB, M.; RONNINGER, M.; SHCHETYNSKY, K.; SCHEYNIUS, A.; KERE, J.; ALFREDSSON, L.; KLARESKOG, L.; EKSTR, T. J. & FEINBERG, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology* 31:142.

LJUNG, L. & RANTAPÄÄ-DAHLQVIST, S. (2016). Abdominal obesity, gender and the risk of rheumatoid arthritis - a nested case-control study. *Arthritis Research and Therapy* 18:277.

LORTIE, C. J. & FILAZZOLA, A. (2020). A contrast of meta and metafor packages for meta-analyses in R. *Ecology and Evolution* 10:10916–10921.

LOWE, R.; SHIRLEY, N.; BLEACKLEY, M.; DOLAN, S. & SHAFEE, T. (2017). Transcriptomics technologies. *PLoS Computational Biology* 13:1–23.

MACGREGOR, A. J.; SNIEDER, H.; RIGBY, A. S.; KOSKENVUO, M.; KAPRIO, J.; AHO, K. & SILMAN, A. J. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data

from twins. *Arthritis and Rheumatism* 43:30–37.

MAGLOTT, D.; OSTELL, J.; PRUITT, K. D. & TATUSOVA, T. (2011). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research* 39:52–57.

MALMSTRÖM, V.; CATRINA, A. I. & KLARESKOG, L. (2016). The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting. *Nature Reviews Immunology* 17(1).

MANZONI, C.; KIA, D. A.; VANDROVCOVA, J.; HARDY, J.; WOOD, N. W.; LEWIS, P. A. & FERRARI, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics* 19:286–302.

MARTIN, P.; MCGOVERN, A.; OROZCO, G.; DUFFUS, K.; YARWOOD, A.; SCHOENFELDER, S.; COOPER, N. J.; BARTON, A.; WALLACE, C.; FRASER, P.; WORTHINGTON, J. & EYRE, S. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature Communications* 6:1–7.

MARTORELL-MARUGÁN, J.; LÓPEZ-DOMÍNGUEZ, R.; GARCÍA-MORENO, A.; TORO-DOMÍNGUEZ, D.; VILLATORO-GARCÍA, J. A.; BARTUREN, G.; MARTÍN-GÓMEZ, A.; TROULE, K.; GÓMEZ-LÓPEZ, G.; AL-SHAHROUR, F.; PEÑA-CHILET, M.; DOPAZO, J.; GONZÁLEZ-RUMAYOR, V.; ALARCÓN-RIQUELME, M. E. & CARMONA-SÁEZ, P. (2020). A comprehensive and centralized database for exploring omics data in autoimmune diseases. *bioRxiv*:1–24.

MAUVAIS-JARVIS, F.; BAIREY MERZ, N.; BARNES, P. J.; BRINTON, R. D.; CARRERO, J. J.; DEMEO, D. L.; DE VRIES, G. J.; EPPERSON, C. N.; GOVINDAN, R.; KLEIN, S. L.; LONARDO, A.; MAKI, P. M.; MCCULLOUGH, L. D.; REGITZ-ZAGROSEK, V.; REGENSTEINER, J. G.; RUBIN, J. B.; SANDBERG, K. & SUZUKI, A. (2020). Sex and gender: modifiers of health, disease, and medicine. *The Lancet* 396:565–582.

MAYNARD, C.; MIKULS, T. R.; CANNON, G. W.; ENGLAND, B. R.; CONAGHAN, P. G.; ØSTERGAARD, M.; BAKER, D. G.; KERR, G.; GEORGE, M. D.; BARTON, J. L. & BAKER, J. F. (2020). Sex Differences in the Achievement of Remission and Low Disease Activity in Rheumatoid Arthritis. *Arthritis Care and Research* 72:326–333.

MENG, W.; ZHU, Z.; JIANG, X.; TOO, C.L.; UEBE, S.; JAGODIC, M.; KOCKUM, I.; MURAD, S.; FERRUCCI, L.; ALFREDSSON, L.; ZOU, H.; KLARESKOG, L.; FEINBERG, A. P.; EKSTRÖM, T. J.; PADYUKOV, L. & LIU, Y. (2017). DNA methylation mediates genotype and smoking interacton in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis. *Arthritis Research and Therapy* 19(1)61:71

MERRIMAN, R.; GALIZIA, I.; TANAKA, S.; SHEFFEL, A.; BUSE, K. & HAWKES, S. (2021). The gender and geography of publishing: a review of sex/gender reporting and author representation in leading general medical and global health journals. *BMJ Global Health* 6:e005672.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G.; ALTMAN, D.; ANTES, G.; ATKINS, D.; BARBOUR, V.; BARROWMAN, N.; BERLIN, J. A.; CLARK, J.; CLARKE, M.; COOK, D.; D'AMICO, R.; DEEKS, J. J.; DEVEREAUX, P. J.; DICKERSIN, K.; EGGER, M.; ERNST, E.; ... & TUGWELL, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6.

MONTANER, D. & DOPAZO, J. (2010). Multidimensional gene set analysis of genomic data. *PLoS ONE* 5.

MOULTON, V. R. (2018). Sex hormones in acquired immunity and autoimmune disease. *Frontiers in Immunology* 9:1–21.

NEKRUTENKO, A. & TAYLOR, J. (2012). Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. *Nature Reviews Genetics* 13:667–672.

7. BIBLIOGRAFÍA

OKADA, Y.; WU, D.; TRYNKA, G.; RAJ, T.; TERAO, C.; IKARI, K.; KOCHI, Y.; OHMURA, K.; SUZUKI, A.; YOSHIDA, S.; GRAHAM, R. R.; MANOHARAN, A.; ORTMANN, W.; BHANGALE, T.; DENNY, J. C.; CARROLL, R. J.; EYLER, A. E.; GREENBERG, J. D.; KREMER, J. M.; ... & PLENGE, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506:376–381.

PHIPSON, B.; LEE, S.; MAJEWSKI, I. J.; ALEXANDER, W. S. & SMYTH, G. (2016). Robust Hyperparameter Estimation Protects. *The Annals of Applied Statistics* 10:946–963.

PLATZER, A.; NUSSBAUMER, T.; KARONITSCH, T.; SMOLEN, J. S. & ALETAHA, D. (2019). Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns. *PLoS ONE* 14:1–23.

R CORE TEAM. (2020). R: A language and environment for statistical computing. Version 4.0. 2 (Taking Off Again). *R Foundation for Statistical Computing, Vienna, Austria.*

REUTER, J. A.; SPACEK, D. V. & SNYDER, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell* 58:586–597.

RITCHIE, M. E.; PHIPSON, B.; WU, D.; HU, Y.; LAW, C. W.; SHI, W. & SMYTH, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43:e47.

ROBINSON, M. D.; MCCARTHY, D. J. & SMYTH, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

ROBINSON, M. D. & OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11.

SALINAS, G. F.; BRAZA, F.; BROUARD, S.; TAK, P. P. & BAETEN, D. (2013). The role of B lymphocytes in the progression from autoimmunity to autoimmune disease. *Clinical Immunology* 146:34–45.

SEAN, D. & MELTZER, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23:1846–1847.

SEDGWICK, P. (2015). How to read a forest plot in a meta-analysis. *BMJ (Clinical research ed.)* 351.

SEYEDNASROLLAH, F.; LAIHO, A. & ELO, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* 16:59–70.

SIEVERT, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. *Chapman and Hall/CRC Florida*.

SIXT, S. U. & DAHLMANN, B. (2008). Extracellular, circulating proteasomes and ubiquitin - Incidence and relevance. *Biochimica et Biophysica Acta - Molecular Basis of Disease* 1782:817–823.

SMOLEN, J. S.; ALETAHA, D.; BARTON, A.; BURMESTER, G. R.; EMERY, P.; FIRESTEIN, G. S.; KAVANAUGH, A.; MCINNES, I. B.; SOLOMON, D. H.; STRAND, V. & YAMAMOTO, K. (2018). Rheumatoid arthritis. *Nature Reviews Disease Primers* 4:1–23.

SMYTH, G. K.; RITCHIE, M. & THORNE, N. (2011). Linear Models for Microarray Data User ' s Guide. *Bioinformatics* 20:3705–3706.

SPIES, C. M.; STRAUB, R. H. & BUTTGEREIT, F. (2012). Energy metabolism and rheumatic diseases: From cell to organism. *Arthritis Research and Therapy* 14:1–10.

STAFFORD, I. S.; KELLERMANN, M.; MOSSOTTO, E.; BEATTIE, R. M.; MACARTHUR, B. D. & ENNIS, S. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *npj Digital Medicine* 3. Springer US.

STERNE, J. A. C.; SUTTON, A. J.; IOANNIDIS, J. P. A.; TERRIN, N.; JONES, D. R.; LAU, J.; CARPENTER, J.; RÜCKER, G.; HARBORD, R. M.; SCHMID, C. H.; TETZLAFF, J.; DEEKS, J. J.; PETERS, J.;

MACASKILL, P.; SCHWARZER, G.; DUVAL, S.; ALTMAN, D. G.; MOHER, D. & HIGGINS, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ (Online)* 343:1–8.

STOLT, P.; KÄLLBERG, H.; LUNDBERG, I.; SJÖGREN, B.; KLARESKOG, L.; ALFREDSSON, L. & STOLT, P. (2005). Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Annals of the Rheumatic Diseases* 64:582–586.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S. & MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102:15545–15550.

SUN, W.; MEEDNU, N.; ROSENBERG, A.; RANGEL-MORENO, J.; WANG, V.; GLANZMAN, J.; OWEN, T.; ZHOU, X.; ZHANG, H.; BOYCE, B. F.; ANOLIK, J. H. & XING, L. (2018). B cells inhibit bone formation in rheumatoid arthritis by suppressing osteoblast differentiation. *Nature Communications* 9.

SUPEK, F.; BOŠNJAK, M.; ŠKUNCA, N. & ŠMUC, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6.

SVENDSEN, A. J.; GERVIN, K.; LYLE, R.; CHRISTIANSEN, L.; KYVIK, K.; JUNKER, P.; NIELSEN, C.; HOUEN, G. & TAN, Q. (2016). Differentially methylated DNA regions in monozygotic twin pairs discordant for rheumatoid arthritis: An epigenome-wide study. *Frontiers in Immunology* 7:17.

TAN, P.; HE, L.; CUI, J.; QIAN, C.; CAO, X.; LIN, M.; ZHU, Q.; LI, Y.; XING, C.; YU, X.; WANG, H. Y. & WANG, R. F. (2017). Assembly of the WHIP-TRIM14-PPP6C Mitochondrial Complex Promotes RIG-I-Mediated Antiviral Signaling. *Molecular Cell* 68:293-307.e5.

TARCA, A. L.; ROMERO, R. & DRAGHICI, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology* 195:373–388.

TASAKI, S.; SUZUKI, K.; KASSAI, Y.; TAKESHITA, M.; MUROTA, A.; KONDO, Y.; ANDO, T.; NAKAYAMA, Y.; OKUZONO, Y.; TAKIGUCHI, M.; KURISU, R.; MIYAZAKI, T.; YOSHIMOTO, K.; YASUOKA, H.; YAMAOKA, K.; MORITA, R.; YOSHIMURA, A.; TOYOSHIBA, H. & TAKEUCHI, T. (2018). Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nature Communications* 9.

TEIXEIRA, V. H.; OLASO, R.; MARTIN-MAGNIETTE, M. L.; LASBLEIZ, S.; JACQ, L.; OLIVEIRA, C. R.; HILLIQUIN, P.; GUT, I.; CORNELIS, F. & PETIT-TEIXEIRA, E. (2009). Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PLoS ONE* 4.

TOBÓN, G. J.; YOUINOU, P. & SARAUX, A. (2010). The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. *Autoimmunity Reviews* 9:A288–A292.

URRUTIA, G. & BONFILL, X. (2010). PRISMA_Spanish.pdf. *Medicina Clínica* 135:507–511.

VAN BAARSEN, L. G. M.; WIJBRANDTS, C. A.; TIMMER, T. C. G.; VAN DER POUW KRAAN, T. C. T. M.; TAK, P. P. & VERWEIJ, C. L. (2010). Synovial tissue heterogeneity in rheumatoid arthritis in relation to disease activity and biomarkers in peripheral blood. *Arthritis and Rheumatism* 62:1602–1607.

VERBRUGGE, E. E.; SCHEPER, R. J.; LEMS, W. F.; DE GRUIJL, T. D. & JANSEN, G. (2015). Proteasome inhibitors as experimental therapeutics of autoimmune diseases. *Arthritis Research and Therapy* 17:1–10.

VIECHTBAUER, W. (2020). metafor: Meta-analysis package for R. R package version 2.4-0. *R package*

version 2.4-0:1–275.

WALSH, A. M.; WECHALEKAR, M. D.; GUO, Y.; YIN, X.; WEEDON, H.; PROUDMAN, S. M.; SMITH, M. D. & NAGPAL, S. (2017). Triple DMARD treatment in early rheumatoid arthritis modulates synovial T cell activation and plasmablast/plasma cell differentiation pathways. *PLoS ONE* 12.

WALSMITH, J. & ROUBENOFF, R. (2002). Cachexia in rheumatoid arthritis. *International Journal of Cardiology* 85:89–99.

WANG, L.; WANG, F. S. & GERSHWIN, M. E. (2015). Human autoimmune diseases: A comprehensive update. *Journal of Internal Medicine* 278:369–395.

WANG, Y.; ZHANG, L.; WEI, N.; SUN, Y.; PAN, W. & CHEN, Y. (2021). Silencing LINC00482 inhibits tumor-associated inflammation and angiogenesis through down-regulation of MMP-15 via FOXA1 in bladder cancer. *Aging* 13:2264–2278.

WEYAND, C. M.; SCHMIDT, D.; WAGNER, U. & GORONZY, J. J. (1998). The influence of sex on the phenotype of rheumatoid arthritis. *Arthritis and Rheumatism* 41:817–822.

WHITACRE, C. C. (2001). Sex differences in autoimmune disease. *Nature immunology* 2:777–780.

WICKHAM, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, IJ. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J. W.; DA SILVA SANTOS, L. B.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; ... & MONS, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:1–9.

YANG, X. Y.; ZHENG, K. DI; LIN, K.; ZHENG, G.; ZOU, H.; WANG, J. M.; LIN, Y. Y.; CHUKA, C. M.; GE, R. S.; ZHAI, W. & WANG, J. G. (2015). Energy metabolism disorder as a contributing factor of rheumatoid arthritis: A comparative proteomic and metabolomic study. *PLoS ONE* 10:1–15.

ZHANG, Y.; SZUSTAKOWSKI, J. & SCHINKE, M. (2009). Bioinformatics analysis of microarray data. *Methods in molecular biology (Clifton, N.J.)* 573:259–284.

8. ANEXOS

Anexo A: Figuras

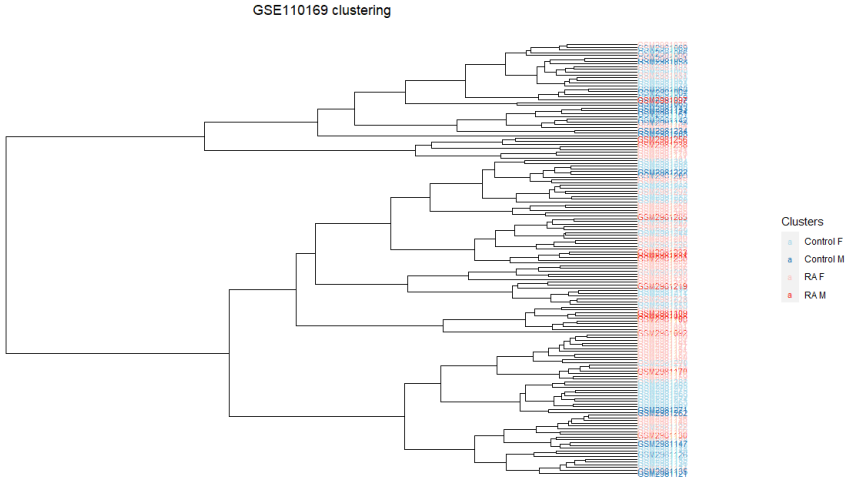

Figura A.1 | *Clustering* jerárquico del estudio GSE110169. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (RA F) y hombre con artritis reumatoide (RA M).

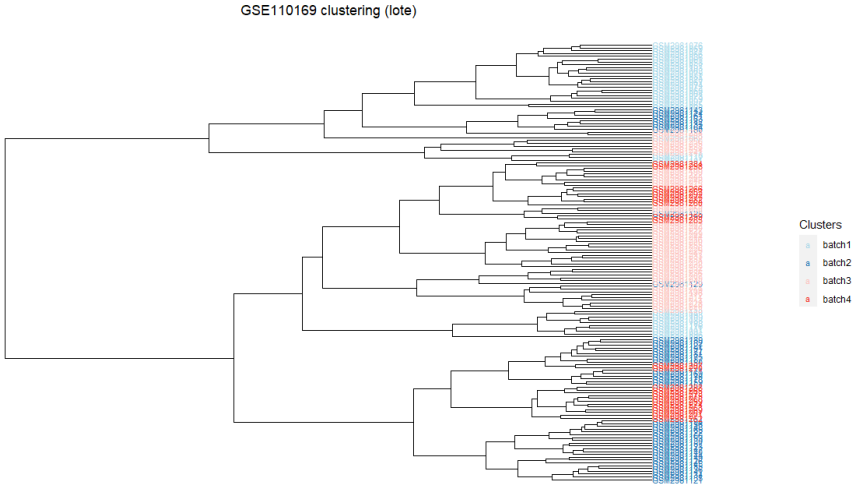

Figura A.2 | *Clustering* jerárquico del estudio GSE110169. Coloreado en función de los lotes (*batch*).

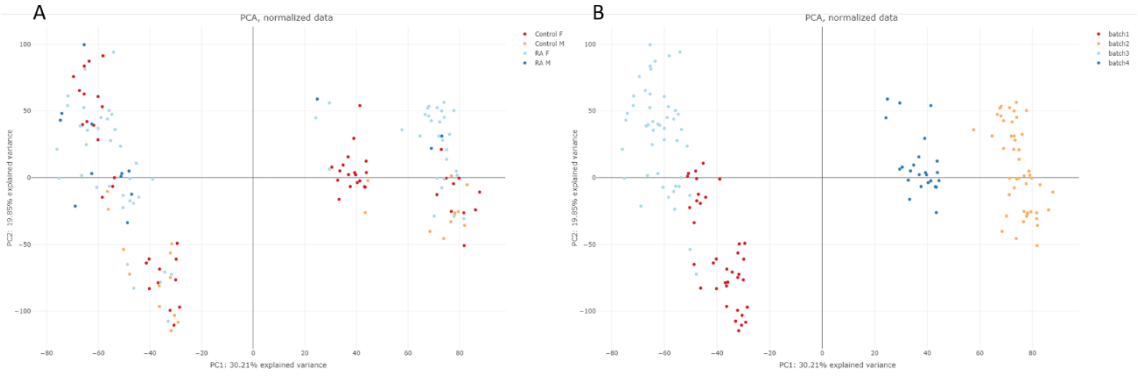

Figura A.3 | Análisis de componentes principales (PCAs) del estudio GSE110169. Se muestran las componentes principales (PC) 1 y 2. A: coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M). B: en función del lote.

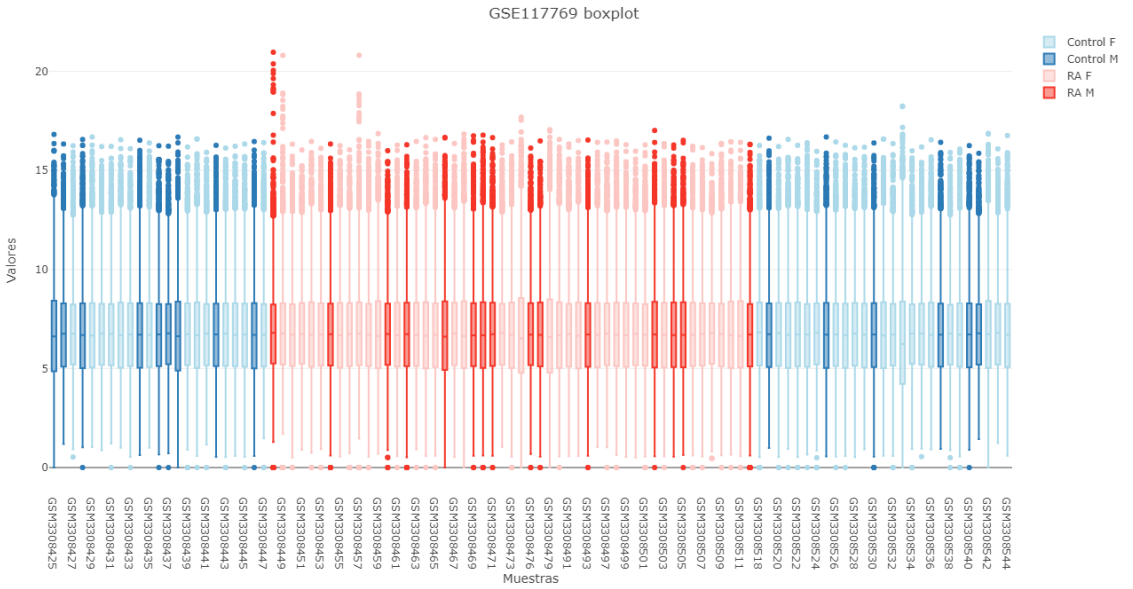

Figura A.4 | Diagrama de cajas del estudio GSE117769 previo a la eliminación de la muestra GSM3308533. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres AR (RA F) y hombres AR (RA M).

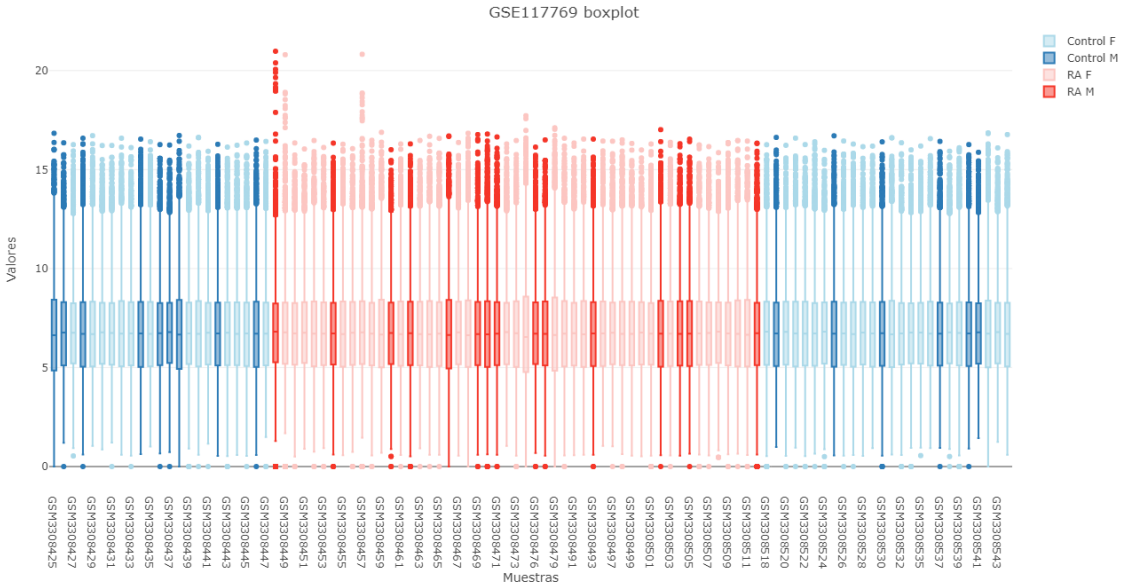

Figura A.5 | Diagrama de cajas del estudio GSE117769 tras la eliminación de la muestra GSM3308533. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

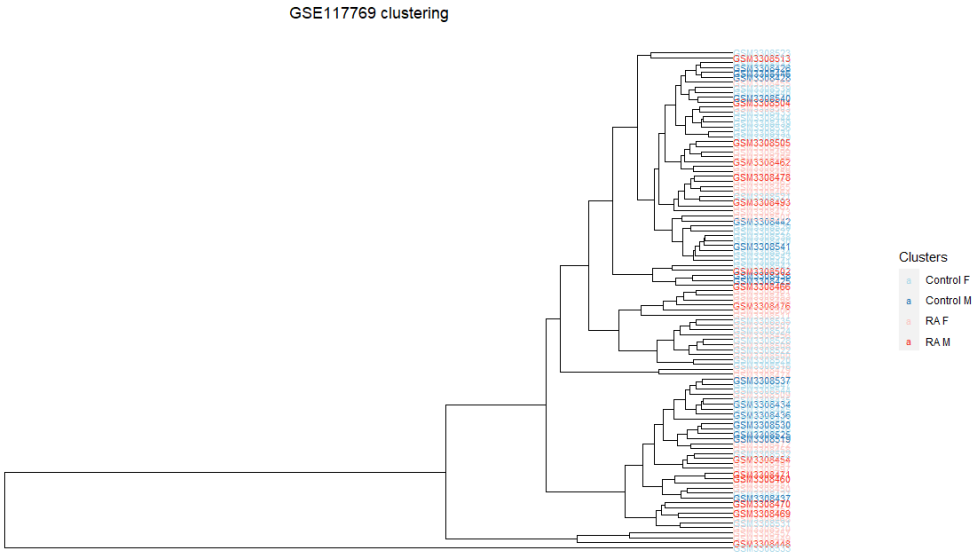

Figura A.6 | *Clustering* jerárquico del estudio GSE117169 antes de eliminar la muestra GSM8330533. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (RA F) y hombre con artritis reumatoide (RA M).

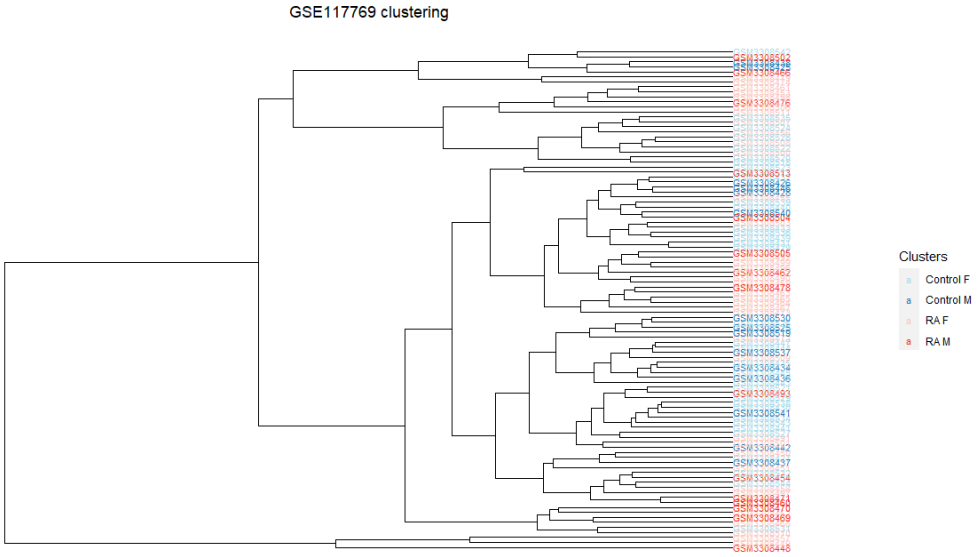

Figura A.7 | *Clustering* jerárquico del estudio GSE117769 tras eliminar la muestra GSM3308533. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (RA F) y hombre con artritis reumatoide (RA M).

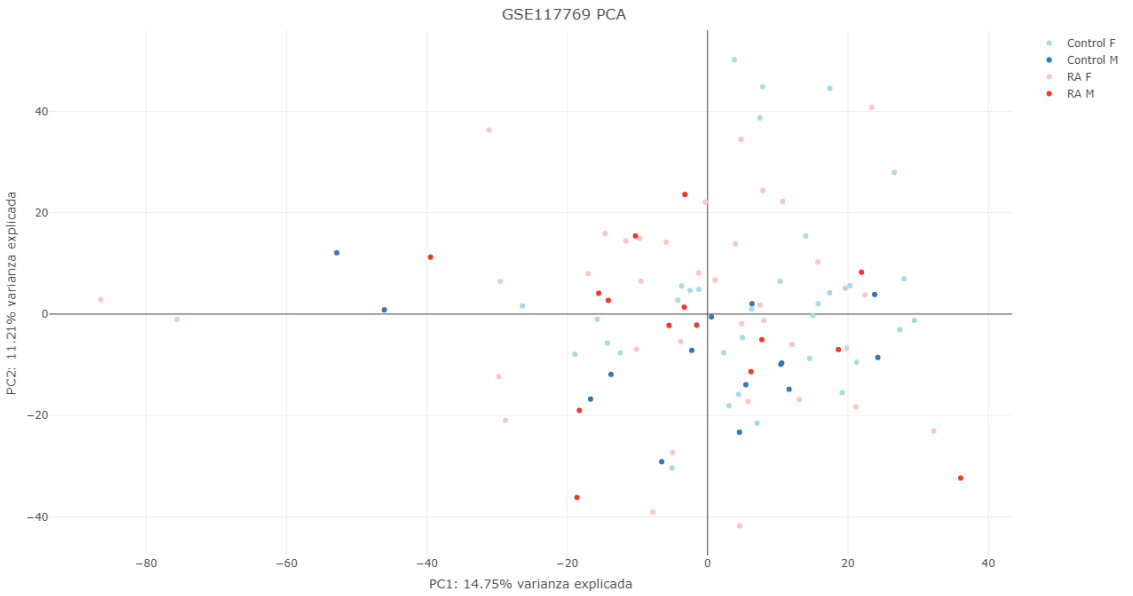

Figura A.8 | Análisis de componentes principales (PCA) del estudio GSE117169 tras eliminar la muestra GSM3308533. Se muestran las componentes principales (PC) 1 y 2. Las muestras se han coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

Figura A.9 | Diagrama de cajas del estudio GSE89408. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

Figura A.10 | Diagrama de cajas conjunto de los estudios GSE89408 y GSE97165. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

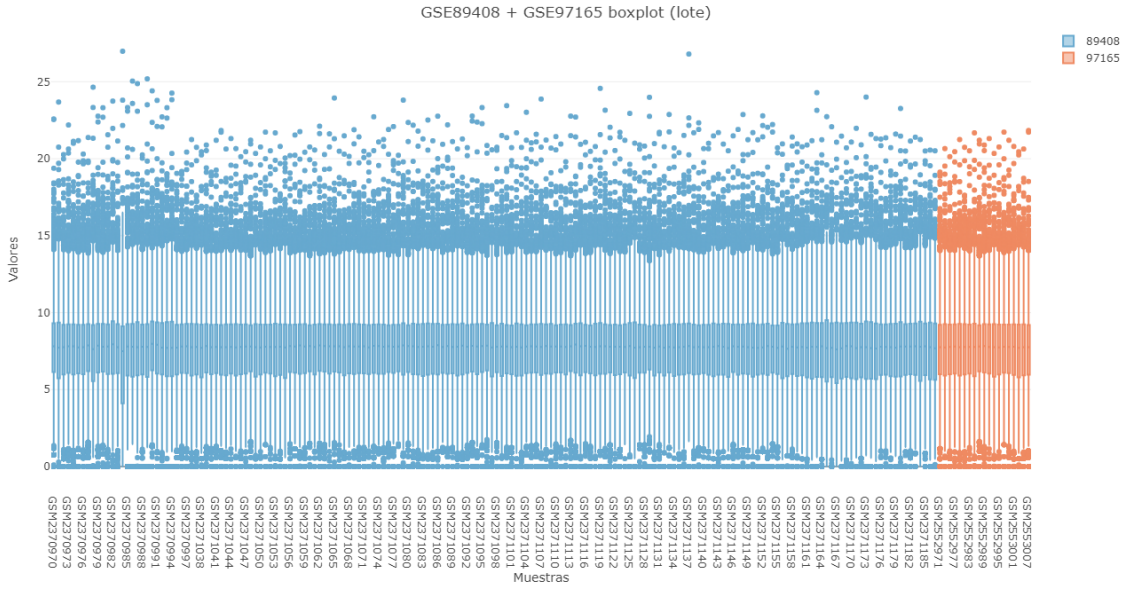

Figura A.11 | Diagrama de cajas conjunto de los estudios GSE89408 y GSE97165. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los *dataset*.

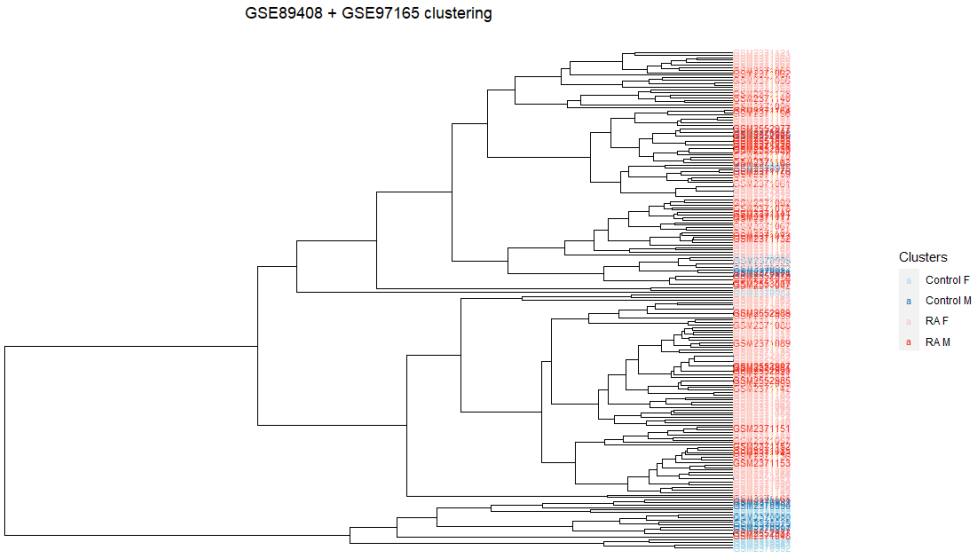

Figura A.12 | *Clustering* jerárquico de los *datasets* GSE89408 y GSE97165. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (RA F) y hombre con artritis reumatoide (RA M).

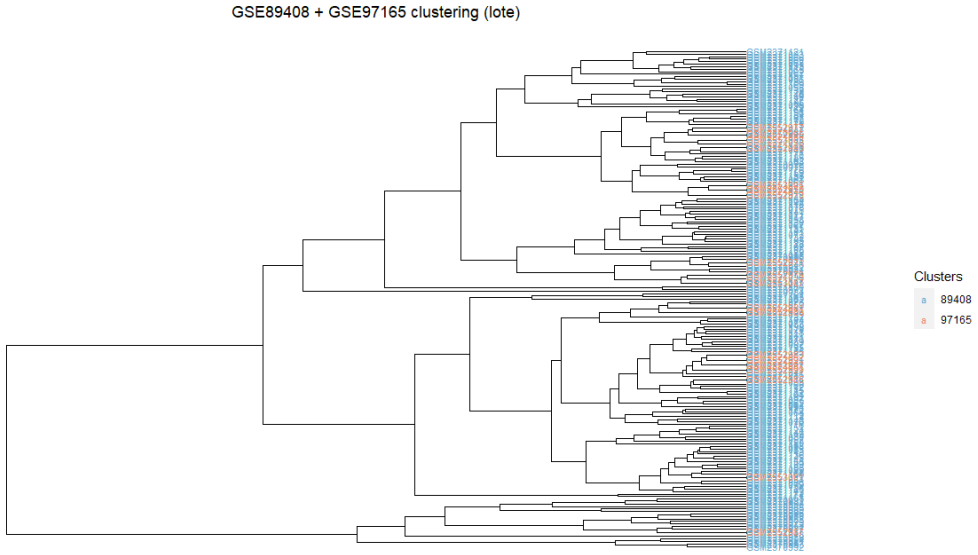

Figura A.13 | *Clustering* jerárquico de los *datasets* GSE89408 y GSE97165. Coloreado en función de los *datasets*.

8. ANEXOS

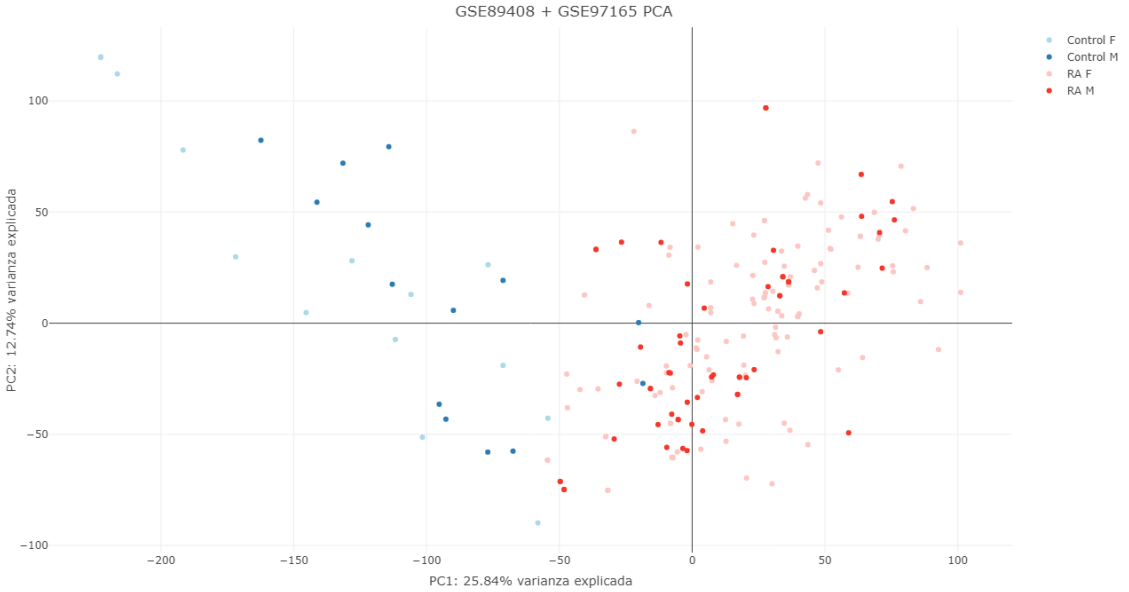

Figura A.14 | Análisis de componentes principales (PCA) de los *datasets* GSE89408 y GSE97165. Se muestran las componentes principales (PC) 1 y 2. Las muestras se han coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

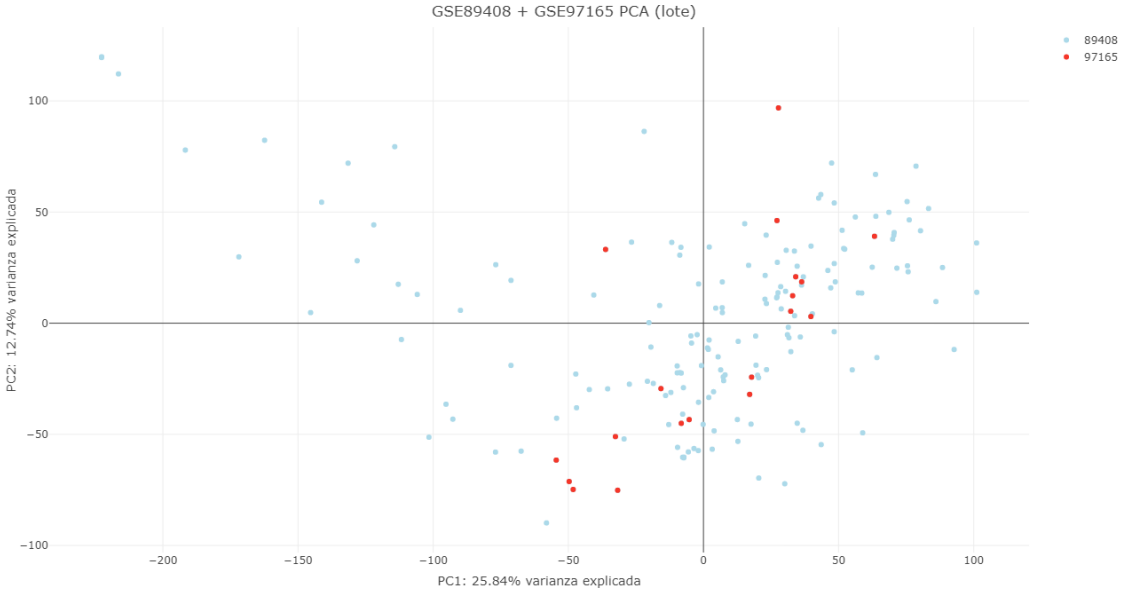

Figura A.15 | Análisis de componentes principales (PCA) de los *datasets* GSE89408 y GSE97165. Se muestran las componentes principales (PC) 1 y 2. Las muestras se han coloreado en función de los *datasets*.

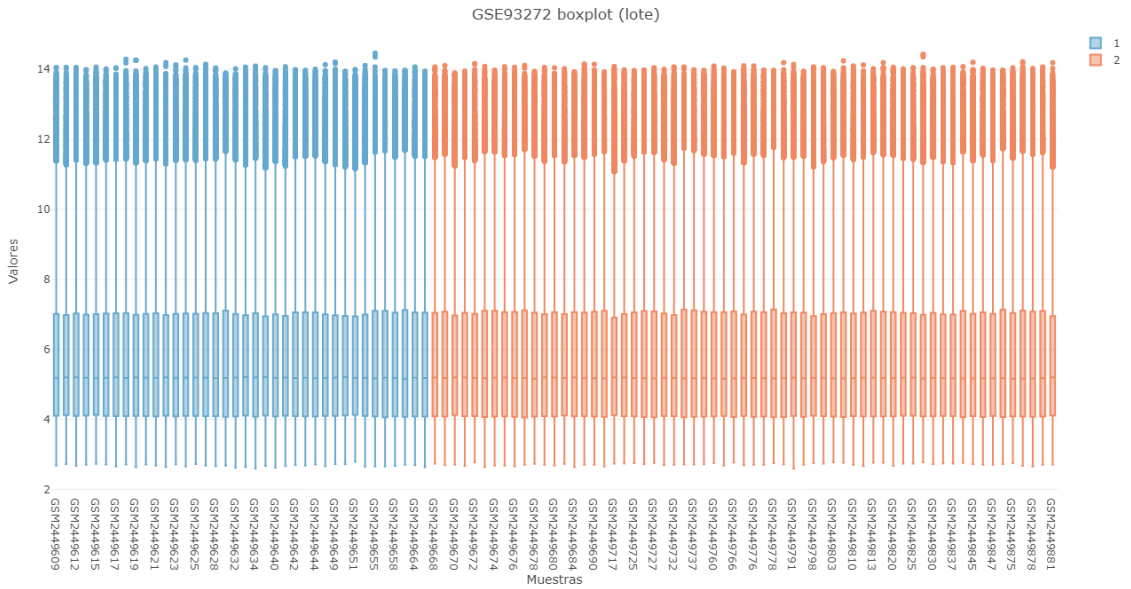

Figura A.16 | Diagrama de cajas del estudio GSE93272. Representa los valores de expresión (eje Y) de cada muestra (eje X). Coloreado en función de los lotes.

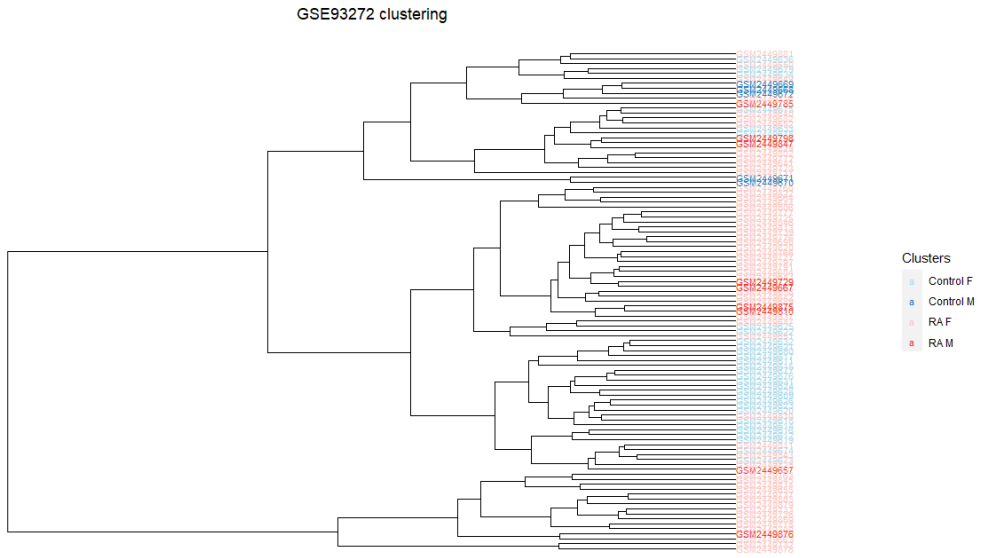

Figura A.17 | *Clustering* jerárquico del estudio GSE93272 en base a la distancia de correlación. Coloreado en función de los grupos experimentales: mujer control (Control F), hombre control (Control M), mujer con artritis reumatoide (RA F) y hombre con artritis reumatoide (RA M).

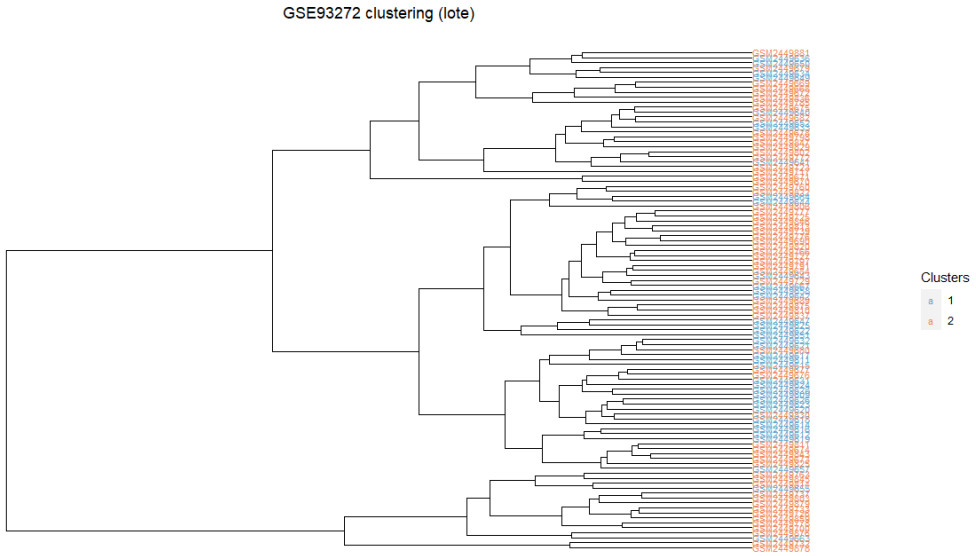

Figura A.18 | *Clustering* jerárquico del estudio GSE93272 en base a la distancia de correlación. Coloreado en función de los lotes.

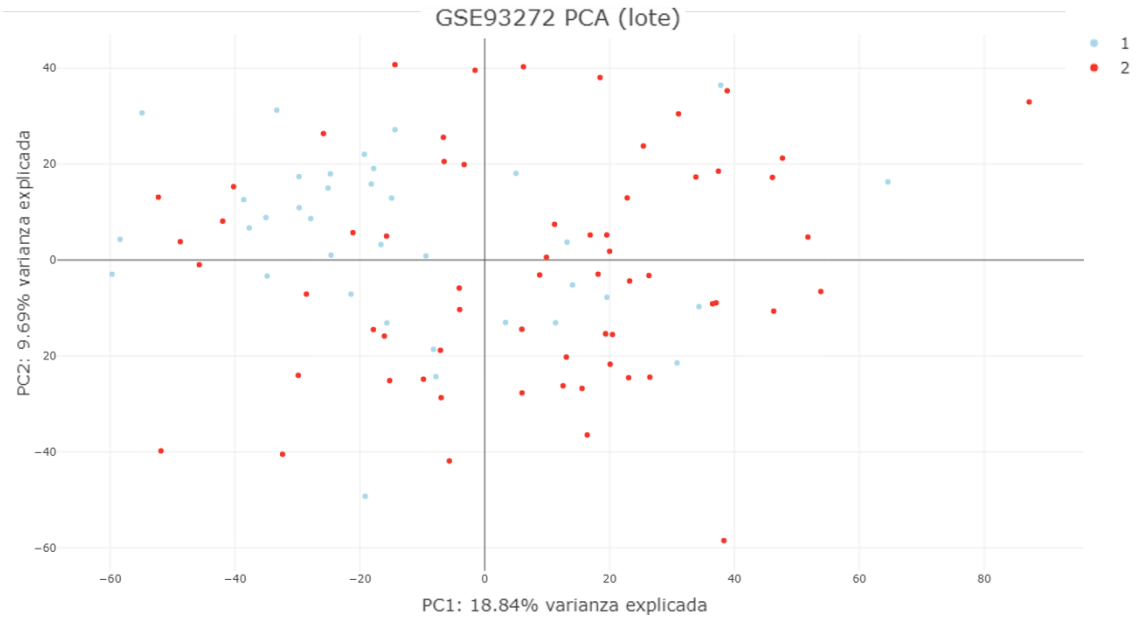

Figura A.19 | Análisis de componentes principales (PCA) de las muestras del GSE93272. Se muestran las componentes principales (PC) 1 y 2. Las muestras se han coloreado en función de los lotes (1 y 2).

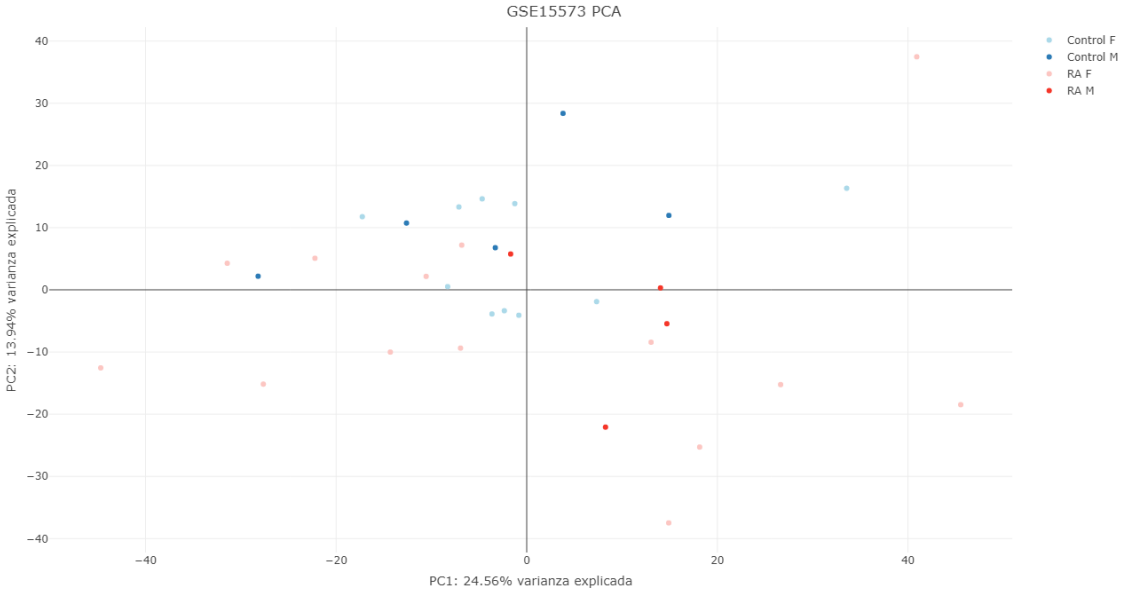

Figura A.20 | Análisis de componentes principales (PCA) del estudio GSE15573. Se muestran las componentes principales (PC) 1 y 2. Las muestras se han coloreado en función de los grupos experimentales: mujeres control (Control F), hombres control (Control M), mujeres con artritis reumatoide (RA F) y hombres con artritis reumatoide (RA M).

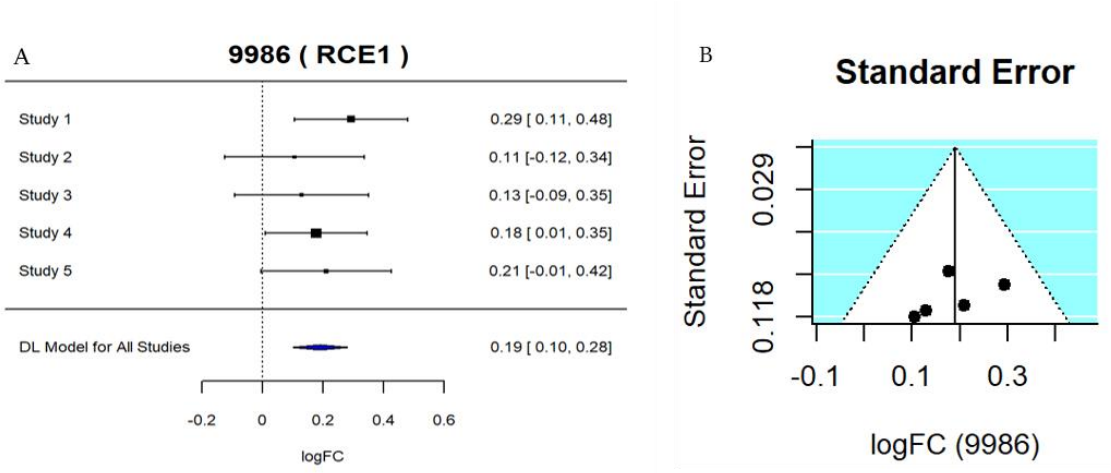

Figura A.21 | Gen RCE1: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

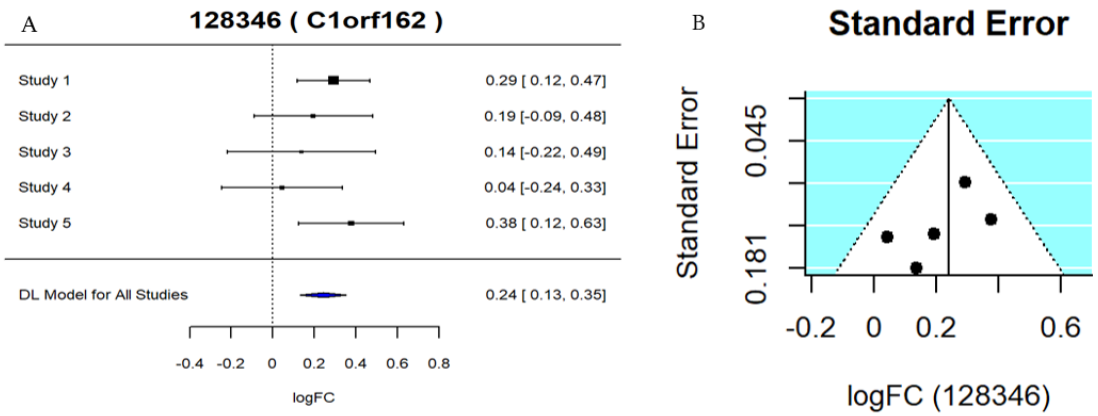

Figura A.22 | Gen C1orf162: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

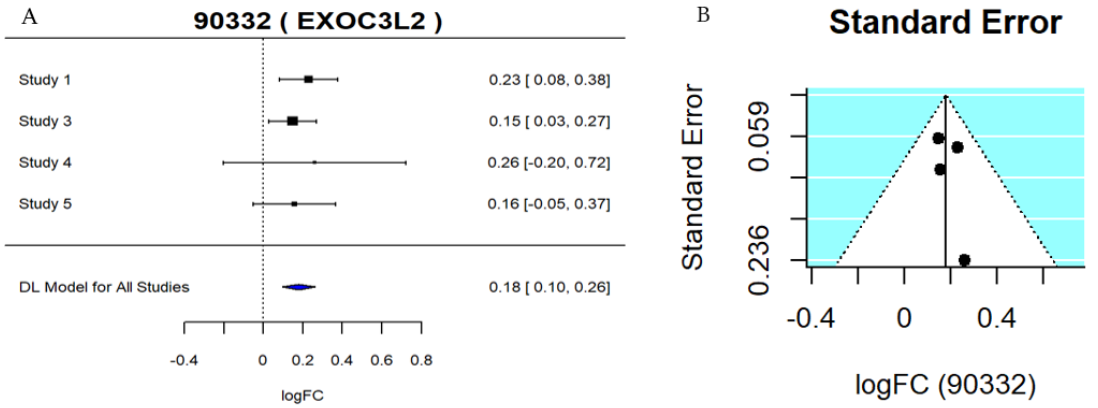

Figura A.23 | Gen EXOC3L2: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

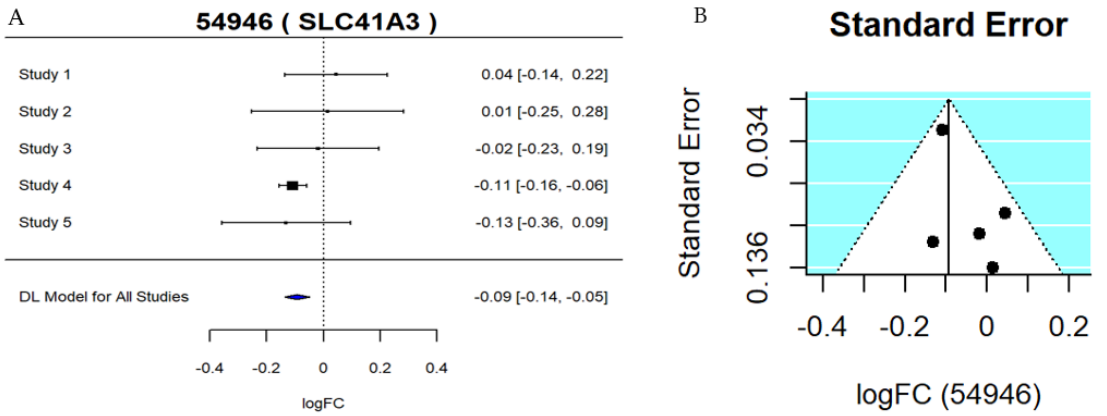

Figura A.24 | Gen SLC41A3: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

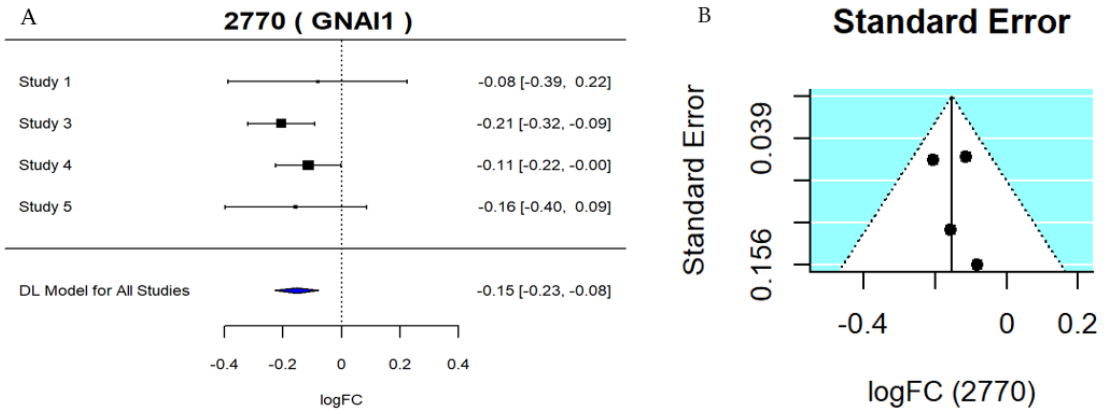

Figura A.25 | Gen GNAI1: A: Gráfico de bosque. Se muestra el tamaño del efecto de cada estudio (parte superior) y la estimación global (diamante azul), junto con su intervalo de confianza. La línea de puntos marca la línea de efecto nulo. **B: Gráfico de embudo.** Muestra la dispersión de las estimaciones de los estudios individuales representando el error estándar en una escala invertida frente al logaritmo de la magnitud de cambio. *Abreviaturas: DL, DerSimonian-Laird; logFC, logaritmo de la magnitud de cambio.*

Anexo B: Tablas

Tabla B.1 | Software empleado y sus versiones.

Software y paquetes	Versión
R	4.0.3
AnnotationDbi	1.52.0
Biobase	2.50.0
biomaRt	2.46.3
ComplexHeatmap	2.6.2
dplyr	1.0.6
edgeR	3.32.1
GEOquery	2.58.0
ggdendro	0.1.22
ggplot2	3.3.3
GO.db	3.12.1
hgu133plus2.db	3.2.3
hgu219.db	3.2.3
illuminaHumanv2.db	1.26.0
kableExtra	1.3.4
KEGG.db	3.2.4
knitr	1.33
limma	3.46.0
mdgsa	1.22.0
metafor	3.0-2
methods	4.0.3
org.Hs.eg.db	3.12.0
plotly	4.9.3
reshape2	1.4.4
rrvgo	1.2.0
scales	1.1.1
stats	4.0.3
tidyverse	1.3.1
UpSetR	1.4.0
utils	4.0.3

Tabla B.2 | Normalizaciones aplicadas a cada *dataset*.

Dataset	Normalización
GSE117769	Trimmed Mean of M-values (TMM)
GSE93272	Frozen robust multiarray análisis (fRMA)
GSE17755	Global ratio median
GSE15573	Quantile normalisation with Beadstudio
GSE110169	Robust Multi-Array Average (RMA)
GSE97165, GSE89408	Trimmed Mean of M-values (TMM)

Tabla B.3 | Selección de términos GO de procesos biológicos significativos en el enriquecimiento del metaanálisis. Se muestra el identificador GO (GO ID), nombre del término GO, los logaritmos de los odds ratio (LOR) y el P-valor ajustado por el método Benjamini-Yekutieli.

GO ID	Nombre	LOR	P-valor ajustado
GO:0043385	mycotoxin metabolic process	1.062	.015
GO:0046222	aflatoxin metabolic process	1.062	.015
GO:1901376	organic heteropentacyclic compound metabolic process	1.062	.015
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	0.504	.024
GO:0042775	mitochondrial ATP synthesis coupled electron transport	0.495	<.001
GO:0042773	ATP synthesis coupled electron transport	0.493	<.001
GO:0006081	cellular aldehyde metabolic process	0.482	.014
GO:0042267	natural killer cell mediated cytotoxicity	0.454	.020
GO:0002228	natural killer cell mediated immunity	0.433	.027
GO:0010257	NADH dehydrogenase complex assembly	0.432	.035
GO:0030514	negative regulation of BMP signaling pathway	-0.500	.011
GO:0002181	cytoplasmic translation	-0.523	<.001
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	-0.525	<.001
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	-0.526	<.001
GO:0045668	negative regulation of osteoblast differentiation	-0.546	.019
GO:0006376	mRNA splice site selection	-0.616	.041
GO:0070166	enamel mineralization	-0.813	.035
GO:0072202	cell differentiation involved in metanephros development	-0.813	.002
GO:0030728	ovulation	-0.849	.015
GO:0009086	methionine biosynthetic process	-1.052	.007