VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

CARACTERIZACIÓN DE LAS DIFERENCIAS DE SEXO EN ESCLEROSIS MÚLTIPLE MEDIANTE ESTRATEGIAS BASADAS EN EL ANÁLISIS MASIVO DE DATOS

AUTORA:
IRENE SOLER SÁEZ

TUTORES:
FRANCISCO GARCÍA GARCÍA
ZORAIDA ANDREU MARTÍNEZ
VICENTE ARNAU LLOMBART

JULIO, 2021

Presidente del Tribunal Evaluador / *President del tribunal avaluador* via **CCA del Máster** / *CCA del Màster*

VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

CARACTERIZACIÓN DE LAS DIFERENCIAS DE SEXO EN ESCLEROSIS MÚLTIPLE MEDIANTE ESTRATEGIAS BASADAS EN EL ANÁLISIS MASIVO DE DATOS

AUTORA:
IRENE SOLER SÁEZ

TUTORES:
FRANCISCO GARCÍA GARCÍA
ZORAIDA ANDREU MARTÍNEZ
VICENTE ARNAU LLOMBART

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

Presidente del Tribunal Evaluador / *President del tribunal avaluador* via **CCA del Máster** / *CCA del Màster*

RESUMEN

La esclerosis múltiple es una enfermedad crónica autoinmune que se caracteriza por la neurodegeneración del sistema nervioso central. Actualmente corresponde con la principal causa de discapacidad entre adultos jóvenes no asociada a accidentes. Su etiología es desconocida, y sus manifestaciones clínicas presentan gran variabilidad en función de las regiones afectadas. Tal y como se ha descrito para otras patologías, en la esclerosis múltiple existen diferencias entre hombres y mujeres (tanto a nivel clínico como epidemiológico).

Caracterizar a nivel molecular la enfermedad con perspectiva de sexo mejoraría su abordaje sanitario mediante una medicina más personalizada y de precisión. Con este objetivo, en el presente trabajo se realiza un análisis bioinformático con datos transcriptómicos de células y núcleos únicos para la búsqueda de potenciales biomarcadores. En detalle, se ha desarrollado un abordaje *in silico* con datos de tres estudios procedentes de repositorios públicos que, en conjunto, permiten evaluar tipos celulares del sistema nervioso y del sistema inmunitario.

Como resultado, se han encontrado diferencias en la proporción de tipos celulares presentes en las muestras en función del sexo de los individuos. Asimismo, numerosos genes presentan un patrón de expresión diferencialmente significativo por tipo celular. Estos genes se han caracterizado funcionalmente desde una perspectiva integrativa. En detalle, se han identificado sobrerrepresentadas funciones biológicas relacionadas con la adhesión celular en el sistema nervioso de mujeres. En hombres, se encuentran sobreexpresados genes comunes a todos los tipos celulares que previamente se han asociado con la enfermedad. Los resultados para el sistema inmunitario también son de interés. Mientras que las mujeres tienen enriquecidas funciones relacionadas con eventos postranscripcionales (en linfocitos T CD4+) y respuestas ante diferentes fuentes de estrés (en linfocitos T CD8+), los hombres presentan sobrerrepresentadas de forma común multitud de funciones inmunológicas, tales como el procesamiento y la presentación de antígenos.

Palabras clave: esclerosis múltiple, sexo, transcriptómica, célula única.

ABSTRACT

Multiple sclerosis is a chronic autoimmune and neurodegenerative disease of the central nervous system. Currently, it is the main cause of non-traumatic disability among young adults. Its aetiology is not well understood, and its clinical manifestations are dependent on the affected regions. As it has been reported for a wide variety of diseases, there is a sex bias in several epidemiological and clinical aspects of multiple sclerosis.

Our aim is to describe the molecular mechanisms of the disease from a sex perspective to find prospective biomarkers that could boost personalized and precision medicine. With that objective in mind, we have implemented a bioinformatic analysis of transcriptomic data coming from single cells and single nuclei. Specifically, we have developed an *in silico* approach for processing data previously downloaded from three public available studies. Altogether, they enable the evaluation of different cell types from the nervous system and the immune system.

As a result, we found sex differences in the proportions of some cell types. Moreover, there were detected significant differential expression patterns in several genes for each cell type evaluated. The results from the latter were integrated among cell types and studies to perform robust functional profiling analysis. With this procedure we identified overrepresentated biological functions in women's nervous system related with cell adhesion. In men, all cell types had commonly upregulated genes previously associated with the disease. The results for the immune system were interesting too. Whilst women had an overrepresentation of postranscriptional functions (in T CD4+ cells) and cellular responses to different stresses (in T CD8+ cells), men had upregulated several immune processes like antigen processing and presentation.

Key words: multiple sclerosis, sex, transcriptomics, single cell.

ÍNDICE

1. ABREVIATURAS

ADN: ácido desoxirribonucleico

ARI: del inglés *Adjusted Rand Index* (índice ajustado de Rand)

ARN: ácido ribonucleico

ARNm: ácido ribonucleico mensajero

BH: Benjamini-Hochberg

BHE: barrera hematoencefálica

CIPF: Centro de Investigación Príncipe Felipe

CPU: del inglés *Central Processing Unit* (unidad central de procesamiento)

DOI: del inglés *Digital Object Identifier* (identificador de objeto digital)

EM: esclerosis múltiple

EMPP: esclerosis múltiple primaria progresiva

EMRR: esclerosis múltiple remitente remanente

EMSP: esclerosis múltiple secundaria progresiva

FAIR: del inglés *Findable, Accesible, Interoperable, Reusable* (encontrable, accesible, interoperable y reutilizable)

FDR: del inglés *False Discovery Rate* (tasa de falsos rechazados)

GEO: del inglés *Gene Expression Omnibus* (recopilación de expresiones génicas)

GO: del inglés *Gene Ontology* (ontología génica)

GWAS: del inglés *Genome-Wide Association Study* (estudios de asociación de genoma completo)

HGNC: *HUGO Gene Nomenclature Committee* (cómite de nomenclatura de genes de HUGO)

HLA: del inglés *Human Leukocyte Antigen* (antígeno leucocitario humano)

HUGO: del inglés *Human Genome Organisation* (organización del genoma humano)

IRM: imágenes por resonancia magnética

lncRNA: del inglés *long non-coding Ribonucleic acid* (ácido ribonucleico largo no codificante)

logFC: del inglés *Fold Change logarithm* (logaritmo de la tasa de cambio)

MAD: del inglés *Median Absolute Deviation* (desviación media absoluta)

PBMC: del inglés *Peripheral Blood Mononuclear Cell* (células mononucleares de sangre periférica)

PCA: del inglés *Principal Component Analysis* (análisis de componentes principales)

PRISMA: del inglés *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (ítems preferentes a informar en las revisiones sistemáticas y los metaanálisis)

RAM: del inglés *Random Access Memory* (memoria de acceso aleatorio)

REVIGO: del inglés *REduce and VIsualize Gene Ontology* (reducir y visualizar ontologías génicas)

microRNA: del inglés *micro Ribonucleic acid* (ácido ribonucleico micro)

RNA-seq: del inglés *Ribonucleic acid Sequencing* (secuenciación de ácido ribonucleico)

scRNA-seq: del inglés *single cell Ribonucleic acid Sequencing* (secuenciación de ácido ribonucleico de célula única)

SLURM: del inglés *Simple Linux Utility for Resource Management* (utilidad básica de Linux para la gestión de recursos)

SNC: sistema nervioso central

SNN: del inglés *Shared Nearest Neighbours* (vecinos comunes más cercanos)

snRNA-seq: del inglés *single nucleus Ribonucleic acid Sequencing* (secuenciación de ácido ribonucleico de núcleo único)

tSNE: del inglés *T-distributed Stochastic Neighbor Embedding* (incrustación de vecinos estocásticos distribuídos en t)

UCSC: del inglés *University of California Santa Cruz* (universidad de California en Santa Cruz)

UMAP: del inglés *Uniform Manifold Approximation and Projection* (distribución de aproximación y proyección uniforme)

2. INTRODUCCIÓN

2.1. Esclerosis múltiple

Una de las principales funciones del sistema inmunitario es la capacidad de discernir entre lo propio y lo ajeno. Concretamente, las células nucleadas de un individuo son capaces de mostrar en su superficie fragmentos de las moléculas que contienen en su interior (antígenos). Mediante esta exposición, los componentes del sistema inmunitario reconocen antígenos que no son propios de la célula y activan mecanismos de respuesta para eliminar al agente nocivo (por ejemplo, al reconocer un péptido de un microorganismo en una célula infectada). Este sistema de presentación de antígenos conlleva un riesgo potencial de estimulación del sistema inmunitario ante la presencia de fragmentos de moléculas del propio individuo (autoantígenos). Por ello, durante el proceso de maduración de las células inmunitarias se establece cuáles son capaces de reconocer con alta afinidad autoantígenos (células autorreactivas) para su destrucción vía apoptosis[1].

Existe un amplio conjunto de patologías heterogéneas caracterizadas por un incremento de las células autorreactivas del sistema inmunitario, las cuales son englobadas bajo el término de **enfermedades autoinmunes.** Como consecuencia de la activación de estas células se desencadenan respuestas inmunitarias inflamatorias, mayoritariamente crónicas, que deterioran los tejidos del individuo. La patología desarrollada dependerá de los autoantígenos reconocidos, que a su vez determinarán los órganos y/o sistemas afectados[2,3].

Entre las enfermedades autoinmunes más comunes destaca la **esclerosis múltiple** (EM), siendo actualmente la principal causa de discapacidad no asociada a accidentes entre la población adulta joven[4]. Se caracteriza por desarrollar respuestas inmunitarias anómalas que destruyen la sustancia grasa que recubre y protege las fibras nerviosas del cerebro y la médula espinal (**sistema nervioso central,** SNC)[5,6]. El consecuente deterioro de este sistema implica, entre otros aspectos, la alteración de los circuitos neuronales y una transmisión del impulso nervioso deficiente[7]. En último término, conlleva a la pérdida neuronal, por lo que la EM es considerada una **enfermedad neurodegenerativa**[5-7].

Durante el transcurso de la enfermedad, las manifestaciones clínicas están supeditadas a las regiones neuronales afectadas; por lo que difieren entre personas y durante su evolución temporal en un mismo individuo. Entre ellas, destacan la neuritis óptica aguda, las alteraciones sensoriales y motoras, los déficits cognitivos, la debilidad y diversas encefalopatías[8]. Como resultado de su extensa variabilidad fenotípica, la EM es conocida coloquialmente como la "enfermedad de las mil caras"[4].

2.1.1. Epidemiología

Atendiendo a los datos reportados en 2020, se calcula que la EM afecta a 2,8 millones de personas a nivel mundial, de las cuales aproximadamente 500.000 son europeas y, a su vez, 50.000 españolas. Su distribución global presenta gran variabilidad entre países, con valores estimados de prevalencia que oscilan entre 1 de cada 300 habitantes y 1 de cada 3.000 (*Figura 1*). Estas diferencias pretenden ser explicadas a través de factores de diversa índole, entre los cuales destacan la concienciación social, la capacidad de acceso al sistema sanitario y aspectos tanto genéticos como ambientales[4-6,9].

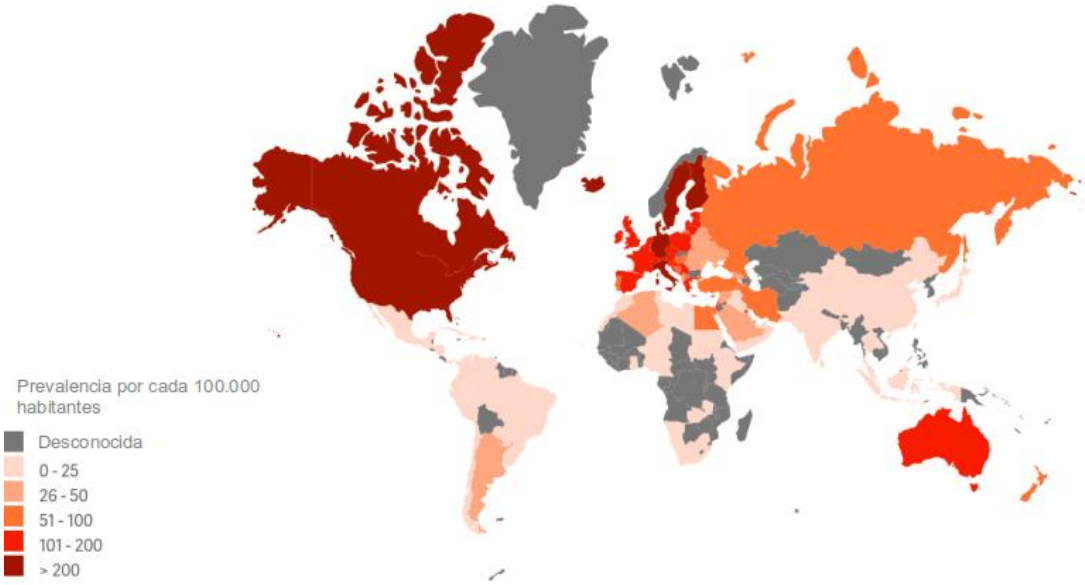


Figura 1. Prevalencia mundial de la EM por países. *Mapa del mundo coloreado por rangos atendiendo al número estimado de personas que sufren EM en cada país por cada 100.000 habitantes. Representación gráfica basada en la integración de datos epidemiológicos recopilados por la federación internacional de EM en 2020. Figura modificada de https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms con fecha 14-05-2021. EM: esclerosis múltiple.*

Se estima que en promedio son diagnosticadas 300 personas a diario; valor que podría estar subestimado debido a la ausencia de datos de múltiples países[9]. En España, la incidencia anual se establece en 4,2 diagnósticos por cada 100.000 habitantes, cifra que corresponde, de forma aproximada, con 2.000 casos nuevos al año[10].

Independientemente de la zona geográfica, el inicio de los síntomas suele producirse entre los 20 y 40 años. Sin embargo, en la última década se ha observado un incremento en la incidencia y prevalencia de EM infantil. Pese a que personas de ambos sexos pueden sufrir la enfermedad, la susceptibilidad es mayor en mujeres respecto a hombres; presentando ratios que oscilan entre 2:1 y 3:1. Esta razón se está ampliando cada año como consecuencia de un aumento en el número de mujeres afectadas[4-6,9,10].

2.1.2. Factores de riesgo

La etiología de la EM es desconocida, pero parece ser debida a una combinación de factores ambientales, que en personas genéticamente predispuestas, desencadenan una respuesta inmunológica anómala.

Los principales factores de riesgo descritos a nivel genético corresponden con diversos alelos del complejo HLA (por sus siglas en inglés *Human Leukocyte Antigen*), tales como *HLA-DRB1*15:01* y *HLA-A*02*[11,12]. Este conjunto de genes está implicado en la presentación de antígenos, por lo que las proteínas codificadas realizan una función esencial en la discriminación de los elementos propios y ajenos del individuo[1]. Asimismo, estudios de asociación de genoma completo (GWAS, por sus siglas en inglés *Genome-Wide Association Study*) han permitido identificar cientos de variantes génicas adicionales al complejo HLA, incluyendo alelos de los genes *IL2RA, IL7RA y STAT3*[6,12]. Adicionalmente, se han descrito efectos de los cromosomas sexuales en los sucesos de autoinmunidad y neurodegeneración en la EM[13].

En términos ambientales también se han identificado numerosos factores de riesgo. Entre ellos, destacan por su evidencia consolidada: presentar deficiencia en vitamina D, vivir en regiones de elevada latitud, ser fumador, ser mujer, y/o haber sufrido una infección del virus Epstein-Barr. Adicionalmente, la obesidad durante la adolescencia, trabajar en horario nocturno y la contaminación ambiental están cobrando mayor relevancia en los últimos años[11,12,14].

2.1.3. Clasificación, diagnóstico y tratamientos

La EM se divide en diferentes subtipos en base a la evolución temporal de la aparición de signos y síntomas, así como de su posible recuperación (*Figura 2*)[15]. La forma más común, denominada **esclerosis múltiple remitente remanente** (EMRR), se caracteriza por presentar manifestaciones clínicas de forma esporádica. De este modo, los pacientes sufren periodos con síntomas (denominados brotes) de duración fluctuante entre días y meses, alternados con fases de recuperación casi completa. Este subtipo de EM puede mantenerse a lo largo de la vida del individuo o agravarse al persistir los síntomas de forma continua y gradual, pasando a denominarse **esclerosis múltiple secundaria progresiva** (EMSP). Por último, la enfermedad puede evolucionar desde el inicio de los síntomas de forma persistente con ausencia de recuperación. Este subtipo minoritario se designa como **esclerosis múltiple primaria progresiva** (EMPP)[6,15].

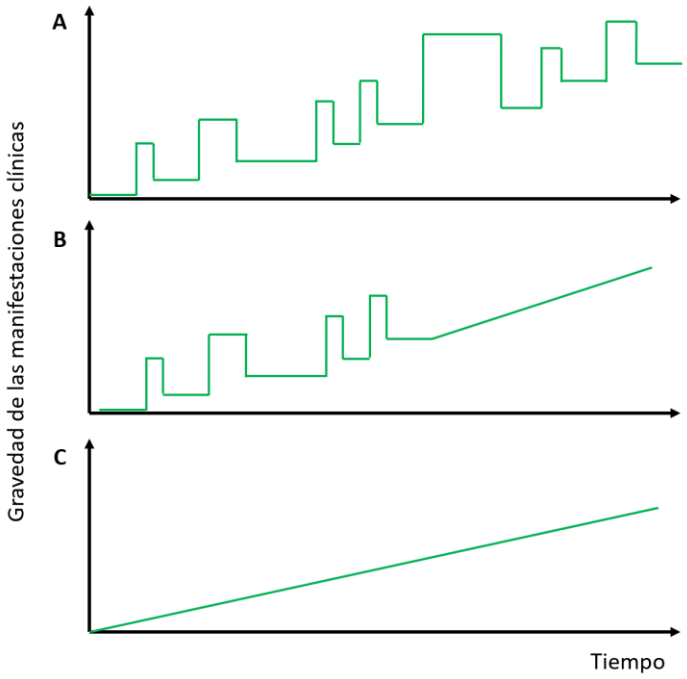


Figura 2. Evolución temporal de las manifestaciones clínicas en los diferentes subtipos de EM. *Representación gráfica que ejemplifica con unidades arbitrarias la gravedad de los signos y síntomas frente al tiempo para las condiciones esclerosis múltiple remitente remanente (A), esclerosis múltiple secundaria progresiva (B), y esclerosis múltiple primaria progresiva (C). EM: esclerosis múltiple.*

El **diagnóstico** de la EM presenta gran complejidad debido al amplio espectro de variabilidad fenotípica que presentan los pacientes. Por ello, la labor del personal médico está enfocada a la realización de pruebas que permitan, de manera conjunta, descartar patologías con manifestaciones clínicas similares e identificar signos y síntomas característicos de la EM. Entre los métodos de diagnóstico comúnmente utilizados destacan[16]:

- Las examinaciones neurológicas para evaluar el funcionamiento del SNC.
- Los análisis de sangre que permitan descartar la presencia de infecciones y/o la carencia de determinados compuestos bioquímicos.
- Los análisis de líquido cerebroespinal para la detección de anticuerpos capaces de reconocer autoantígenos asociados a la EM.
- La neuroimagen a través de imágenes por resonancia magnética (IRM) para la identificación de lesiones en el SNC que permitan discernir entre EM y otro tipo de enfermedades neurológicas.

Tras obtener el historial clínico y los resultados de las pruebas realizadas, en caso de no identificar otra patología se aplican los criterios revisados de McDonald en 2017 para

confirmar el diagnóstico y establecer el subtipo de la enfermedad[17]. En el caso de la forma EMRR existen múltiples combinaciones de resultados que permiten establecer el diagnóstico (*Tabla 1*). Por el contrario, para las condiciones progresivas es indispensable que el individuo presente la sintomatología sin remisión durante un periodo de tiempo superior a un año. Adicionalmente, debe cumplir al menos 2 de los siguientes criterios: (1) presentar como mínimo una lesión en regiones determinadas del encéfalo, (2) presentar dos o más lesiones en la médula espinal y/o (3) detectar anticuerpos que reconocen autoantígenos en el líquido cerebroespinal[5,6,17].

Número de brotes	Número de lesiones en SNC	Evidencias adicionales requeridas
mínimo 2	mínimo 2	Ninguna
mínimo 2	1	Historial clínico con brote previo que haya afectado a una región diferente del SNC respecto a la lesión reportada
mínimo 2	1	Lesión multifocal *
1	mínimo 2	Presencia de anticuerpos que reconocen autoantígenos en el líquido cerebroespinal
1	mínimo 2	Brote adicional que afecte a una región diferente del SNC respecto a la lesión reportada
1	1	Lesión multifocal * y presencia de anticuerpos que reconocen autoantígenos en el líquido cerebroespinal
1	1	Lesión multifocal * y brote adicional que afecte a una región diferente del SNC respecto a la lesión reportada

Tabla 1. Criterios revisados de McDonald en 2017 para el diagnóstico de la esclerosis múltiple remitente remanente. *Representación por filas de las posibles combinaciones en el número de brotes (columna 1), número de lesiones que afecten al sistema nervioso central (columna 2) y evidencias adicionales (columna 3) que determinen el diagnóstico de EM. * Lesión multifocal: conjunto de lesiones con origen común que afectan a diferentes áreas del encéfalo. SNC: sistema nervioso central, EM: esclerosis múltiple. Tabla adaptada de J. A. Cohen et al[17].*

En la actualidad **no existe ningún tratamiento que permita curar la EM**. Por ello, las terapias desarrolladas están enfocadas a paliar la sintomatología, así como a ralentizar el agravamiento de la enfermedad.

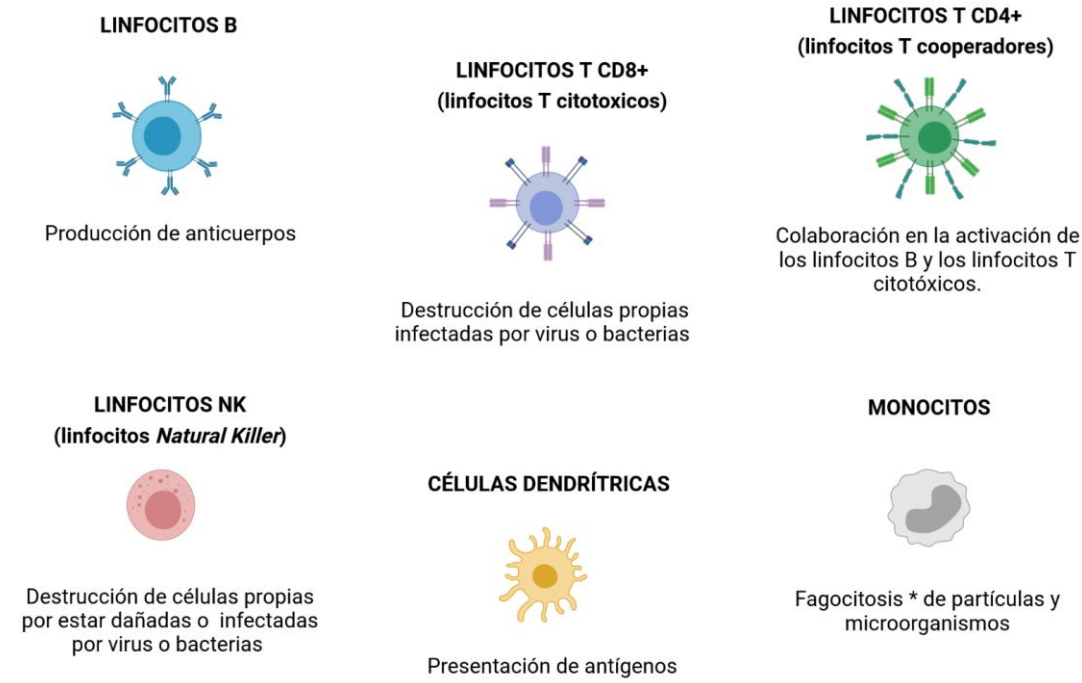
Entre los tratamientos farmacológicos más comunes se encuentran los corticoides (gracias a sus propiedades antiinflamatorias) y los compuestos inmunomoduladores (sustancias que permiten modificar el comportamiento del sistema inmunitario para reducir la aparición de brotes y su severidad)[18]. Los nuevos avances en esta disciplina están dirigidos a silenciar la respuesta inmunitaria de tipos celulares concretos, con el objetivo de alcanzar tratamientos más específicos que reduzcan los efectos adversos[19].

También son relevantes los tratamientos rehabilitadores, efectuados en ámbitos como la fisioterapia y la psicología, que contribuyen a disminuir el impacto de las manifestaciones clínicas mejorando la calidad de vida de los pacientes[20].

2.1.4. Bases celulares y moleculares de la enfermedad

La EM es una patología que afecta primordialmente a los **sistemas nervioso e inmunitario.** Cada uno de ellos está conformado a su vez por múltiples tipos celulares, los cuales desempeñan actividades específicas e interconectadas para realizar las funciones atribuidas al correspondiente sistema (*Figura 3*)[1,21,22]. Por ello, la comprensión detallada de la EM incluye la descripción de las alteraciones que se producen en los diferentes tipos celulares.

A



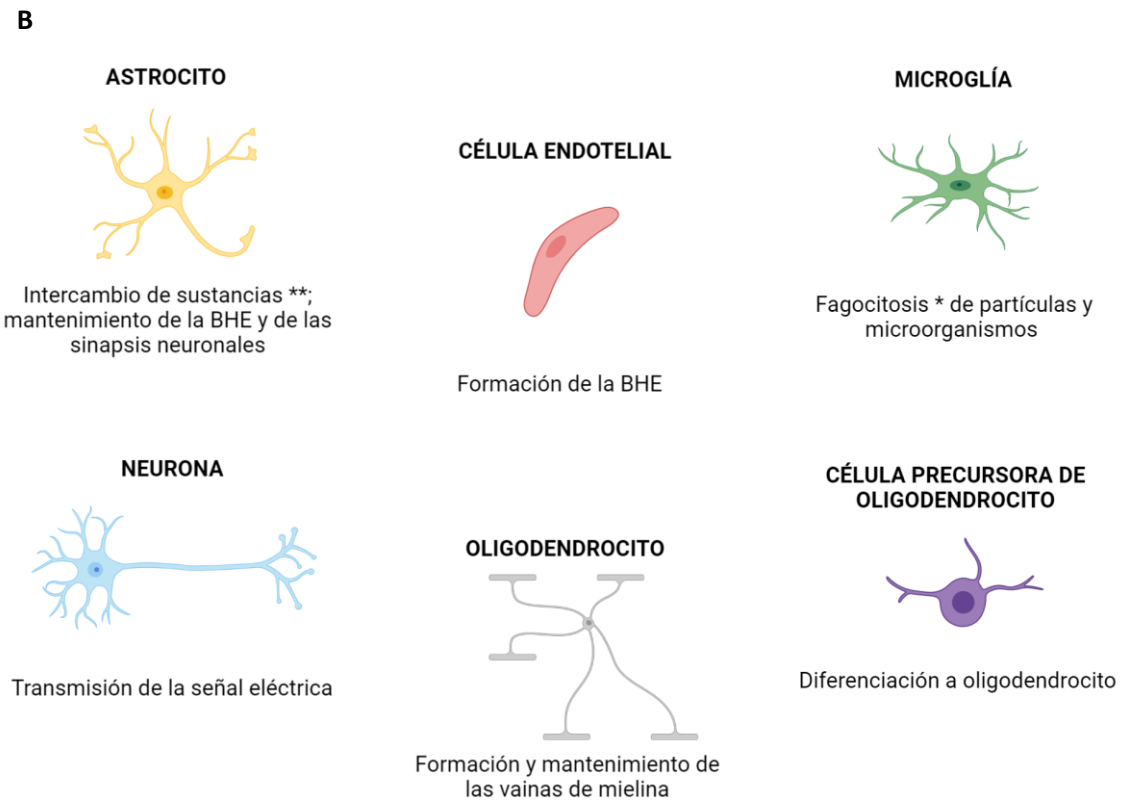


Figura 3. Funciones principales de diversos tipos celulares del sistema inmunitario (A) y del sistema nervioso (B). *(A) Células mononucleares del sistema inmunitario (linfocitos B, linfocitos T CD4+, linfocitos T CD8+, linfocitos NK, células dendríticas y monocitos). (B) Células destacadas del tejido nervioso (astrocitos, células endoteliales, microglía, neuronas, oligodendrocitos y células precursoras de oligodendrocitos). BHE: barrera hematoencefálica. * Fagocitosis: endocitosis de elementos vía prolongaciones de la membrana plasmática; ** El término "sustancias" refiere a metabolitos y restos de desecho. Figuras diseñadas en BioRender.com*

Concretamente, los axones de las neuronas del SNC en estado fisiológico se encuentran recubiertos por la membrana plasmática de los oligodendrocitos. Esta envoltura lipoproteica en forma de vaina, denominada **mielina**, actúa como aislante eléctrico durante la transmisión del impulso nervioso y participa en la correcta formación de los circuitos neuronales[21]. En la enfermedad de EM la mielina es destruida por un proceso cuya etiología continúa siendo desconocida, como se ha comentado anteriormente (*Figura 4*).

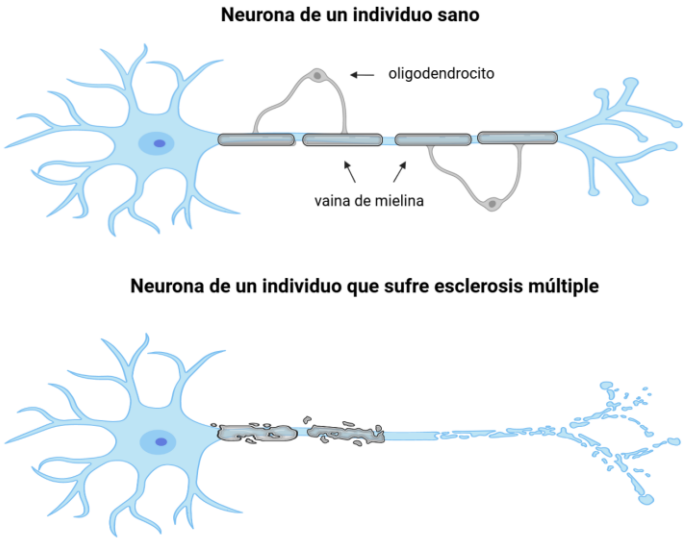


Figura 4. Neuronas en estado fisiológico y patológico por la enfermedad de EM. *Representación gráfica de neuronas (célula azul) recubiertas por la membrana plasmática de los oligodendrocitos (célula gris) en estado de salud y alteradas en la patología de EM. EM: esclerosis múltiple. Figura diseñada en BioRender.com.*

La participación de células inmunitarias autorreactivas contra la mielina procedentes de la sangre periférica (principalmente linfocitos T CD4+ y linfocitos T CD8+) está ampliamente documentada. Para desvelar el desencadenante de este proceso se han propuesto múltiples autoantígenos, así como distintas alteraciones en la presentación de los mismos. Sin embargo, ninguno de ellos ha sido confirmado como desencadenante de la enfermedad[22-25].

Las células inmunitarias procedentes de la sangre periférica acceden al tejido nervioso atravesando zonas alteradas de la barrera hematoencefálica (BHE), estructura formada por células endoteliales que permite cribar físicamente los elementos que acceden al SNC. En concreto, las lesiones observadas por IRM suelen localizarse adyacentes a esta barrera; principalmente cerca de roturas y/o regiones de mayor permeabilidad[8].

Adicionalmente, los linfocitos, los monocitos y la microglía se caracterizan por encontrarse en un estado proinflamatorio, secretando citoquinas, quimiocinas y radicales libres. Como resultado, se genera un microambiente que favorece la desmielinización y la degeneración del tejido nervioso. Por ello, numerosas terapias farmacológicas tratan de ejercer actividades neuroprotectoras reduciendo la inflamación[22,26,27].

Asimismo, está ampliamente descrito el fenómeno de gliosis reactiva en pacientes de EM. En condiciones inflamatorias, los astrocitos se activan modificando significativamente su morfología y patrón de expresión molecular; generando una estructura conocida como cicatriz glial. El objetivo principal de este proceso consiste en ejercer funciones protectoras (por ejemplo, fagocitando restos de mielina y reparando la BHE). Sin embargo, en esta conformación también secretan citocinas y quimiocinas que promueven la inflamación[28].

2.2. Tecnologías de alto rendimiento: técnicas ómicas

Las tecnologías de alto rendimiento, también conocidas como técnicas ómicas, constituyen un conjunto de metodologías destinadas a la obtención masiva de datos del tipo de molécula de interés. Estas herramientas conforman un cambio de paradigma en las investigaciones científicas, ya que permiten identificar y cuantificar elementos presentes en las células que previamente no habían sido descritos. Adicionalmente, sustentan a los abordajes *in silico* al poder reutilizar los datos de forma reiterada para la evaluación de diferentes hipótesis tras la obtención de los mismos[29,30].

El almacenamiento de los datos procedentes de las tecnologías de alto rendimiento supone un gran reto para la comunidad científica. Esta circunstancia se debe, en parte, a la gran popularidad del uso de estas metodologías, así como a la ingente cantidad de datos generados cada vez que son aplicadas. Asimismo, existe un gran impulso hacia la estandarización y la accesibilidad de forma gratuita de los recursos generados, por lo que actualmente existen numerosas bases de datos públicas que permiten la descarga de los datos generados en diferentes líneas de investigación[31].La primera técnica ómica descrita en el ámbito científico fue la genómica; procedimiento que permite conocer la secuencia de todas las moléculas de ácido desoxirribonucleico (ADN) presentes en una muestra[32]. La función principal del ADN es almacenar la información genética que se transmite a lo largo de las generaciones. En la célula, las moléculas de ADN se localizan en el núcleo y las mitocondrias; y están compuestas por dos cadenas antiparalelas de nucleótidos sucesivos, los cuales se diferencian de acuerdo a la base nitrogenada que los componen [adenina (A), citosina (C), guanina (G) o timina (T)][33]. Por ello, secuenciar el ADN consiste en identificar la disposición ordenada de estos nucleótidos en las moléculas presentes en la muestra analizada[32].

El dogma central de la biología molecular propuesto por Francis Crick en 1970 establece los sucesos que acontecen en la célula para la obtención de proteínas; moléculas encargadas de realizar la gran mayoría de funciones celulares (*Figura 5*)[34]. Concretamente, se generan copias de las moléculas de ADN a través del proceso de replicación. Mediante la transcripción, fragmentos de la secuencia de ADN se utilizan como molde para generar moléculas de ácido ribonucleico (ARN), también conocidos como transcritos. Estos están conformados por una secuencia de nucleótidos cuyas bases nitrogenadas pueden ser adenina (A), citosina (C), guanina (G) o uracilo (U). Si la fracción del ADN utilizada corresponde con un gen, el ARN obtenido se denomina ARN mensajero (ARNm), el cual se utiliza de nuevo como molde para codificar proteínas mediante el proceso de traducción. Las tecnologías de alto rendimiento conocidas como transcriptómica y proteómica permiten obtener datos, respectivamente, del conjunto de moléculas de ARN y proteínas que se encuentran presentes en la muestra analizada[33]. En concreto, este trabajo se centra en el análisis de datos de **transcriptómica** procedentes células y núcleos individuales.

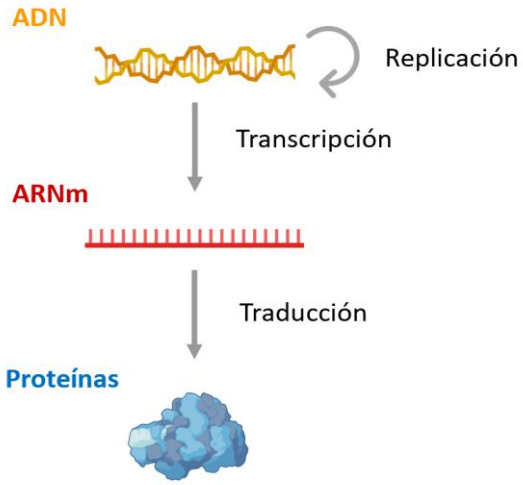


Figura 5. Dogma central de la biología molecular. *Sucesos que conforman el proceso de transmisión de la información genética desde la molécula de ADN hasta la codificación de proteínas. Fueron planteados por primera vez por Watson Crick en 1970. Figura adaptada de J. A. Lasky-Su et al. 2019*[35]*.*

2.2.1. Transcriptómica

La transcriptómica es la disciplina que permite conocer la secuencia de las **moléculas de ARN** presentes en las muestras, tanto codificantes como no codificantes de proteínas. Entre las técnicas utilizadas destaca la secuenciación de ARN (RNA-seq, por sus siglas en inglés *Ribonucleic acid Sequencing*)[32], la cual es descrita a continuación (*Figura 6*).

El primer paso de la metodología de RNA-seq consiste en la extracción del ARN presente en una muestra de interés. Seguidamente, las moléculas obtenidas pueden ser fragmentadas y son utilizadas como molde para la obtención de moléculas de ADN mediante un proceso conocido como transcripción reversa. Asimismo, se realizan una serie de técnicas experimentales adicionales que varían en función de los requerimientos de la tecnología utilizada. Como resultado, se obtienen moléculas que presentan las características necesarias para el proceso de secuenciación. En este paso las muestras procesadas se incorporan en una máquina denominada secuenciador, la cual determina para cada molécula la sucesión ordenada de nucleótidos por la que está conformada. Este procedimiento se realiza automáticamente y de forma simultánea para múltiples moléculas y muestras; situación que permite la obtención masiva de datos[36,37].

Los siguientes pasos del análisis se realizan a nivel bioinformático. La secuencia obtenida de cada molécula evaluada se denomina lectura. Conocer a qué transcrito (o región del ADN) del organismo corresponde cada lectura se efectúa a través de alineamientos. Para ello, se utiliza un genoma (conjunto de genes) o transcriptoma (conjunto de transcritos) de referencia de la misma especie dónde se encuentran anotados los elementos de interés de la secuencia. Por similitud de secuencias se localiza cada lectura en el genoma o transcriptoma de referencia, asignándole el nombre de la región correspondiente. En

caso de no disponer de referencias, las lecturas se van uniendo de forma ordenada de acuerdo con los solapamientos de secuencia encontrados (ensamblaje *de novo*)[36,37].

Una vez identificado el transcrito al que corresponde cada lectura se cuantifica cuántas lecturas se han obtenido por transcrito (y/o gen); valor utilizado como nivel de expresión de la respectiva región del genoma. Por último, los datos se procesan en función del propósito de los investigadores. Por ejemplo, si se analizan muestras sujetas a diferentes condiciones, sería de interés identificar diferencias de expresión génica entre ambos estados a través de un análisis de expresión diferencial[36,37].

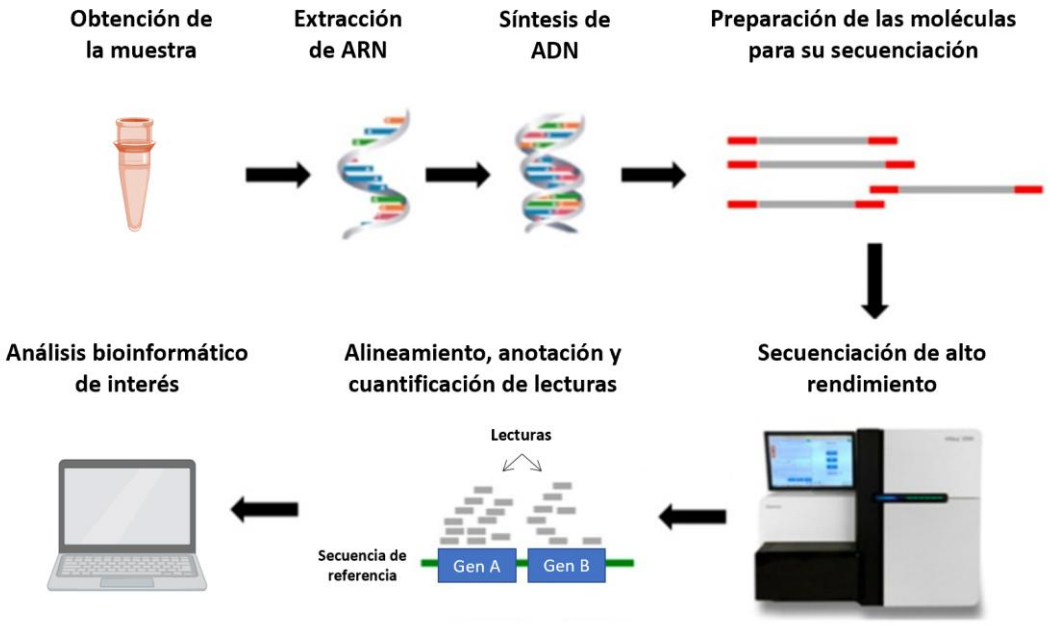


Figura 6. Representación esquemática de un procedimiento estandarizado de RNA-seq. *Ejemplo de abordaje para la identificación y cuantificación de las moléculas de ARN presentes en las muestras de interés mediante secuenciación de alto rendimiento. Figura adaptada de A. Natsume et al. 2019[38].*

2.2.2. Aplicación de las técnicas transcriptómicas en células individuales

El interés creciente en estudiar a nivel celular los sistemas biológicos bajo condiciones concretas ha permitido el desarrollo de metodologías destinadas al **aislamiento de células** procedentes de una muestra, así como a la adaptación de las tecnologías de alto rendimiento para su posterior procesamiento[39].

En concreto, la técnica **scRNA-seq** (por sus siglas en inglés *single cell Ribonucleic acid Sequencing*) se fundamenta en la secuenciación masiva de las moléculas de ARN presentes en células individuales de manera independiente. Cuantificar el nivel de expresión de los transcritos con este nivel de resolución presenta múltiples ventajas. Entre ellas, destaca la evaluación de la heterogeneidad celular al poder detectar células presentes en baja proporción o con niveles de expresión génica reducidos. Además,

mejora la precisión en la descripción de los fenómenos biológicos, siendo posible asignar a cada tipo celular las funciones y/o alteraciones estudiadas de forma específica[39,40].

El procedimiento experimental desarrollado en scRNA-seq se corresponde al descrito previamente para RNA-seq (*Figura 6*) con diversas modificaciones que permiten el estudio de células únicas. La principal diferencia se encuentra en la necesidad de aislar las células y proporcionar a cada una de ellas los reactivos necesarios. Existen numerosas metodologías para lograr este objetivo, entre las cuales destaca la microfluídica[41]. En este trabajo se utilizan datos obtenidos a partir de la tecnología comercial *10x genomics,* basada en un sistema de microfluídica en la que se aplican las propiedades físicas del agua y el aceite para obtener pequeñas gotas (*Figura 7*). De acuerdo con las concentraciones previas de las suspensiones incorporadas al sistema, se maximiza la posibilidad de que cada una de las gotas generadas contenga una célula aislada y los reactivos necesarios para su procesado[42].

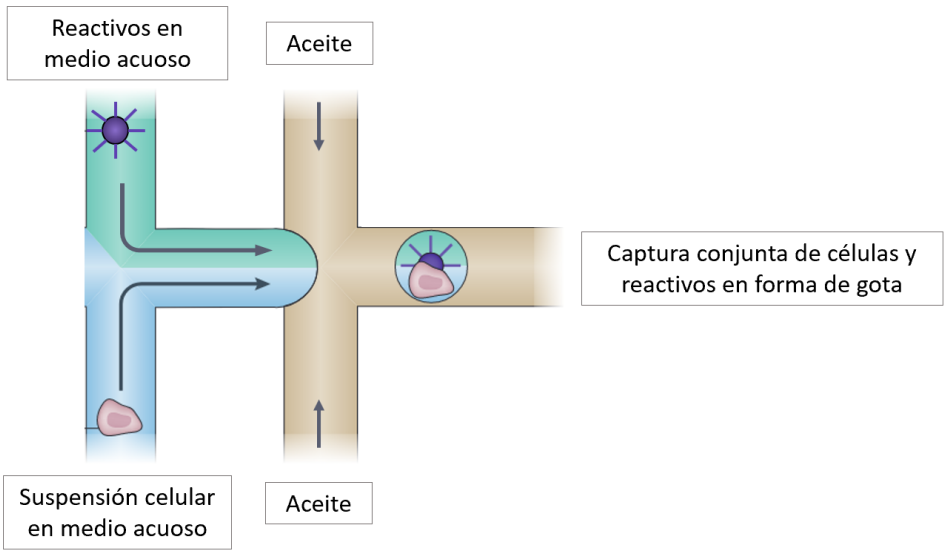


Figura 7. Sistema de aislamiento de células a través de la formación de gotas por microfluídica. *Dispositivo formado por un sistema de tuberías que combina suspensiones de reactivos y de células en medio acuoso para la formación posterior de gotas basada en la interferencia física de un fluido oleoso. Figura adaptada de D. A. Weitz et al. 2017[42].*

El éxito del aislamiento de las células depende en gran medida del tejido a procesar. Ante muestras que presentan dificultades para mantener la integridad física de las células individuales, como las procedentes del SNC, pueden aplicarse alternativas como **snRNA-seq** (por sus siglas en inglés *single nucleus Ribonucleic acid Sequencing*). En esta técnica, el objetivo radica en el aislamiento de los núcleos de las células para secuenciar masivamente las moléculas de ARN que contienen en su interior[43]. Aunque el contenido citoplasmático queda excluido del análisis, diversos estudios establecen que los resultados entre ambas herramientas podrían ser comparables[44,45].

3. ANTECEDENTES Y OBJETIVOS DEL PROYECTO

En la actualidad, la existencia de **diferencias epidemiológicas y clínicas en función del sexo de los individuos** está ampliamente documentada tanto en enfermedades autoinmunitarias como neurológicas. Estas variaciones se han descrito desde múltiples perspectivas, incluyendo la prevalencia e incidencia, el diagnóstico, las manifestaciones clínicas, la progresión de la patología, y la respuesta ante determinados tratamientos[46-48].

Atendiendo a la EM, la prevalencia en mujeres duplica o triplica (en función de la zona geográfica estudiada) el número de hombres afectados. Al evaluar esta ratio a lo largo del tiempo se observa como la proporción de mujeres va incrementando; siendo esta tendencia el resultado de un mayor número de mujeres diagnosticadas y no una disminución en la cifra de hombres afectados[4,9].

En contraposición, la progresión de la enfermedad con sus correspondientes manifestaciones clínicas presenta mayor severidad en hombres. Concretamente, se han identificado alteraciones cognitivas y atrofias del SNC de mayor gravedad[49]. Asimismo, la conversión del subtipo de EMRR a EMSP se produce de forma más temprana y a edades más jóvenes en hombres respecto de mujeres[50]. Por otra parte, la velocidad y severidad de la progresión en mujeres depende de la etapa de la vida en la que se encuentren. Existen indicios de fluctuaciones en los síntomas asociados al ciclo menstrual, así como cambios en el curso de la enfermedad durante la menopausia y el embarazo[49,51].

El abordaje de la perspectiva de sexo en EM desde el ámbito de la biología molecular es limitado y suele plantearse a través de dos vertientes interconectadas: los cromosomas y las hormonas sexuales.

Existen diversos estudios que han analizado el efecto de los **cromosomas sexuales** a través de organismos modelo. Como resultado, se han identificado genes del cromosoma X implicados en procesos de inmunidad, epigenética y neurodegeneración que podrían estar involucrados en las diferencias de sexo reportadas. Adicionalmente, se ha demostrado que se produce un aumento de citoquinas antiinflamatorias en organismos modelo que disponen los cromosomas XY[13,49,52]. La presencia o ausencia del cromosoma Y, y las diferencias de dosis y de expresión específica en los genes del cromosoma X son tres factores de gran interés que aún no disponen de evidencia consolidada para la caracterización de la variabilidad de la EM con perspectiva de sexo[49,50,52].

Por otra parte, las **hormonas sexuales** podrían tener un papel fundamental en la etiología y el mecanismo patogénico de la enfermedad, tanto desde la perspectiva del sistema nervioso como del inmunitario[49,50,52]. Concretamente, la testosterona, la progesterona y estrógenos como el estriol se asocian con efectos neuroprotectores. Por ello, en la actualidad se está analizando su uso como potenciales tratamientos. En las evaluaciones de la testosterona y el estriol, ensayos clínicos revelan una ralentización y mejora de la progresión de la EM en hombres y mujeres, respectivamente[49,52].

Con todo ello, el **objetivo principal** de este trabajo es **mejorar la caracterización a nivel molecular de las diferencias de sexo presentes en la enfermedad de EM.** Para lograr esta finalidad se realiza un abordaje *in silico* que evalúa los niveles de expresión de los diferentes tipos celulares humanos que conforman el sistema inmunitario y el sistema nervioso. Estos son analizados para las condiciones de salud y de enfermedad con perspectiva de sexo a través del cumplimiento de los siguientes objetivos secundarios:

1. Leer libros, revisiones y artículos de diversa índole para lograr una comprensión detallada de la enfermedad de EM, así como del proceso de análisis bioinformático de datos procedentes de las tecnologías scRNA-seq y snRNA-seq.

2. Revisar las bases de datos públicas para identificar datos disponibles de scRNA-seq y/o snRNA-seq referentes a la enfermedad de EM. Posteriormente seleccionar mediante criterios de inclusión y exclusión los estudios de interés para su descarga de los repositorios correspondientes.

3. Analizar de forma individual el conjunto de datos procedentes de cada estudio seleccionado. Concretamente, se desarrollan los siguientes abordajes para cada uno de ellos:
    a. Asignar el tipo celular a cada célula o núcleo en base a los niveles de expresión génica.
    b. Conocer si hay diferencias estadísticamente significativas en la distribución de las poblaciones celulares entre los estados de salud y enfermedad con perspectiva de sexo.
    c. Realizar análisis de expresión diferencial por tipo celular considerando si las células y los núcleos proceden de pacientes de EM o de individuos controles; así como el sexo de los mismos.

4. Integrar los resultados obtenidos del análisis independiente de cada estudio.

4. MATERIALES Y MÉTODOS

Para cumplir con los objetivos específicos dispuestos en el apartado anterior se establece un flujo de trabajo que consta de las siguientes etapas (*Figura 8*): (1) revisión sistemática, (2) análisis primario e (3) integración de los resultados.

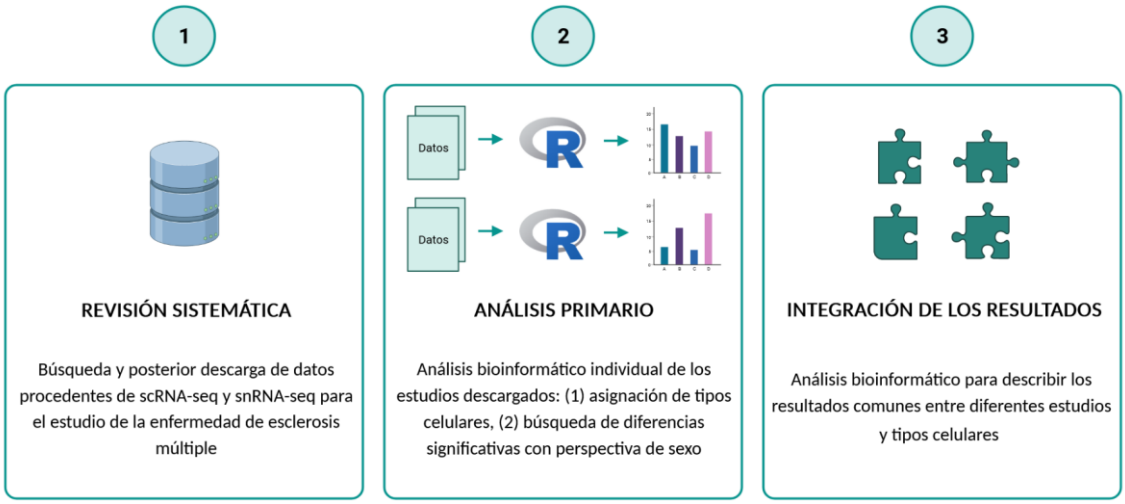


Figura 8. Flujo de trabajo. *Representación esquemática del proceso llevado a cabo en el presente trabajo, junto con una descripción resumida de cada apartado. Figura diseñada en BioRender.com.*

Antes de iniciar la descripción metodológica, en los siguientes párrafos se expone información de interés para mejorar la comprensión del manuscrito.

En primer lugar, **ZENODO** (https://zenodo.org/) es un repositorio que impulsa a nivel europeo el acceso público a la información generada durante el proceso de investigación. Este recurso se ha utilizado para depositar el código diseñado durante el trabajo (https://doi.org/10.5281/zenodo.5068598) y los anexos con las tablas y figuras suplementarias del manuscrito (https://doi.org/10.5281/zenodo.5068587). A través del DOI indicado entre paréntesis, cualquier usuario puede acceder a estos materiales.

En concreto, el código para realizar el análisis bioinformático ha sido desarrollado con el **lenguaje de programación *R***[53] (versión 4.0.2). La versión de los paquetes utilizados en el trabajo, los cuales son citados a lo largo de este manuscrito, puede consultarse en la tabla I.1. del anexo I (10.5281/zenodo.5068587).

Por otra parte, el gran volumen de datos generado en las estrategias de scRNA-seq y snRNA-seq requiere de una capacidad de almacenamiento y de procesamiento difícilmente disponible en los ordenadores personales. Por ello, el análisis bioinformático se ha realizado en la **infraestructura computacional ubicada en el Centro de Investigación Príncipe Felipe (CIPF)** (https://bioinfo.cipf.es/ubb/cluster/). Este *clúster* de computadores está conformado por 44 nodos que permiten disponer de 600 unidades centrales de procesamiento (CPUs, por sus siglas en inglés *Central Processing Unit*). En conjunto, presenta una capacidad de almacenamiento de 1 *PetaByte* y una memoria RAM (por sus siglas en inglés *Random Access Memory*) de 11 *TeraBytes*. Para poder trabajar

de forma ordenada con el resto de los usuarios, los trabajos se han ejecutado a través del sistema de colas SLURM (por sus siglas en inglés *Simple Linux Utility for Resource Management*).

Por último, para la lectura de los materiales y métodos (apartado 4) y resultados (apartado 5) es necesario realizar dos apreciaciones lingüísticas. La primera de ellas es que se ha empleado la palabra "célula" referenciando indistintamente a células (scRNA-seq) o a núcleos (snRNA-seq). Asimismo, se ha utilizado el término "gen" para aquellas regiones del genoma de las que se han cuantificado transcritos, independientemente de si estos codifican o no proteínas. En ambos casos se indicará de forma explícita si las palabras presentan una connotación distinta a la planteada.

4.1. Revisión sistemática

La revisión sistemática es el procedimiento que permite recopilar la evidencia científica descrita hasta el momento sobre un tema de interés[54]. Este proceso se ha implementado en el trabajo atendiendo a las **directrices PRISMA** (por sus siglas en inglés *Preferred Reporting Items for Systematic reviews and Meta-Analyses*), que establecen un marco común para favorecer la estandarización y reproducibilidad de la revisión sistemática, así como de la posterior síntesis de los resultados[55].

El primer paso a realizar corresponde con la definición de una pregunta de investigación de forma clara y concisa. Una vez establecida, se determinan criterios de inclusión que enmarcan las características esenciales que deben cumplir los estudios. Seguidamente, se realiza una búsqueda exhaustiva en diferentes bases de datos, filtrando de acuerdo con los requerimientos previamente establecidos (fase de identificación). Tras su recopilación, los estudios son validados de forma individual comprobando que no existe ningún motivo para su exclusión (fases de revisión y elegibilidad). Finalmente, los datos de los estudios seleccionados son descargados para realizar el análisis bioinformático que permita responder a la pregunta inicialmente planteada (fase de inclusión)[54,55].

En detalle, se han recopilado estudios que permiten caracterizar las bases moleculares por tipo celular de la enfermedad de EM en función del sexo del individuo. La búsqueda fue realizada en las bases de datos públicas **GEO** (por sus siglas en inglés *Gene Expression Omnibus*)[56], **ArrayExpress**[57] y **UCSC Cell Browser** (por sus siglas en inglés *University of California Santa Cruz Cell Browser)*[58], así como en la herramienta de búsqueda **Google***.*

Los **criterios de inclusión** utilizados fueron adaptados a los formatos de los repositorios:

- Para GEO y ArrayExpress se realizó una búsqueda avanzada con las palabras clave "*multiple sclerosis"* y "*single cell"* o "*single nuclei"* o "*single nucleus",* y como organismo "*Homo sapiens".* En el caso de GEO se indicó como tipo de estudio "*Expression profiling by high throughput sequencing"* mientras que en ArrayExpress se seleccionó en el apartado de tecnologías "*RNA assay"* y *"sequencing assay".*

- En la plataforma UCSC Cell Browser se revisó el listado de datos disponibles seleccionando aquellos relacionados con la enfermedad EM procedentes de humanos.

- Por último, se introdujeron en el buscador Google las palabras clave "*multiple sclerosis"* y "*scRNA-seq*" en una búsqueda, y "*multiple sclerosis"* y "*snRNA-seq*" en otra. En ambos casos, se revisó la bibliografía resultante con el objetivo de identificar los estudios realizados en humanos que contuviesen datos asociados.

Tras este proceso, cada uno de los estudios seleccionados fue evaluado de forma manual, eliminando aquellos que cumplían con alguno de los siguientes **criterios de exclusión:**

- Metodología: no utilizar la técnica de scRNA-seq o snRNA-seq.

- Enfermedad evaluada: no presentar como objetivo principal el estudio de la enfermedad de EM.

- Diseño experimental: disponer solamente de individuos controles o de pacientes de EM.

- Sexo: no conocer el sexo del individuo del que procede cada muestra.

- Tamaño muestral: estudiar un número insuficiente de individuos. Como mínimo, cada grupo (mujer enferma, hombre enfermo, mujer sana y hombre sano) deberá contener al menos 3 individuos.

Los datos de los estudios que finalmente fueron seleccionados se descargaron y almacenaron en la infraestructura computacional del CIPF. Concretamente, se obtuvo el nivel de expresión de las distintas regiones del genoma a través de la **matriz de conteos**. En esta estructura bidimensional cada fila corresponde a un gen y cada columna a una célula. Cada entrada fila-columna de la matriz es un valor entero (conteo) que indica el número de lecturas obtenidas en el proceso de secuenciación para ese gen en la correspondiente célula (explicación asociada al apartado de la introducción 2.2.1.).

Asimismo, se descargaron 3 ficheros adicionales para cada registro, los identificadores de las células, los metadatos de los genes y los metadatos de las células respectivamente. El término metadatos hace referencia al conjunto de variables que permiten caracterizar los elementos de interés (por ejemplo, los datos sobre el sexo del individuo del que procede la célula).

4.2. Análisis primario de los estudios

Este apartado se centra en el análisis bioinformático individual de los datos descargados de cada estudio seleccionado tras la revisión sistemática (*Figura 9*). En resumen, el algoritmo establecido pretende caracterizar a qué tipo celular pertenece cada célula en base a sus perfiles de expresión. Una vez asignados, se evalúa si existen diferencias estadísticamente significativas en las proporciones de los diferentes tipos celulares, así como en los niveles de expresión génica de forma generalizada y por tipo celular. Con este objetivo, los análisis son realizados atendiendo a la condición (EM, control) y al sexo (hombre, mujer) de los individuos utilizando la siguiente nomenclatura:

- EM_Mujer: mujeres enfermas que sufren EM.
- Control_Mujer: mujeres sanas.
- EM_Hombre: hombres enfermos que sufren EM.
- Control_Hombre: hombres sanos.

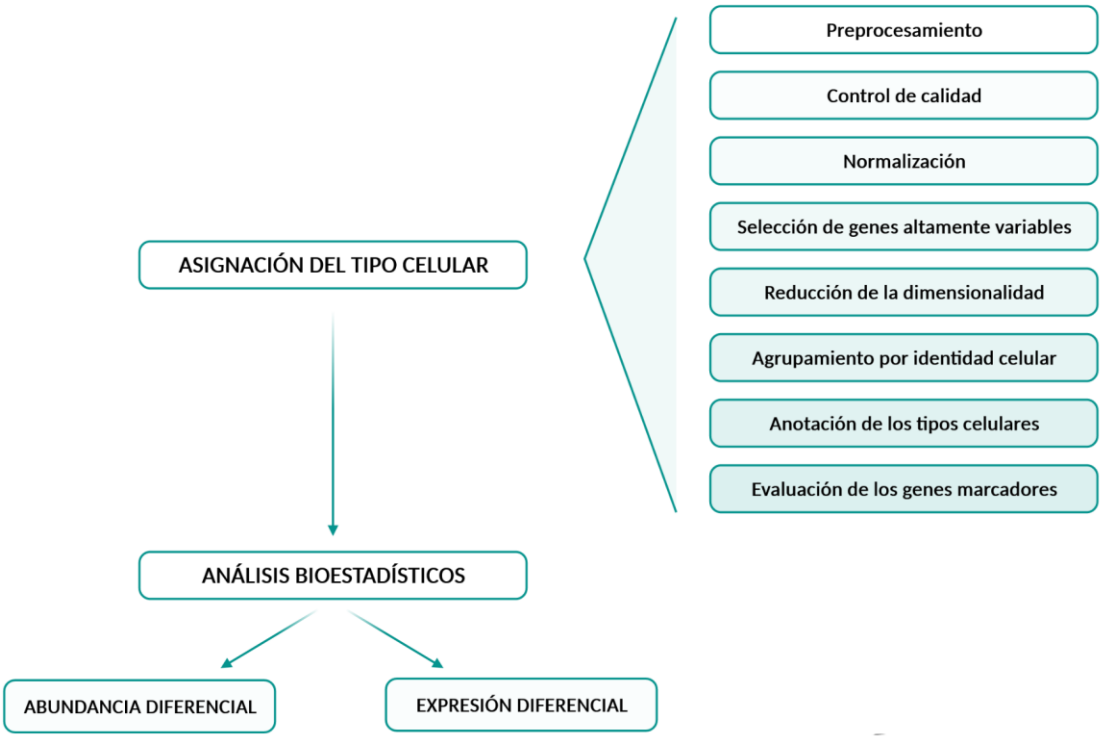


Figura 9. Flujo de trabajo para el análisis primario. *Esquematización de los pasos realizados para el análisis individual de cada uno de los estudios. Los análisis establecidos para la asignación del tipo celular se disponen de forma secuencial. Figura diseñada en BioRender.com.*

Los análisis bioinformáticos de los datos procedentes de los abordajes scRNA-seq y snRNA-seq son prácticamente equivalentes. La principal diferencia radica en considerar

la ausencia del citoplasma en la segunda metodología. Del mismo modo, el abordaje computacional utilizado en cada estudio debe ser lo más homogéneo posible; evaluando simultáneamente sus peculiaridades de forma concreta. Por tanto, en los siguientes subapartados si no se realiza una mención específica, la metodología expuesta habrá sido aplicada en el análisis de todos los estudios; independientemente de si proceden de scRNA-seq o snRNA-seq.

La creciente popularidad de los análisis transcriptómicos de célula y núcleo únicos ha permitido el desarrollo de numerosas herramientas bioinformáticas específicas para estas tecnologías. Sin embargo, esta área de investigación sigue caracterizándose por la falta de estandarización en las metodologías empleadas[59]. Entre los paquetes de R más utilizados se encuentra por una parte *Seurat*[60], y por otra s*cater*[61] junto a s*cran*[62]*.* En ambos casos, los paquetes integran numerosas funciones con el objetivo de proporcionar las herramientas necesarias para unificar la ejecución de los diferentes pasos del algoritmo.

En este trabajo se ha optado por elaboración del código con las funciones procedentes de los paquetes s*cater* y *scran* por los siguientes motivos: (1) permite el trabajo con datos almacenados en el objeto S4 de la clase *SingleCellExperiment*[63]*,* el cual es utilizado en numerosos paquetes de Bioconductor; situación que favorece la interoperabilidad entre paquetes generando un código más robusto y (2) son utilizados en la bibliografía enfocada a la estandarización del procesamiento de datos de scRNA-seq empleada en este trabajo[59,63,64].

4.2.1. Preprocesamiento

El primer paso para el análisis individual de los estudios es almacenar los datos descargados en un objeto de la clase *SingleCellExperiment* (*Figura 10*). Concretamente, la matriz de conteos se deposita en el apartado *assays;* los metadatos de los genes en *rowData* y los metadatos de las células en *colData.* Para ello, se utiliza la función *SingleCellExperiment* ubicada en el paquete de nombre homónimo[63]. Es de suma importancia indicar previamente los identificadores de las filas (genes) y columnas (células) en la matriz de conteos; asegurando que estos coinciden con los utilizados en los metadato*s.*

Además de recopilar todos los datos de partida en un solo objeto, la clase *SingleCellExperiment* permite almacenar muchos de los resultados obtenidos en los pasos posteriores. Adicionalmente, multitud de funciones se encuentran diseñadas para identificar los datos de entrada que necesitan y depositar los datos de salida en los correspondientes apartados. Esta automatización de las funciones facilita en gran medida el procedimiento desarrollado por el investigador.

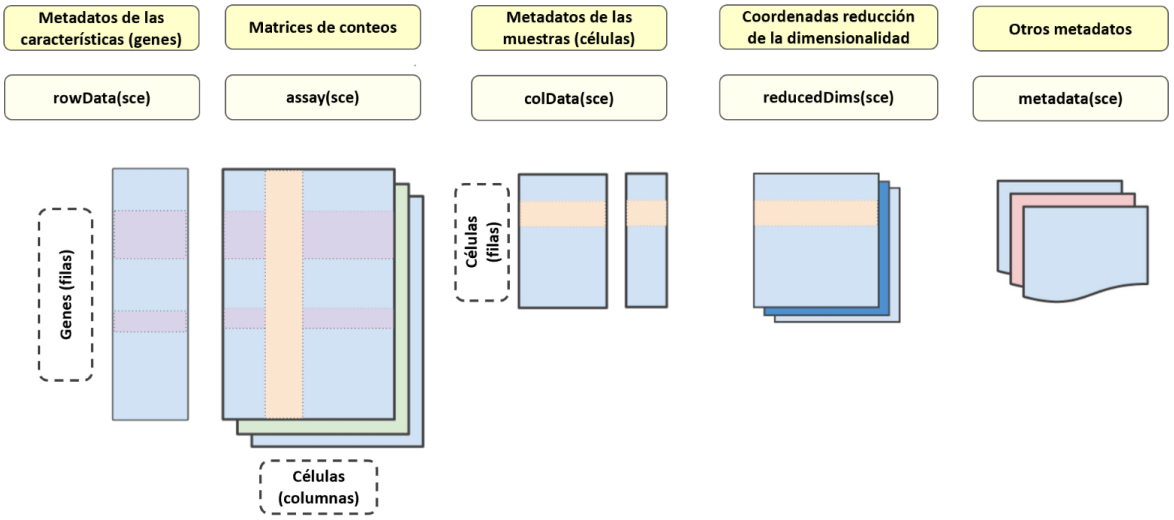


Figura 10. Categorías para el almacenamiento de datos en la clase SingleCellExperiment. *Ordenación en diferentes compartimentos de acuerdo con la naturaleza de los datos (rótulos amarillos), los cuales pueden ser extraídos ejecutando el código correspondiente (rótulos amarillos claros). Representación gráfica de la disposición necesaria de las filas y columnas para cada compartimento. Franja morada: datos de un gen para cada una de las células; franja beige: datos de todos los genes para una célula; sce: nombre de ejemplo proporcionado a un objeto SingleCellExperiment. Figura adaptada del curso M. Hemberg et al.*[65].

4.2.2. Control de calidad

Tras incorporar los datos en el formato adecuado, se realiza una evaluación de la calidad de las células. El principal fin de este paso consiste en **incorporar en el análisis solamente aquellas columnas de la matriz de conteos que procedan de una única célula viable**.

Por diversos motivos tanto biológicos como técnicos, puede ocurrir que durante el aislamiento de las células se hayan procesado gotas vacías, células dañadas o agregados celulares. Asimismo, durante la preparación de las muestras y la secuenciación de las moléculas de ARN pueden darse fallos que, en último término, representen datos de expresión génica totalmente diferentes a la realidad biológica de las células[59,63].

Toda esta casuística se puede estudiar a través de diversos **indicadores** que permitan identificar aquellos genes y/o células que no cumplan con los umbrales de calidad establecidos, que serán eliminados del análisis. Es importante resaltar que el establecimiento de los puntos de corte para cada indicador depende en gran medida de cómo se haya realizado cada experimento, por lo que debe adaptarse de forma individual a cada estudio evaluado[59,63].

Los indicadores calculados para identificar gotas vacías, muestras procedentes de células dañadas o que han sufrido algún problema durante la fase de preparación y/o secuenciación son: el tamaño de la librería (suma total de conteos), el número de genes

que disponen de conteos, el porcentaje de genes mitocondriales y el porcentaje de genes ribosomales; todos ellos calculados por célula. Concretamente, se considera que aquellas células que presentan un número bajo de conteos y/o de genes expresados han perdido parte del citoplasma o son gotas vacías con presencia de contaminación ambiental. El mismo razonamiento es aplicado cuando se detectan células con un porcentaje elevado de genes mitocondriales. Por su parte, un alto porcentaje de genes ribosomales puede ser indicativo de que una fracción de ARN ha sido degradada[59,63].

Para la identificación de agregados celulares se ha utilizado la función *scDblFinder* incorporada en el paquete del mismo nombre[66]. Esta función está programada para clasificar a las células en grupos en base a sus perfiles de expresión. Seguidamente, simula en repetidas ocasiones que en una gota se encuentran dos células (doblete) sumando los conteos de 2 en 2 células; tanto pertenecientes al mismo grupo como a distintos. Finalmente compara los perfiles de expresión de las células del estudio con los dobletes creados artificialmente. De este modo se conoce si el ARN procede de una célula o de un agregado.

A nivel de gen, para cada uno de ellos se determina el número de células en las que se ha detectado su expresión. Si este valor es muy reducido, el gen es descartado al considerarse que no presenta transcritos en las muestras analizadas. Asimismo, se eliminan todos aquellos elementos del genoma que no sean genes codificantes de proteínas, microRNAs *o* lncRNAs. Para su identificación se han utilizado las guías de nomenclatura del identificador *Gene Symbol* establecidas por HGNC (por sus siglas en inglés *HUGO Gene Nomenclature Committee*)[67].

Con todo ello, para cada estudio se han calculado los indicadores citados en los párrafos anteriores. Seguidamente, estos se han representado gráficamente para explorar su distribución por condición, sexo, conjuntamente por condición y sexo, y por muestra de procedencia. A continuación, se han determinado y aplicado los puntos de corte de cada indicador para identificar las células a ser descartadas. Los umbrales pueden ser valores fijos que determine el investigador al interpretar el análisis exploratorio previo. También pueden ser adaptativos, haciendo uso de funciones como *isOutlier* del paquete *scuttle*[61]. Concretamente, esta función calcula la desviación media absoluta (MAD, por sus siglas en inglés *Median Absolute Deviation*) (promedio de todas las diferencias entre la media aritmética y cada valor en términos absolutos). Por defecto, se descartan las células que presentan valores del indicador evaluado más alejados de 3 MADs.

Finalmente, se descartan las células que queden excluidas por encontrarse fuera de los puntos de corte para cualquiera de los indicadores evaluados. Asimismo, se vuelven a representar gráficamente los indicadores confirmando que el proceso se ha realizado correctamente.

Cabe destacar que en este apartado se ha determinado la **fase del ciclo celular** en la que se encontraba cada célula cuando fue procesada. Pese a que no se ha utilizado como indicador de la calidad, este metadato es empleado posteriormente para conocer qué fracción de la variabilidad de la expresión génica puede deberse a que las células se encuentren en distintas fases del ciclo celular. Para ello, se ha utilizado la función *cyclone*[68] del paquete *scran*[62]. Esta función cuenta con un listado de parejas de genes

para las fases G1, S, G2-M. En cada pareja, el primer gen presenta una mayor expresión en la fase asignada respecto al resto de fases. Con ello, se determina para cada célula las proporciones de parejas de genes que cumplen con dicha situación en todas las fases evaluadas. Los valores obtenidos son transformados en puntuaciones, las cuales se utilizan para determinar la fase del ciclo celular.

4.2.3. Normalización

Cada paso del procedimiento experimental destinado a la obtención de las lecturas es repetido para cada célula o núcleo analizado. Suponiendo que dos células idénticas fueran procesadas, los resultados que observaríamos en la matriz de conteos no serían equivalentes. Estas diferencias, ajenas a las características biológicas de las células, se conocen como **variabilidad técnica**. El objetivo principal de la normalización consiste en mitigar su efecto, para que los responsables de la diversidad caracterizada sean efectos biológicos y no fuentes de variabilidad intrínsecas al procesamiento de las muestras[59,63].

La metodología más habitual para realizar la normalización es corregir los valores de la matriz de conteos por un **factor de escalado** (denominado comúnmente en inglés *size factor*). Este factor es específico para cada célula y representa el sesgo técnico de su procesamiento. Al aplicar la normalización, se dividen los conteos de todos los genes de cada célula por su correspondiente factor de normalización, asumiendo que la variabilidad técnica afectará a todos ellos por igual.

El factor de escalado puede calcularse a través de diversos procedimientos. Entre ellos, destaca por su simplicidad el método de **normalización por el tamaño de la librería**. Esta estrategia, utilizada con asiduidad en los análisis de RNA-seq, asume que el número de moléculas de ARN presentes en cada célula (y, por tanto, el tamaño de las librerías en la matriz de conteos) es el mismo. Con ello, las diferencias en este valor serían consecuencia de la variabilidad técnica introducida durante el procesamiento de las mismas[59,63]. El cálculo del factor de escalado por el tamaño de la librería se ha realizado en este trabajo con la función *librarySizeFactors* del paquete *scuttle*[61].

Sin embargo, las matrices de conteos no son normalizadas con este factor de escalado. Pese a que para determinadas aplicaciones se utiliza en los datos de scRNA-seq y snRNA-seq, este método de normalización suele ser insuficiente por dos motivos. El primero de ellos es un sesgo técnico específico de estas tecnologías conocido como *dropout.* Al trabajar como unidades experimentales células individuales, la cantidad de ARN a procesar es muy limitada. Esta situación conlleva que la expresión de numerosos genes no sea detectada, por lo que las matrices de conteos se caracterizan por presentar una elevada proporción de 0. Por otra parte, para cada célula los niveles de expresión, tanto a nivel cualitativo como cuantitativo, serán diferentes. Como consecuencia, no es asumible esclarecer *a priori* que el tamaño de la librería debe ser el mismo en todas las células[59,63].

Para considerar ambos aspectos en el proceso de normalización, se calcula el factor de escalado por el **método de deconvolución** (*Figura 11*)[69]. En primer lugar, se generan

grupos de células con la función *quickCluster* del paquete *scran*[62]. Seguidamente, se calculan los factores de escalado con la función *computeSumFactors* del mismo paquete. Este procedimiento se basa en sumar los conteos por gen de las células que conforman cada grupo creando *pseudocélulas.* Adicionalmente, se generan otras *pseudocélulas* sumando conteos por gen entre células de distintos grupos. A continuación, se calcula el factor de escalado para cada *pseudocélula* como la ratio del tamaño de la librería de la *pseudocélula* respecto a la media. A partir de este valor se determinan los factores de escalado para cada célula por un sistema de ecuaciones lineales.

Tras calcular el factor de escalado, se normalizan los datos de la matriz de conteos y se transforman a escala logarítmica en base 2 con la función *logNormCounts* del paquete *scuttle*[61].

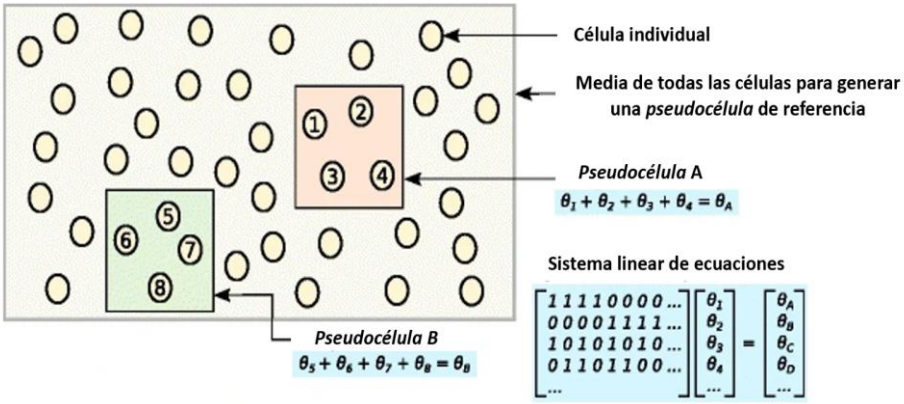


Figura 11. Representación gráfica del cálculo de los factores de escalado por el método de deconvolución. *Ejemplificación de la creación de dos pseudocélulas, cada una de ellas conformada por la suma de los niveles de expresión de cuatro células diferentes. Se calcula el factor de escalado para cada pseudocélula a partir del tamaño de la librería respecto de la pseudocélula de referencia; "deconvoluyendo" posteriormente el valor para cada célula real del estudio a través de un sistema de ecuaciones lineales. Figura adaptada de J. C. Marioni et al*[69]*.*

4.2.4. Selección de genes altamente variables

Las dimensiones de las matrices de conteos suelen ser de gran magnitud, por lo que hay que tener presente que en un único estudio pueden evaluarse los niveles de expresión de miles de genes para cientos de miles de células. Esta situación ocasiona que las ejecuciones posteriores para determinar el tipo celular tengan un coste computacional ingente y elevados requerimientos de almacenamiento. Asimismo, no todos los datos presentes en las matrices de conteos son de interés, ya que el análisis de gran cantidad de genes introduce ruido junto con señal biológica no relevante[59,63].

Por todo ello, es necesario realizar una **selección previa de los genes** cuyos niveles de expresión presentan **mayor variabilidad biológica** entre las células de cada estudio. De

forma estandarizada, suele definirse un intervalo entre 1.000 y 5.000 genes. Si *a priori* se tuviese conocimiento sobre los genes más informativos para la variabilidad biológica a estudiar, estos podrían ser seleccionados. Sin embargo, esta situación no se da en el trabajo. Por ello, la aproximación empleada consiste en (1) ordenar los genes en función de la cantidad de heterogeneidad biológica que pueden explicar sus niveles de expresión para (2) seleccionar el 20% que mayor variabilidad representan[64]. Para ello, se han utilizado diversas funciones del paquete *scran*[62].

Concretamente, se ha calculado la variabilidad de cada gen para posteriormente separarla en sus componentes técnico y biológico. La medida de variabilidad utilizada en este trabajo es la varianza. En detalle, este proceso se ha realizado con la función *modelGeneVar.* Tras calcular la media y la varianza de los valores de expresión por gen normalizados y en escala logarítmica, la función modeliza el componente técnico de la varianza con la función *fitTrendVar*. Seguidamente, calcula la fracción de varianza biológica como la diferencia entre la varianza total y su componente técnico. Del resultado obtenido se determina el 20% de genes que mayor variabilidad biológica representan con la función *getTopHVGs.*

En este punto es importante resaltar que el proceso de normalización está destinado a eliminar la variabilidad técnica y el ruido estocástico que se han introducido durante el procesamiento individual de las células. Sin embargo, estas pueden haber sido procesadas por lotes, aplicando posteriormente diferentes estrategias para conocer qué dato corresponde a cada célula. En este procesamiento conjunto pueden darse diferencias técnicas en los niveles de expresión que afecten al grupo de células analizadas. El fenómeno que describe dicha variabilidad se conoce como **efecto *batch***.

Con todo ello, en el trabajo se ha realizado el procedimiento de selección de los genes altamente variables bloqueando efecto *batch* para descartar la heterogeneidad debida a los lotes de procesamiento[63]. Cuando se proporciona esta información a la función *modelGeneVar* se calcula la descomposición de la varianza por cada lote de forma independiente. El resultado final por estudio, a partir del cual se determina el 20% de los genes altamente variables, corresponde con los valores medios de la varianza total, de sus componentes técnico y biológico, así como de los valores medios de expresión entre los lotes.

4.2.5. Reducción de la dimensionalidad

Tras seleccionar aquellos genes que representan la mayor parte de la variabilidad biológica, se continúa compactando la información que proporcionan los datos a través del proceso de reducción de la dimensionalidad. Este paso consiste en combinar los niveles de expresión de distintos genes para abreviar el número de características analizadas.

El primer motivo que refleja su necesidad es conseguir **resumir la variabilidad biológica** que representan los genes altamente variables seleccionados en el paso anterior. De este modo se simplifica la complejidad y se reduce la redundancia debida, por ejemplo, a

patrones de coexpresión. Concretamente, los pasos siguientes del análisis consideran cada característica analizada como una dimensión independiente. Por tanto, si se trabaja a nivel de los genes altamente variables en cada ejecución se establecerían miles de dimensiones, atendiendo a los miles de genes evaluados. Tal y como indica el nombre del apartado, reducir las dimensiones a las características de mayor relevancia mejoraría en gran medida el proceso computacional sin alterar las conclusiones biológicas obtenidas[59].

Por otro lado, la representación gráfica de las características obtenidas tras la reducción de la dimensionalidad permite la **visualización de los datos**. Lograr interpretar los niveles de expresión procedentes de miles de genes es una tarea casi imposible para el ojo humano. Aglutinar la variabilidad biológica que estos representan en un número muy reducido de características favorece enormemente la labor investigadora[59].

En este trabajo, con motivo de resumir y favorecer la visualización de los datos, se ha realizado un análisis de componentes principales (PCA, por su acrónimo en inglés *Principal Component Analysis*). Adicionalmente, se han aplicado dos estrategias adicionales de visualización: tSNE (por su sigla en inglés *T-distributed Stochastic Neighbor Embedding*) y UMAP (por su sigla en inglés *Uniform Manifold Approximation and Projection*).

4.2.5.1. Análisis de componentes principales (PCA)

La metodología PCA pretende minimizar el número de características a evaluar a través de una combinación lineal de los niveles de expresión génica. Es decir, se calcula un número limitado de valores por célula, cada uno de los cuales está constituido por una combinación lineal de los niveles de expresión de un conjunto de genes correlacionados. Estos valores se denominan **componentes principales**, y presentan como objetivo recopilar la mayor variabilidad posible de los datos. En detalle, los componentes principales están ordenados en función del porcentaje de variabilidad explicado: el componente principal 1 reúne la mayor proporción de variabilidad, seguido por el componente principal 2, etc.

Se asume que la heterogeneidad debida a los aspectos biológicos relevantes está englobada en los niveles de expresión de multitud de genes, muchos de los cuales actuarán de forma coordinada. Por tanto, se establece que los primeros componentes principales recopilarán variabilidad biológica de interés, mientras que los restantes incluirán ruido técnico y diversidad biológica no relevante[63]. En este punto surge la pregunta ¿cuántos componentes principales son necesarios?

Esta situación plantea un compromiso entre arriesgar la pérdida de información biológica (selección de insuficientes componentes principales) e incluir aspectos no relevantes que puedan enmascarar la variabilidad de interés (selección de un exceso de componentes principales). La estrategia aplicada en este trabajo se conoce comúnmente como el **método del codo;** correspondiente a seleccionar los componentes principales que llegan hasta la curva que se observa en el gráfico de sedimentación[63]. Esta representación es un

diagrama que refleja el porcentaje de variabilidad explicada por cada componente principal.

A nivel computacional, se han calculado los primeros 50 componentes principales para cada estudio con la función *runPCA* del paquete *scater*[61]. Como dato de entrada, se han utilizado los niveles de expresión normalizados a escala logarítmica de los genes altamente variables. Seguidamente, se ha extraído el porcentaje de varianza que explica cada componente principal y, con la función *findElbowPoint* del paquete *PCAtools*[70], se ha determinado el número de componentes principales de interés por el método del codo. Finalmente, se representaron de forma gráfica los resultados obtenidos.

4.2.5.2. *T-distributed Stochastic Neighbor Embedding* (tSNE)

Esta metodología tiene como objetivo conseguir que la disposición de las células en los gráficos de dimensiones reducidas sea lo más similar a su distribución en función de los niveles de expresión de cada gen. En primer lugar, se calcula la distancia entre las células en el gráfico multidimensional, donde cada dimensión corresponde al nivel de expresión de un gen. Para las células i,j, la distancia se establece como la probabilidad de que la célula j sea la vecina más próxima de la célula i en base a una distribución-t. Seguidamente, se distribuyen las células de forma aleatoria en el espacio de dimensiones reducidas (generalmente bidimensional). Mediante un proceso iterativo, se van modificando sus posiciones en el gráfico y recalculando las distancias hasta que el resultado converge con el del espacio multidimensional (*Figura 12*)[71].

Por todo ello, al representar las coordenadas bidimensionales calculadas para cada célula se puede interpretar que las células cercanas en el espacio bidimensional también lo están en el espacio multidimensional. Sin embargo, esta situación no puede ser asumida para las células que se encuentren separadas a mayor distancia.

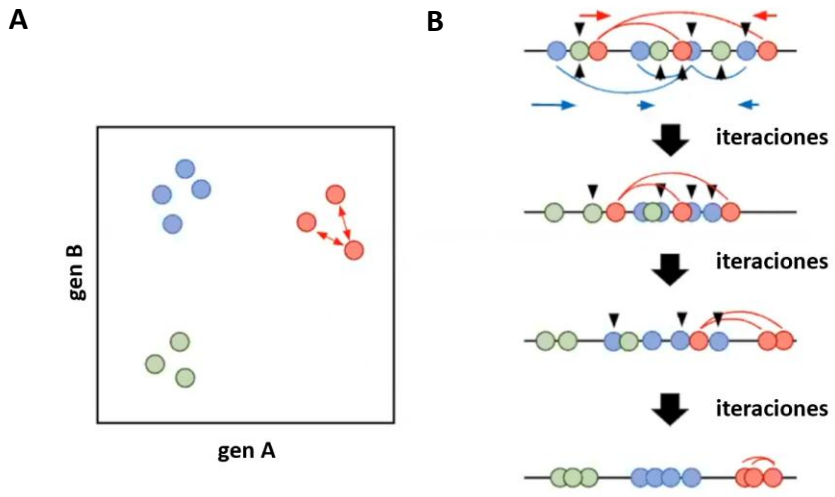


Figura 12. Ejemplificación gráfica del algoritmo de reducción de la dimensionalidad tSNE. *(A) Espacio bidimensional que desea ser reducido a una dimensión. (B)*

Procedimiento de reducción de la dimensionalidad. Disposición inicial de las células en el espacio unidimensional de forma arbitraria para que, tras sucesivas iteraciones, cada célula presente los mismos vecinos que en el espacio bidimensional. Figura adaptada del curso scRNA-seq[72].

Este proceso se ha ejecutado con la función *runTSNE* del paquete *scater*[61] para reducir la variabilidad a dos dimensiones. Debido a que el procedimiento tendría gran coste de recursos y de tiempo, la función está optimizada para utilizar la información de los componentes principales seleccionados por el método del codo. Asimismo, no se calcula la distancia entre todas las células, sino que, en cada iteración, son evaluados un número limitado de vecinos. Este parámetro, denominado *perplexity,* es seleccionado por el investigador.

Como *a priori* no se conoce el número de vecinos que permiten una mejor visualización de los resultados, se ha ejecutado el proceso para valores de *perplexity* 5, 20, 80 y 120. Seguidamente se ha determinado un intervalo y se han testado valores más cercanos entre sí con el objetivo de encontrar el idóneo. Adicionalmente, se ha ejecutado el proceso inicializando la semilla a 100 y 12345. Estas ejecuciones se han utilizado para comprobar que las coordenadas obtenidas para las células pueden variar, ya que inicialmente se disponen de forma aleatoria.

4.2.5.3. *Uniform Manifold Approximation and Projection* (UMAP)

UMAP es una estrategia alternativa a tSNE para la visualización de las células en dimensiones reducidas. En concreto, su base metodológica se centra en calcular la distancia que separa a las células en el espacio multidimensional. En este caso, la distancia puede calcularse con el método que decida el investigador. Seguidamente, se conectan aquellas células cuya distancia es menor a un umbral previamente establecido, generando estructuras topológicas denominadas *simplex* (*Figura 13A, 13B*). Finalmente, se conectan los diferentes *simplex* en base a la distancia relativa que los separa, de modo que las células puedan disponerse en las dimensiones de interés (*Figura 13C*)[73].

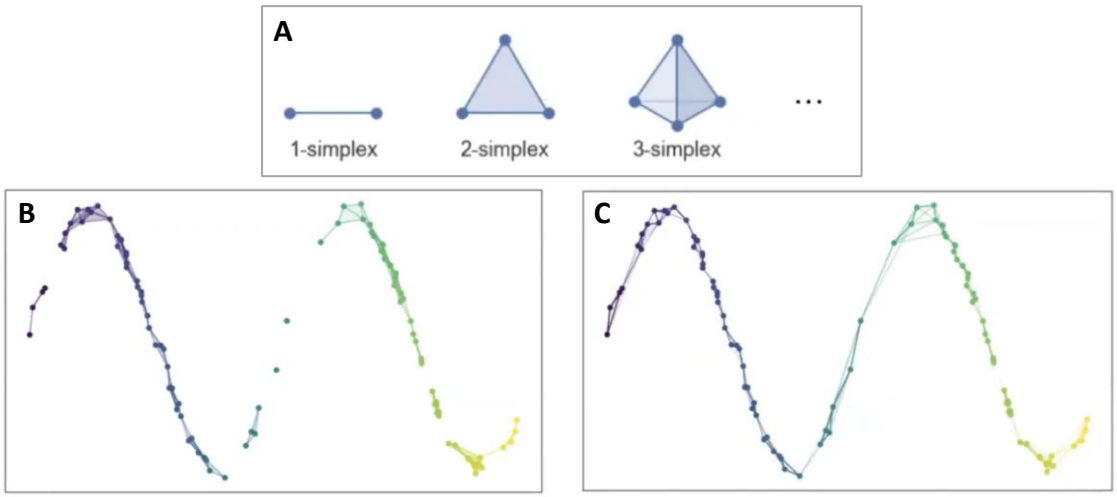


Figura 13. Ejemplificación gráfica del algoritmo de reducción de la dimensionalidad UMAP. *(A) Simplex construidos a partir de dos células (1-simplex), tres células (2-simplex), y cuatro células (3-simplex). (B) Representación gráfica de simplex de ejemplo en dos dimensiones (C) los cuales son conectados entre sí para determinar sus coordenadas relativas en el gráfico. Figuras obtenidas del curso scRNA-seq*[72]*.*

Los gráficos de reducción de la dimensionalidad basados en la estrategia UMAP han sido realizados con la función *runUMAP* del paquete *scater*[61]. Tal y como se ha descrito en la metodología tSNE, el algoritmo utiliza los componentes principales previamente seleccionados para optimizar el proceso. Asimismo, las distancias se calculan entre los vecinos más cercanos, cuyo número es seleccionado por el investigador.

Con todo ello, se han realizado diversas ejecuciones valorando qué número de vecinos y qué semilla son más adecuados utilizar para facilitar la visualización de los resultados.

4.2.6. Agrupamiento por identidad celular

Los componentes principales de mayor relevancia biológica, seleccionados por el método del codo, son utilizados como datos de entrada para clasificar a las células en grupos. De este modo, las células son categorizadas en función de la similitud de sus patrones de expresión génica; **considerando cada grupo como una identidad celular diferente**. Los resultados obtenidos permiten explorar la heterogeneidad biológica al ser representados sobre los gráficos generados en el paso de reducción de la dimensionalidad.

Independientemente de la estrategia utilizada, el objetivo de agrupar las células es conseguir que la variabilidad descrita dentro de cada grupo sea más pequeña que la presente entre grupos. En este trabajo, se ha utilizado la **metodología basada en grafos** por ser escalable al número de células procesadas en los diferentes estudios. Este método de agrupamiento de las células puede dividirse en dos partes. En primer lugar, debe construirse un grafo formado por nodos (células) conectados a través de aristas de acuerdo a la similitud compartida. Seguidamente, se procede a la identificación de

comunidades en el grafo. Estas son definidas como grupos de células que presentan mayor probabilidad de conectarse entre ellas respecto al resto de nodos del grafo[59,63].

En términos computacionales, se ha generado un grafo de acuerdo al algoritmo *Shared Nearest Neighbours* (SNN) con la función *buildSNNGraph* del paquete *scran*[62]. Concretamente, los nodos se distribuyen en base a los componentes principales seleccionados por el método del codo, determinando los *n* vecinos más próximos de cada célula. Seguidamente, las células se conectan si comparten vecinos entre ellas. A mayor número de vecinos compartidos, mayor peso de la arista, y por tanto mayor similitud entre las correspondientes células. En este procedimiento el número de vecinos a evaluar también es seleccionado por el investigador; parámetro que define de forma indirecta el número de grupos identificados.

No existe un número correcto de grupos a definir; ya que el nivel de detalle que se desee alcanzar dependerá de los intereses del investigador. Si se desea obtener una mayor resolución, correspondiente a definir un mayor número de grupos, el número de vecinos a contrastar debe ser pequeño. Por el contrario, si el objetivo es obtener un menor número de grupos estableciendo una clasificación más general, el número de vecinos a testar debe aumentar. En este trabajo, tras realizar varias pruebas, se ha decidido evaluar para todos los estudios los 20 vecinos más cercanos de cada célula.

Tras generar el grafo se buscan las comunidades; definiendo grupos como aquellas regiones con más conexiones (y de mayor peso) entre las células que lo conforman respecto al resto del grafo. Para ello, se ha utilizado la función *cluster_walktrap* del paquete *igraph*[74] aplicando el algoritmo del paseo aleatorio.

Una vez definidos los grupos y las células pertenecientes a los mismos, se procede a la evaluación del resultado para conocer cuán separados están los grupos. En primer lugar, se calcula la **pureza del agrupamiento** con la función *neighborPurity* del paquete *bluster*[75]. Para ello, se determina a qué grupo pertenecen los vecinos de cada célula. Seguidamente, se calcula la pureza como la proporción en tanto por 1 de los vecinos que pertenecen al mismo grupo de la célula evaluada respecto al total de vecinos. Cuando el resultado se representa gráficamente, se asigna un color a cada célula de acuerdo al grupo mayoritario al que pertenecen sus vecinos[63].

Otra medida para describir la separación de los grupos es la **modularidad.** Este concepto se define para parejas de grupos como la diferencia entre el número de aristas observado y esperado en base a conexiones arbitrarias. Por tanto, cuanto mayor sea el valor de la modularidad, más definido se encontrará el grupo. En el caso de los gráficos SNN, el número de aristas corresponde a la suma de los pesos de las aristas conectadas[63]. La modularidad para todas las parejas de grupos obtenidas en este trabajo se ha calculado con la función *pairwiseModularity* del paquete *bluster*[75].

Por último, se determina la **estabilidad** de los diferentes grupos. Con este objetivo, se han realizado 20 iteraciones de muestreo por reemplazamiento. Para cada uno de los nuevos *pseudodatos* generados se vuelve a construir el grafo SNN y a identificar las comunidades. A continuación, se valora si los resultados obtenidos son equivalentes al original[63]. Este proceso se ha realizado con la función *bootstrapStability* del paquete

bluster[75]. Como resultado, se obtiene un valor derivado de la puntuación ARI (por su sigla en inglés *Adjusted Rand Index*) por parejas de grupos.

4.2.7. Anotación de los tipos celulares

Tras agrupar a las células de acuerdo a sus niveles de expresión se procede a asignar el **tipo celular** al que pertenecen. El abordaje empleado en este trabajo consiste en utilizar datos procedentes de repositorios públicos en los que se conoce el tipo celular de cada muestra[64]. Estos se utilizan como referencia para compararlos con los niveles de expresión normalizados de los genes altamente variables en los estudios analizados. Con ello, las células que presenten perfiles de expresión similares a las muestras de referencia se anotarán con el correspondiente tipo celular.

Para el estudio *Multiple sclerosis* se ha utilizado la función *brainCells* del paquete *BRETIGEA*[76]. Esta función permite clasificar las células en los 6 tipos celulares principales del tejido nervioso: neuronas, astrocitos, células endoteliales, microglía, oligodendrocitos y precursores de oligodendrocitos. Concretamente, *BRETIGEA* contiene un listado de 1.000 genes asociados a cada tipo celular. Este listado fue generado por los autores mediante un consenso de los genes sobreexpresados que mejor caracterizan con sus niveles de expresión cada tipo celular. La función *brainCells* busca en la matriz de conteos una fracción de los genes del listado (proporción que puede ser modulada por el investigador) y evalúa sus niveles de expresión en cada célula. En este trabajo, tras ejecutar la función para distintos valores, se decide analizar los niveles de expresión de 60 genes. Como resultado, *brainCells* establece una puntuación por tipo celular a cada célula, asignando aquel que presenta una puntuación más elevada.

Para las cohortes 1 y 3 del estudio GSE144744, que contienen muestras de células mononucleares de sangre periférica (PBMC, por sus siglas en inglés *Peripheral Blood Mononuclear Cell*), se ha utilizado la función *SingleR* ubicada en el paquete con el mismo nombre[77]. A diferencia de *BRETIGEA*, esta función permite evaluar todo tipo de datos y no solamente aquellos procedentes de un tejido concreto. En primer lugar, se debe elegir la matriz de conteos de las células de referencia a utilizar en el análisis. Como en este trabajo no se disponía de datos propios, se han proporcionado los datos *MonacoImmuneData* del paquete *celldex*[77]. Estos constituyen la referencia estándar para analizar las células inmunitarias presentes en muestras de PBMC. En concreto, incluyen datos pertenecientes a linfocitos B, linfocitos T (tanto CD4+ como CD8+), linfocitos NK, células dendríticas y monocitos. Asimismo, contienen referencias de neutrófilos, basófilos y progenitores para identificar si alguna célula pertenece a estos tipos celulares al no haber sido eliminada durante la obtención de PBMC.

Tras la obtención de la referencia de interés, la función *SingleR* calcula la correlación de Spearman para los niveles de expresión de los genes presentes tanto en las células de referencia como en las células de los estudios a analizar. Seguidamente, determina una puntuación por tipo celular de acuerdo al resultado del análisis. De nuevo, se asigna el tipo celular que ha logrado una puntuación más elevada. En ocasiones puede ocurrir que dos o más tipos celulares presenten una puntuación elevada para una misma célula. Para

mejorar la resolución, se eliminan los tipos celulares que no han superado un determinado umbral y las puntuaciones vuelven a ser calculadas.

SingleR permite realizar el análisis de forma individual para cada célula o bien por los agrupamientos de identidad celular obtenidos en el paso anterior. En este último caso, la función suma los perfiles de expresión celulares para obtener un perfil de expresión por grupo. La matriz de conteos resultante se utiliza como dato de entrada para el análisis, por lo que a todas las células de cada grupo se les asigna el mismo tipo celular. En este trabajo se han realizado ambas ejecuciones para seleccionar la vertiente que mejor se adapte a los datos analizados.

4.2.8. Evaluación de los genes marcadores

Una vez definidas las identidades y los tipos celulares para cada estudio, se procede a identificar qué genes se encuentran sobreexpresados respecto al resto en cada categoría establecida. El objetivo principal es encontrar **combinaciones de genes específicas que permitan describir sin ambigüedad cada grupo y tipo celular**. Estos genes se denominan **genes marcadores**, ya que al detectar altos niveles de su expresión se puede "marcar" una célula con un tipo o identidad celular concreto.

La búsqueda de genes marcadores se ha realizado con la función *findMarkers* del paquete *scran*[62]. Concretamente, se realizan contrastes de expresión diferencial para los genes altamente variables considerando los tipos (o identidades) celulares por parejas aplicando el test de Wilcoxon. Este test estadístico se ha seleccionado por ser no paramétrico y, por tanto, no es necesario asumir que la distribución de los niveles de expresión es normal. Para designar un gen como marcador de un tipo (o identidad) celular, el gen debe estar sobreexpresado y presentar un p-valor ajustado significativo (FDR < 0.05) en todos los contrastes por parejas entre ese tipo celular (o identidad) y el resto. El p-valor ajustado ha sido calculado por el método de Benjamini-Hochberg (BH)[78].

Adicionalmente, los genes marcadores detectados para cada tipo celular se han buscado en la bibliografía para conocer si previamente han sido asignados al tipo celular correspondiente. La finalidad de este procedimiento consiste en aportar documentación contrastada que permita confirmar que los tipos celulares se han establecido correctamente.

4.2.9. Análisis bioestadísticos

Una vez esclarecidos los tipos celulares para cada célula y núcleo evaluado, se procede a la caracterización de las diferencias de sexo presentes en la enfermedad de EM. Para ello se realizan dos abordajes, los cuales son aplicados en cada uno de los estudios. Concretamente, el primer análisis está destinado a evaluar si se han modificado las proporciones de los tipos celulares a través de un análisis de abundancia diferencial (subapartado 4.2.9.1). Por otra parte, la segunda estrategia consiste en identificar alteraciones en los niveles de expresión génica por tipo celular (subapartado 4.2.9.2). En

ambos casos se plantean tres comparaciones atendiendo a la condición y al sexo de los individuos:

- **EM_Mujer – Control_Mujer**: conocer las diferencias existentes entre pacientes de EM e individuos sanos siendo mujer.

- **EM_Hombre – Control_Hombre**: conocer las diferencias existentes entre pacientes de EM e individuos sanos siendo hombre.

- **(EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre):** conocer las diferencias existentes en la enfermedad de EM en función del sexo del individuo, sin considerar las variaciones presentes entre hombres y mujeres en estado de salud.

4.2.9.1. Análisis de abundancia diferencial

Tal y cómo se ha citado anteriormente, el objetivo del análisis de abundancia diferencial es detectar **diferencias significativas en el número de células por tipo celular** para cada una de las comparaciones previamente establecidas. Este abordaje puede ser realizado a través de diferentes estrategias. En detalle, en este trabajo se ha optado por contrastar los resultados obtenidos mediante dos análisis diferentes; los cuales son nombrados en el manuscrito en función de su fuente bibliográfica.

El primero de ellos, denominado "método *orchestrating*"[63], se ha ejecutado haciendo uso de las funciones incluidas en el paquete *edgeR*[79]. Concretamente, consiste en la adaptación de un abordaje comúnmente utilizado para realizar análisis de expresión diferencial. En primer lugar, se cuantifica el número de células que tiene cada tipo celular en cada muestra (por ejemplo, un valor se obtendría al contabilizar el número de neuronas identificadas en la muestra procedente del paciente 1). El siguiente paso consiste en filtrar aquellos tipos celulares con un número bajo de células haciendo uso de la función *filterByExpr*. Seguidamente, se estiman las dispersiones con la función *estimateDisp*, y con *glmQLFit* se aplica un modelo general lineal asumiendo una distribución binomial negativa. Finalmente, se calculan los estadísticos de las comparaciones pertinentes con la función *glmQLFTest*.

La segunda estrategia se designa en este manuscrito como **"método *Schimer et al. 2019*"**[80]. El primer paso es idéntico al abordaje anterior: contabilizar el número de células por tipo celular y por muestra. A continuación, se calcula el número de células totales presentes en cada muestra (por ejemplo, el número total de células procedentes del paciente 1 independientemente del tipo celular). Los valores obtenidos se utilizan para calcular un factor de escalado por muestra a través de la siguiente expresión:

$$Factor\ de\ escalado = \frac{Número\ total\ de\ células\ en\ la\ muestra\ evaluada}{Número\ total\ de\ células\ en\ la\ muestra\ que\ mayor\ cantidad\ de\ células\ presenta}$$

Seguidamente, se normalizan los valores de la matriz inicial dividiéndolos por el correspondiente factor de escalado. Los términos resultantes son utilizados para buscar

diferencias de abundancia a través en las comparaciones establecidas aplicando el test U de Mann-Whitney. Este test estadístico se ha utilizado mediante la función *wilcox.test* del paquete *stats*[81].

Para ambos abordajes se ha considerado que un tipo celular presenta una abundancia diferencial significativa cuando el p-valor obtenido del contraste es menor a 0.05. Además, se ha determinado en qué sentido se detecta un aumento o reducción del número de células para cada tipo celular mediante el valor logFC. Este es calculado como el número medio de conteos por millón del primer término del contraste dividido por el correspondiente valor medio del segundo término. Finalmente, el resultado es convertido a escala logarítmica en base 2.

4.2.9.2. Análisis de expresión diferencial

El objetivo principal de este análisis consiste en identificar para cada tipo celular el número de genes que presentan un patrón de expresión alterado en la enfermedad de EM con perspectiva de sexo. En este momento es relevante recordar que se trabaja con datos transcriptómicos procedentes de las técnicas scRNA-seq y snRNA-seq. Por tanto, los datos de partida son el número de moléculas de ARN detectadas por gen y por célula o núcleo evaluado. La pregunta que se desea responder en este subapartado es ¿qué genes presentan diferencias significativas en el número de moléculas de ARN detectadas dependiendo de la condición y el sexo del individuo de procedencia? Esta cuestión es planteada para cada estudio y de forma independiente para cada tipo celular.

Existen numerosas estrategias bioinformáticas que permiten resolver la pregunta planteada. Estas pueden dividirse en dos grandes grupos. Por una parte, se encuentran los abordajes que inicialmente fueron diseñados para la técnica RNA-seq (donde se analizan conjuntos de células), y que han sido adaptados para procesar datos procedentes de célula y núcleo único. Por otra parte, en los últimos años se han generado herramientas bioinformáticas específicas para el análisis de los datos procedentes de scRNA-seq o snRNA-seq. En este trabajo se ha decidido implementar un abordaje de cada grupo, para posteriormente poder contrastar los resultados.

Atendiendo a los **análisis de RNA-seq adaptados a las estrategias de célula y núcleo único**, el proceso se ha realizado utilizando las funciones del paquete *edgeR* (*Figura 14*)[79]. El primer paso consiste en obtener una matriz de conteos donde las filas representan los genes y las columnas *pseudocélulas*; las cuales simbolizan a un tipo celular por muestra. En detalle, cada valor de la matriz se obtiene al sumar los conteos sin normalizar por gen y por muestra de las células que pertenecen al mismo tipo celular. En el ejemplo, el valor de la matriz de conteos para el gen A y la *pseudocélula* 1 se obtendría al sumar los conteos del gen A en todas las células de microglía procedentes del paciente 1. Esta aproximación permite disponer los datos de célula única en el mismo formato en el que se obtienen las lecturas de RNA-seq: (1) una columna por muestra (y tipo celular) y (2) reducción de la presencia de 0 en la matriz de conteos.

Tras generar la matriz de conteos se selecciona el tipo celular a analizar. A partir de este momento, el procedimiento expuesto se realiza para todos los tipos celulares de forma independiente. En primer lugar, se filtran los datos eliminando aquellas *pseudocélulas* que están formadas por menos de 10 células. Asimismo, se descartan los genes con baja expresión aplicando la función *filterByExprs*. A continuación, se normalizan los datos con la función *calcNormFactors.*

Los conteos normalizados se utilizan para representar gráficos de escalado multidimensional con la función *plotMDS*. Estas representaciones son equivalentes a realizar PCA, donde las coordenadas de cada *pseudocélula* se establecen en base a los dos primeros componentes principales. El objetivo consiste en identificar de forma visual si las *pseudocélulas* se separan en base a la condición y al sexo del individuo de procedencia; o bien si otro factor podría estar enmascarando la variabilidad de interés.

Seguidamente, se estima la dispersión con la función *estimateDisp.* A continuación, se emplea la función *glmQLFit* para aplicar el modelo general lineal mediante una distribución binomial negativa. En este momento, se puede bloquear el efecto de las covariables cuya variabilidad no desea ser considerada. Una vez establecido el modelo, se realizan las tres comparaciones para calcular los estadísticos pertinentes mediante la función *glmQLFTest,* corrigiendo el p-valor por el método de BH[78].

Finalmente, se establece que un gen presenta una expresión diferencial significativa cuando su p-valor ajustado es menor a 0.05. Asimismo, se determina si el gen está sobreexpresado o infraexpresado en cada condición mediante el signo del logFC; término calculado con la misma estrategia expuesta en el análisis de abundancia diferencial.

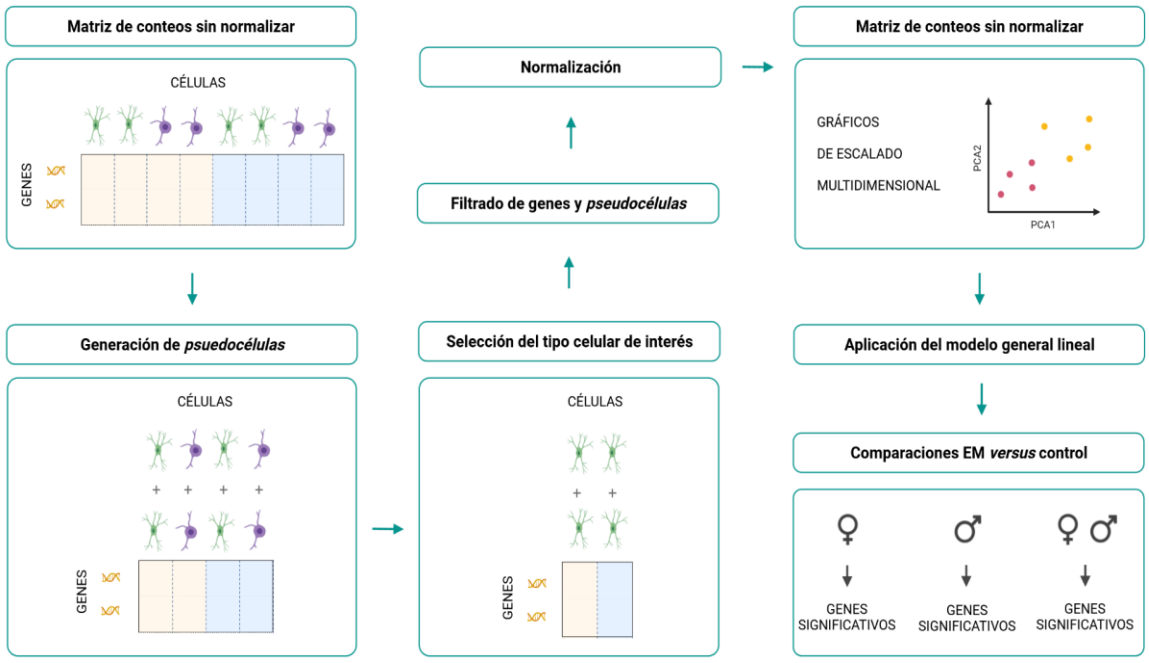


Figura 14. Flujo del análisis de expresión diferencial mediante la estrategia adaptada de los abordajes originales para RNA-seq. *Partiendo de la matriz de conteos con los*

datos de expresión génica sin normalizar, las columnas son combinadas para la obtención de pseudocélulas (sumatorio de conteos por gen, tipo celular y muestra). Los datos generados por tipo celular son analizados a través de sucesivos pasos para realizar los contrastes de expresión diferencial de interés: (♀) EM_Mujer – Control_Mujer, (♂) EM_Hombre – Control_Hombre, y (♀♂) (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). Célula verde: célula de microglía; célula morada: precursor de oligodendrocito; columnas naranjas: datos procedentes del individuo 1; columnas azules: datos procedentes del individuo 2. EM: esclerosis múltiple. Figura creada en BioRender.com.

Por otra parte, la **estrategia específica para los datos de scRNA-seq y snRNA-seq** se ha elaborado mediante las funciones del paquete MAST (*Figura 15*)[82]. En este caso se trabaja directamente con las células de forma individual; ya que el abordaje está preparado para considerar sus características concretas. De nuevo, el proceso se realiza para cada tipo celular de forma independiente.

En primer lugar, se elabora un análisis exploratorio para conocer qué porcentaje de variabilidad de los 10 primeros componentes principales puede explicarse mediante cada covariable. Estos valores son calculados con la función *getExplanatoryPCs* del paquete *scater*[61], y representados gráficamente a través de diagramas de densidad con la función *plotExplanatoryPCs* del mismo paquete. La evaluación de las gráficas obtenidas permite identificar qué fuentes de variabilidad pueden enmascarar las diferencias existentes entre la condición y el sexo de los individuos.

Seguidamente, se escala el número de genes por célula para los que se detectan moléculas de ARN, ya que esta variable tiene que proporcionarse al modelo. Concretamente, se aplica un modelo lineal generalizado a través de la función *zlm* del paquete MAST tras ser solventado un error en su configuración (https://github.com/RGLab/MAST/issues/158). La función *zlm* aplica un modelo de vallas en el que se combina una distribución discreta y una distribución continua. Posteriormente, se ajusta el resultado del modelo a través del método empírico de Bayes. En detalle, la distribución discreta corresponde con un modelo regresión logística para modelizar si un gen se expresa o no. Por su parte, en la distribución continua se aplica un modelo lineal gaussiano condicionado a que el gen se exprese.

Con todo ello, para la función *zlm* se utiliza como datos de entrada la matriz de conteos sin normalizar en escala logarítmica. Asimismo, se proporciona la variable conjunta de condición y sexo, el número escalado de genes, y todas aquellas variables control que se hayan identificado en el análisis exploratorio para descartar el efecto de su variabilidad. Tras obtener el resultado del modelo, se determinan los estadísticos de las tres comparaciones con la función *lrTest.*

De la misma forma que en la estrategia anterior, los p-valores son corregidos por el método de BH[78]. Asimismo, conocer en qué sentido se encuentra el gen sobreexpresado se determina calculando el valor logFC. La significatividad de los resultados se establece en p-valores ajustados inferiores a 0.05.

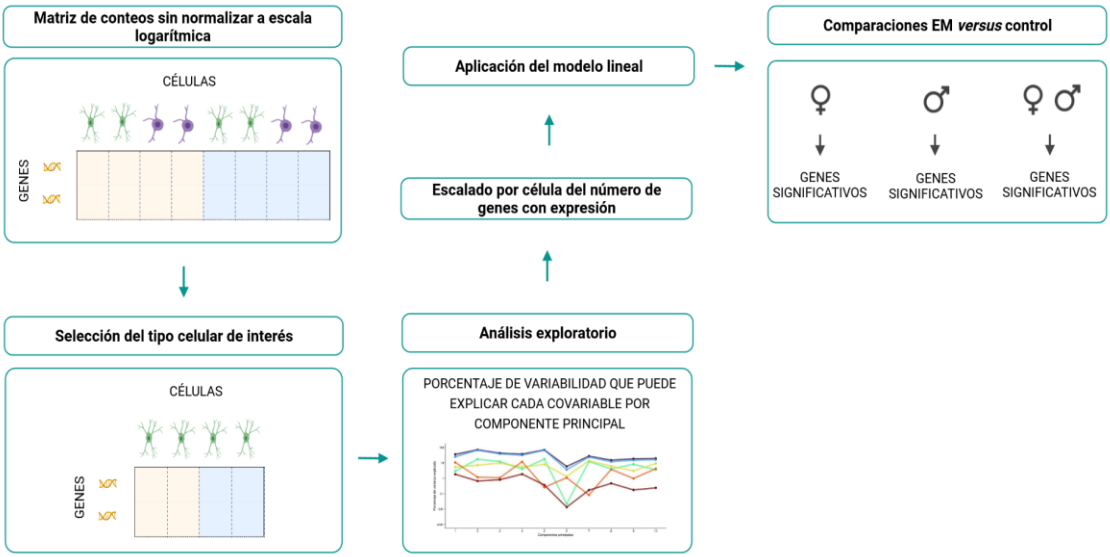


Figura 15. Flujo del análisis de expresión diferencial mediante la estrategia específica para datos de scRNA-seq y snRNA-seq. *Realización del procedimiento por cada tipo celular evaluado; para finalmente poder desarrollar los contrastes de expresión diferencial de interés: (♀) EM_Mujer – Control_Mujer, (♂) EM_Hombre – Control_Hombre, y (♀♂) (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). Célula verde: célula de microglía; célula morada: precursor de oligodendrocito; columnas naranjas: datos procedentes del individuo 1; columnas azules: datos procedentes del individuo 2. EM: esclerosis múltiple. Figura creada en BioRender.com*

4.3. Integración de los resultados del análisis primario

Tras el análisis individual de cada uno de los estudios, el siguiente paso es encontrar qué tienen en común los resultados obtenidos. En este punto, es importante recordar la procedencia de los datos que van a ser utilizados.

Concretamente, en el análisis de expresión diferencial (apartado 4.2.9.2.) se han realizado tres comparaciones por tipo celular: EM_Mujer – Control_Mujer; EM_Hombre – Control_Hombre; y (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). Del resultado obtenido en cada contraste se ha generado un listado de genes, los cuales presentan un patrón de expresión estadísticamente significativo entre las condiciones evaluadas.

A su vez, este listado puede dividirse en dos atendiendo al signo del valor logFC. Con ello, se dispone de dos listas por estudio, tipo celular y comparación: la primera con los genes sobreexpresados en la primera condición del contraste (o infraexpresados en la segunda condición); y la segunda con la situación inversa. De este modo, los análisis se pueden plantear de forma independiente para cada una de las dos situaciones.

Los conjuntos de genes resultantes son utilizados como datos de entrada para llevar a cabo la integración de los resultados. Con este objetivo, se plantean dos abordajes:

1. Caracterizar los **genes significativos comunes para todos los tipos celulares dentro de cada estudio.** Con este análisis se pretende conocer qué genes podrían mejorar la descripción de la enfermedad sin considerar el tipo celular.

2. Caracterizar los **genes significativos comunes por tipo celular entre las cohortes 1 y 3 del estudio GSE144744.** Los resultados individuales de los dos estudios pueden integrarse al disponer de los mismos tipos celulares, logrando así encontrar información consenso para las células del sistema inmunitario de sangre periférica.

Ambos abordajes se realizan con la misma metodología modificando los datos de entrada. Para el primer caso, en cada análisis se utilizan los datos de todos los tipos celulares procedentes del mismo estudio. Por otra parte, la segunda estrategia se aplica para cada tipo celular con los datos de los dos estudios que contienen muestras de sangre periférica.

El primer paso de ambos abordajes es calcular las correspondientes intersecciones, representando los resultados de forma gráfica mediante **diagramas de Venn.** Para ello, se ha utilizado la función *venn* del paquete con el mismo nombre[83]. De los resultados obtenidos se extraen los genes comunes para todos los tipos celulares (primer abordaje) o por tipo celular entre los dos estudios (segundo abordaje).

A continuación, se determina si los genes presentes en las intersecciones han sido descritos previamente para la enfermedad de EM. Con este objetivo, se descargan de la base de datos ***Open Targets Platform***[84] los genes asociados a la patología bajo el término "*multiple sclerosis".* El listado obtenido se compara con los resultados de este trabajo, para conocer qué genes han sido asociados a la enfermedad con anterioridad.

Asimismo, los genes obtenidos de las intersecciones se caracterizan a nivel funcional mediante un **análisis de sobrerrepresentación.** Este procedimiento se basa en comparar las funciones biológicas anotadas para un grupo de genes de interés (en este caso, los genes procedentes de las intersecciones) respecto al total de genes evaluados en el análisis de expresión diferencial. En detalle, se determina para cada función si la proporción de genes asociados es significativamente mayor en el grupo de genes de interés respecto al grupo de referencia[85].

Este procedimiento, conocido como test de Fisher (o test hipergeométrico), se ha implementado con la función *enrichGO* del paquete *clusterProfiler*[86]*.* Concretamente, se han evaluado como funciones los procesos biológicos descritos en el repositorio GO (por sus siglas en inglés *Gene Ontology*) obtenidos del paquete *org.Hs.eg.db*[87]. Para calcular el p-valor ajustado se ha utilizado el método de BH[78], considerando significativas aquellas funciones con valores inferiores a 0.05.

5. RESULTADOS

Este apartado está destinado a la exposición de los resultados obtenidos tras la aplicación de las metodologías y tratamientos estadísticos desarrollados en el apartado anterior.

5.1. Revisión sistemática

5.1.1. Flujo de trabajo atendiendo a las directrices PRISMA

La revisión sistemática se realizó durante el mes de febrero de 2021. En este periodo de tiempo se identificaron 33 estudios en las diferentes bases de datos y a través del buscador Google, de los cuales dos de ellos se encontraban repetidos. Tras realizar su revisión individual, 14 fueron descartados por no utilizar las metodologías scRNA-seq o snRNA-seq, 4 por estar enfocados en la evaluación de otras enfermedades neurológicas presentando a los pacientes de EM como controles, y 6 por no disponer de suficiente tamaño muestral (*Figura 16*).

Con todo ello, 7 estudios cumplían los requisitos de elegibilidad establecidos para poder realizar el análisis. Tras una evaluación más profunda, se eliminaron 3 estudios por presentar como controles a individuos con otras enfermedades neurológicas (pseudo-controles), 1 por no cumplir con el tamaño muestral requerido tras recibir los datos de los autores, y 1 por no lograr el acceso a los datos. Por tanto, 2 estudios fueron finalmente incluidos en el análisis (*Figura 16*).

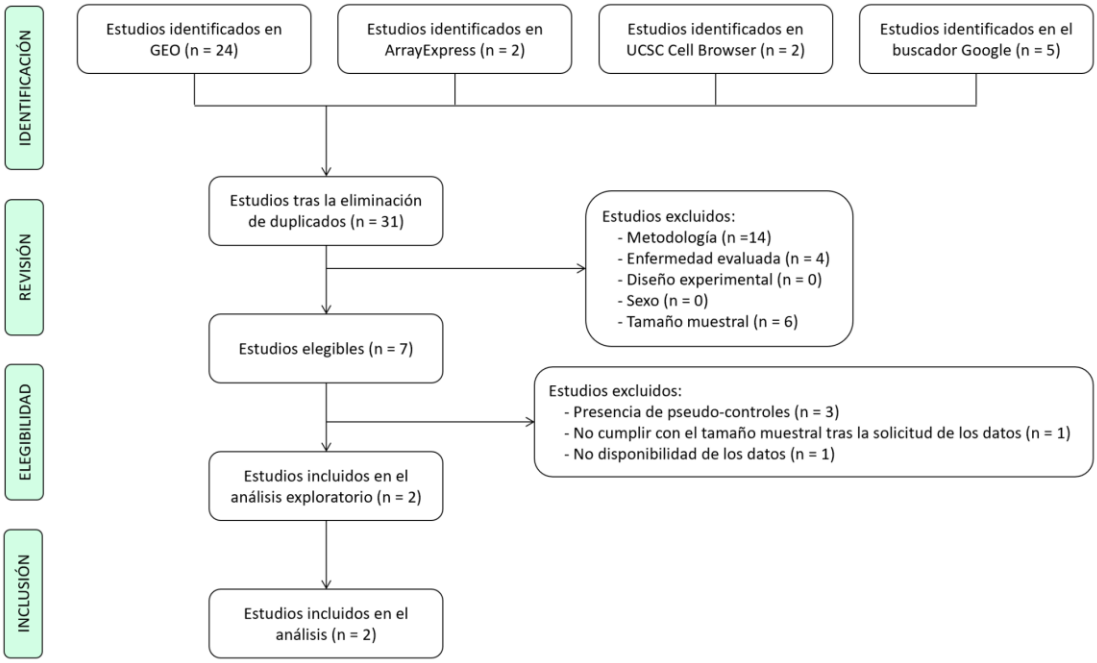


Figura 16. Fases del proceso de la revisión sistemática acordes a la declaración PRISMA[55]. *Especificación del número de estudios (n) mantenidos en las fases de*

identificación, revisión, elegibilidad e inclusión; así como los correspondientes motivos de exclusión.

5.1.2. Descripción de los estudios seleccionados

Los datos de los estudios que participaron en el análisis fueron (1) el designado como *Multiple sclerosis* en el repositorio UCSC Cell Browser y (2) el registrado con el identificador GSE144744 en la base de datos GEO.

Los datos procedentes de ***Multiple sclerosis*** **en la base de datos UCSC Cell Browser** fueron identificados en primera instancia en *Google Scholar* a través de su publicación asociada[80]. En ella, se referencia solamente el identificador para acceder a las lecturas obtenidas tras el proceso de secuenciación. Por ello, se decidió contactar con los investigadores responsables para solicitar la matriz de conteos, la cual fue proporcionada, junto con los ficheros adicionales necesarios, a través de la base de datos UCSC Cell Browser.

En este trabajo se procesaron muestras biológicas *postmortem* de tejido nervioso encefálico mediante el abordaje snRNA-seq. En conjunto, se evaluaron 21 muestras procedentes de 10 pacientes de EM progresiva (1 individuo EMPP y 9 individuos EMSP) y 9 sujetos controles. El aislamiento de los núcleos y la preparación de las moléculas de ARN para el proceso de secuenciación se realizó con la metodología comercial *10x Genomics,* y como secuenciador se utilizó Illumina HiSeq 2500[80].

Por su parte, los datos depositados bajo el identificador **GSE144744 en el repositorio GEO** incluyen muestras procesadas de 3 cohortes independientes (*Tabla 1*). Todas corresponden con muestras de PBMC de pacientes de EM y de sujetos controles procesadas mediante el abordaje scRNA-seq. Esta metodología se desarrolló con la tecnología comercial *10x Genomics* y las lecturas se obtuvieron con el secuenciador Illumina Nextseq500[88].

La primera cohorte incluye muestras de pacientes de EMRR en un estudio longitudinal en el que se proporciona el fármaco *Natalizumab,* la segunda también evalúa a pacientes de EMRR mientras que la tercera se centra en el subtipo EMPP[88] *(Tabla 1).* Atendiendo a los criterios de inclusión y exclusión establecidos para este trabajo, se seleccionó la cohorte 1 descartando las muestras de pacientes tratados. Asimismo, la cohorte 2 fue eliminada por contener solamente muestras de mujeres. Por último, se seleccionó la cohorte 3 por cumplir con todos los requerimientos.

En conjunto, las características de los **3 estudios** que finalmente fueron utilizados en el trabajo se resumen en la tabla 2.

Cohorte	Tipo de pacientes	Tipo de controles	Muestras (caso-mujer: caso-hombre : control-mujer : control-hombre)
1	EMRR	individuos sanos	20 (10 -nat, 10 +nat) : 18 (8 –nat, 10 +nat) : 10 : 10
2	EMRR	individuos sanos	0 : 13 : 0 : 13
3	EMPP	individuos sanos	42 : 18 : 42 :18

Tabla 1. Características de las cohortes presentes en el estudio GSE144744. *Se indica la cohorte (columna 1), el subtipo de la enfermedad de esclerosis múltiple sufrida por los pacientes (columna 2), el tipo de sujetos controles utilizado (columna 3) y el grupo de muestras procesadas en base a la condición y sexo de los individuos (columna 4). EMRR: esclerosis múltiple remitente remanente; EMPP: esclerosis múltiple primaria progresiva; -nat: muestra previa al tratamiento con Natalizumab; +nat: muestra posterior al tratamiento con Natalizumab.*

Estudio	Metodología	Tipo de muestra	Casos	Controles
Multiple sclerosis	snRNA-seq	Tejido nervioso *postmortem*	10 pacientes de EM progresiva (1 EMPP y 9 EMSP)	9 individuos sanos
GSE144744 cohorte 1	scRNA-seq	PBMC	5 pacientes de EMRR	5 individuos sanos
GSE144744 cohorte 3	scRNA-seq	PBMC	10 pacientes de EMPP	10 individuos sanos

Estudio	Muestras (caso-mujer: caso-hombre : control-mujer : control-hombre)	Número total de células	Número total de genes
Multiple sclerosis	4 : 8 : 5 : 4	48.919	65.217
GSE144744 cohorte 1	10 : 8 : 10 : 10	71.592	15.354
GSE144744 cohorte 3	42 : 18 : 42 : 18	265.342	15.354

Tabla 2. Descripción de los estudios considerados en el trabajo. *Metodología utilizada (columna 2), procedencia de la muestra biológica (columna 3), número de pacientes de esclrosis múltiple (columna 4) y controles (columna 5) considerados, número de muestras*

desglosadas por la condición y sexo del individuo (columna 7) y número de células (columna 8) y genes (columna 9) de partida en la matriz de conteo para cada estudio. scRNA-seq: single cell RNA sequencing; snRNA-seq: single nucleus RNA sequencing; PBMC: peripheral blood mononuclear cell; EMRR: esclerosis múltiple remitente remanente; EMPP: esclerosis múltiple primaria progresiva; EMSP: esclerosis múltiple secundaria progresiva; EM: esclerosis múltiple.

5.2. Análisis primario de los estudios

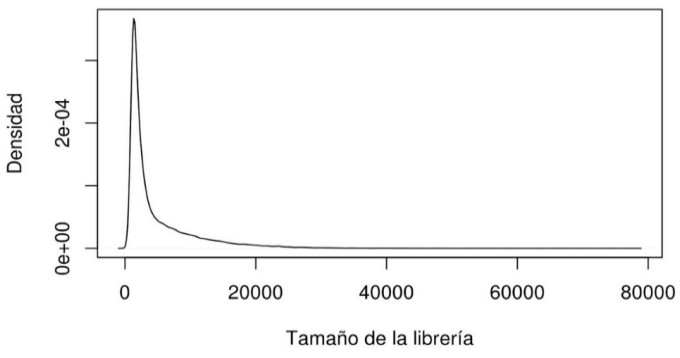
5.2.1. Control de calidad

Tal y como se ha indicado en el apartado de materiales y métodos, la evaluación del control de calidad se adaptó individualmente a las peculiaridades de cada estudio. En este subapartado se incorporan las representaciones gráficas más significativas, mientras que las restantes pueden ser consultadas en el anexo II (*Figuras II.1 - II.21, https://doi.org/10.5281/zenodo.5068587*).

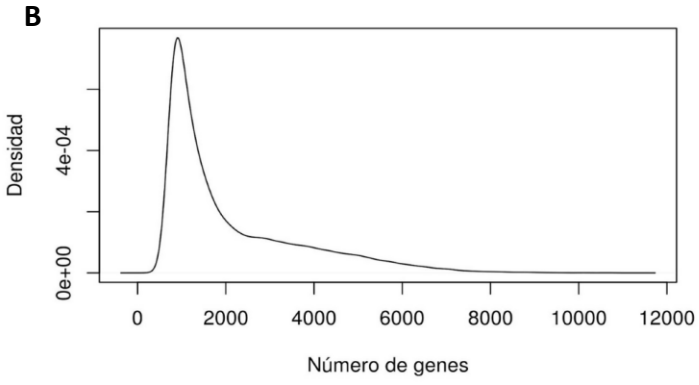
5.2.1.1. Control de calidad para el estudio *Multiple sclerosis*

Tras observar las representaciones gráficas de los indicadores antes del filtrado *(Figura 17)* se establecieron los umbrales para cada uno de ellos. Se decidió eliminar aquellas células con tamaño de la librería inferior a 1000 y/o número de genes expresados menor a 500, tal y como realizaron los autores del estudio[80]. Para los porcentajes de conteos mitocondriales y ribosomales se establecieron puntos de corte adaptativos, descartando las células consideradas atípicas tras la ejecución de la función *isOutlier.* Esta actuación, más restrictiva que establecer un umbral fijo, fue realizada por tratarse de un experimento de snRNA-seq (en una situación ideal el núcleo se habría separado completamente del citoplasma; por lo que no se detectaría la expresión de genes mitocondriales). Por otra parte, la eliminación de agregados celulares se realizó descartando los dobletes detectados por la función *scDblFinder*. Adicionalmente, se descartaron los genes expresados en menos de 3 células (punto de corte indicado en el estudio[80]) y que no fuesen genes nucleares codificantes de proteínas, *microRNAs* o *lncRNAs*.

A



Resumen	
Mínimo	79
Primer cuartil	1418
Mediana	2381
Media	4850
Tercer cuartil	6056
Máximo	77853



Resumen	
Mínimo	68
Primer cuartil	959
Mediana	1434
Media	2120
Tercer cuartil	2882
Máximo	11286

Figura 17. Diagramas de densidad para el tamaño de las librerías (A) y el número de genes (B) del estudio Multiple sclerosis antes del filtrado. *Representación gráfica de la distribución del indicador evaluado (derecha) y los valores de sus medidas estadísticas descriptivas: valores mínimo, primer cuartil, mediana, media, tercer cuartil y máximo (izquierda).*

Con la inspección detallada de las representaciones gráficas, tanto antes como después del filtrado, se determinó que ninguna muestra obtenida de los participantes presentaba un comportamiento excesivamente divergente cómo para ser eliminada (*Figuras 18 y 19*).

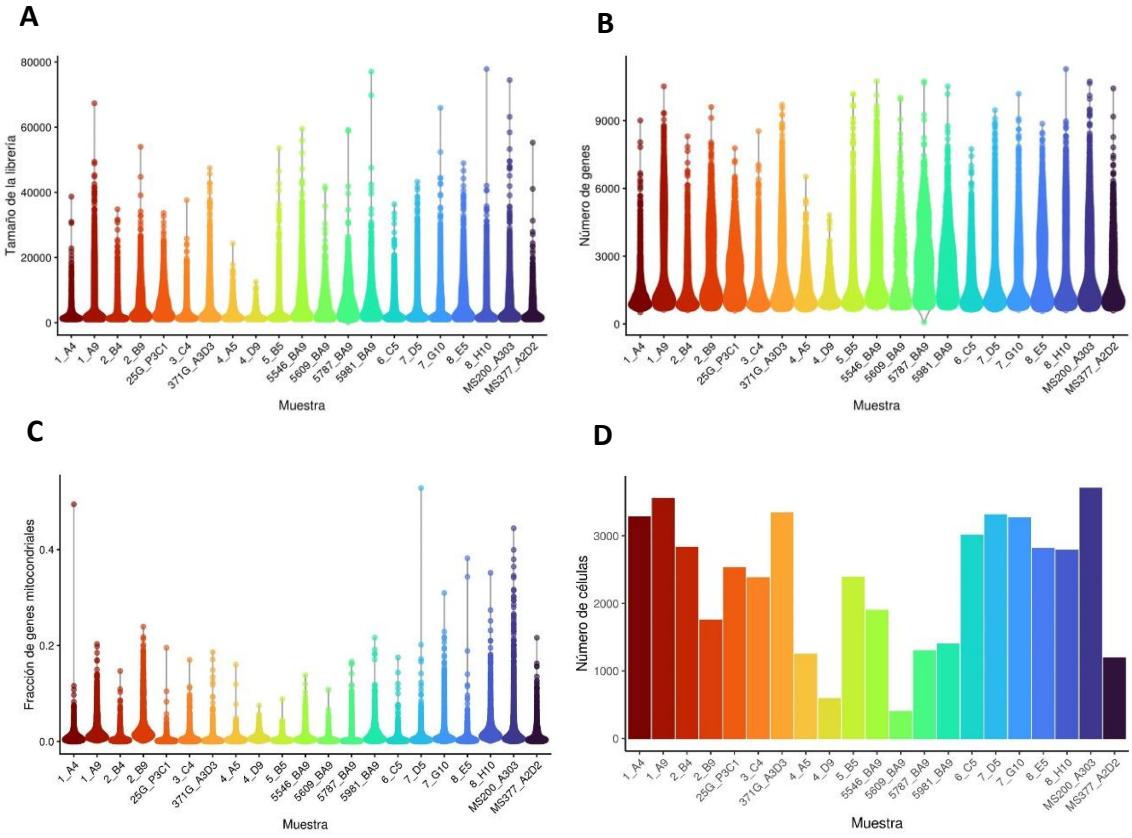


Figura 18. Descripción gráfica de la distribución de las células por muestra del estudio Multiple sclerosis antes del filtrado en base al tamaño de la librería (A), el número de genes (B), y la proporción de conteos mitocondriales (C); así como del número de

células evaluadas (D). *Representación de las células (puntos) de acuerdo con la muestra de procedencia (eje X) en función de los valores de los indicadores calculados (eje Y): tamaño de la librería (A), número de genes expresados (B) y fracción de conteos mitocondriales (C). Distribución del número de células obtenidas por muestra en formato de diagrama de barras (D). Cada punto representa una célula en los gráficos (A), (B) y (C).*

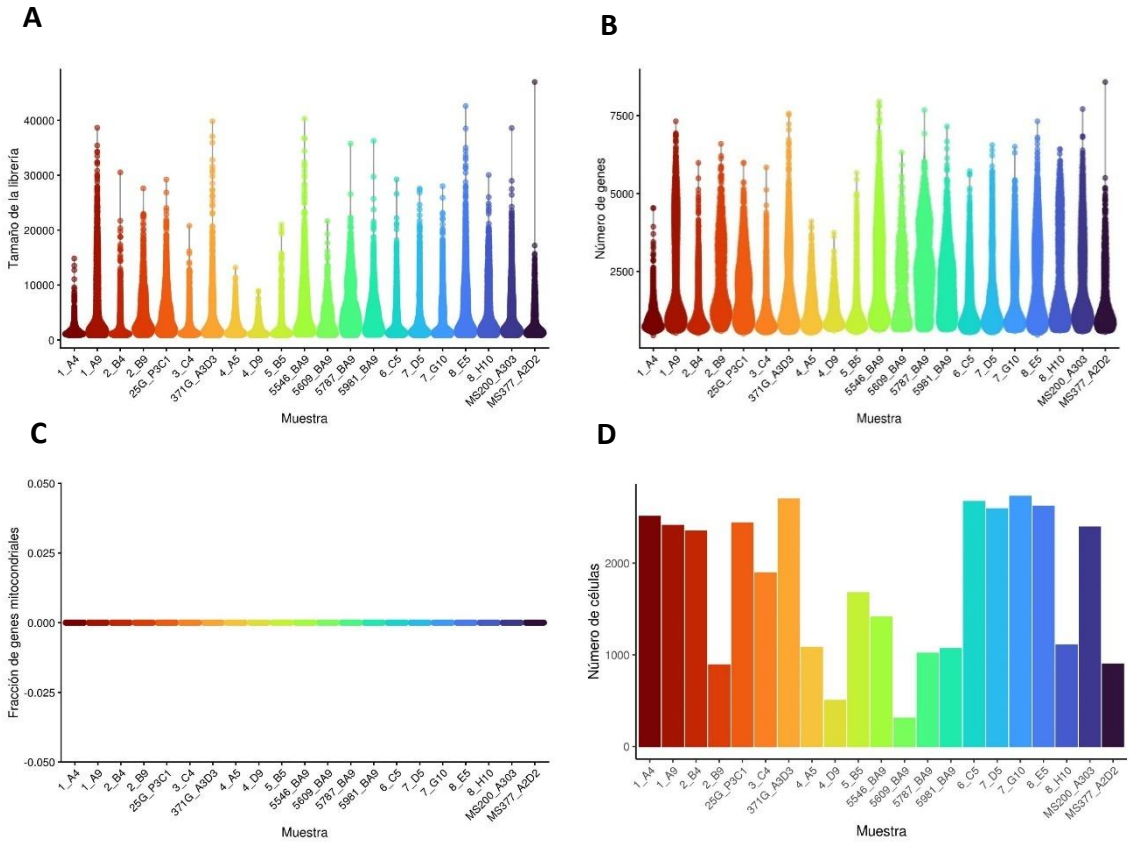


Figura 19. Descripción gráfica de la distribución de las células por muestra del estudio Multiple sclerosis después del filtrado en base al tamaño de la librería (A), el número de genes (B), y la proporción de conteos mitocondriales (C); así como del número de células evaluadas (D). *Representación de las células (puntos) de acuerdo a la muestra de procedencia (eje X) en función de los valores de los indicadores calculados (eje Y): tamaño de la librería (A), número de genes expresados (B) y fracción de conteos mitocondriales (C). Distribución del número de células obtenidas por muestra en formato de diagrama de barras (D). Cada punto representa una célula en los gráficos (A), (B) y (C).*

Con todo ello, antes del filtrado se disponía de 48.919 células y 65.216 genes. Tras descartar las células y los genes que no cumplían con los criterios de calidad para alguno de los indicadores se mantienen en el análisis **37.116 células y 21.842 genes**. La distribución de células tras el filtrado en base a la condición y el sexo de la muestra de procedencia puede visualizarse en la figura 20.

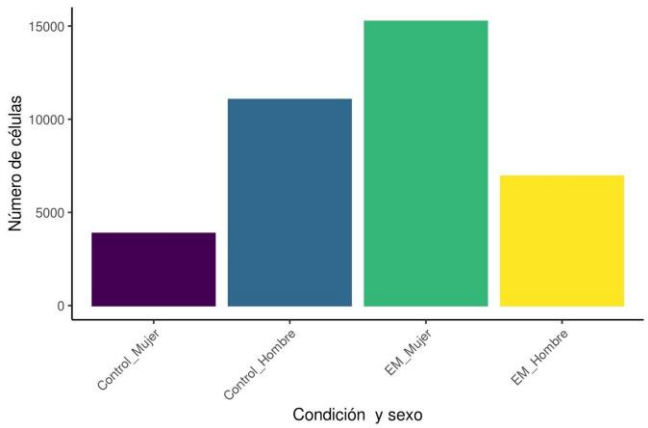
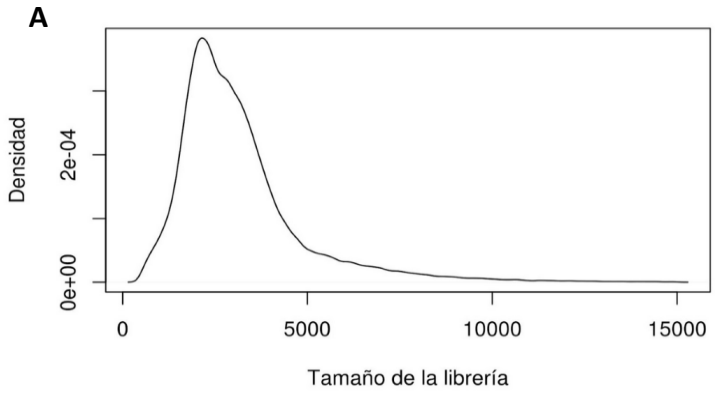


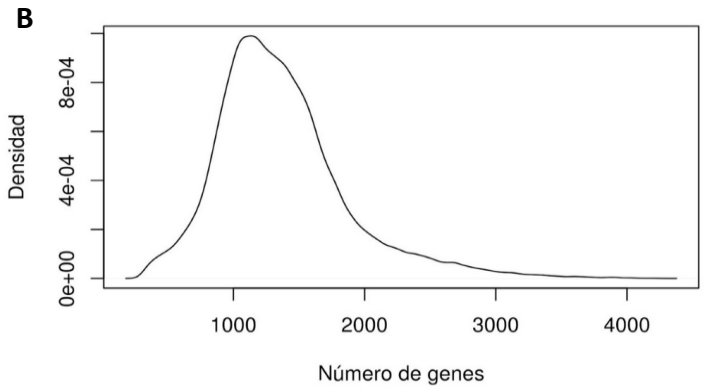
Figura 20. Número de células mantenidas en el estudio Multiple sclerosis tras el control de calidad. *Distribución segregada por la condición y el sexo del total de células que permanecen en el análisis al cumplir los requerimientos de calidad establecidos. EM: esclerosis múltiple.*

5.2.1.2. Control de calidad para las dos cohortes del estudio GSE144744

Al representar gráficamente los indicadores, se observó en ambas cohortes que los datos descargados habían sido filtrados con los umbrales establecidos por los autores (*Figuras 21 y 22 para las cohortes 1 y 3, respectivamente*). Por ello, se decidió no aplicar ningún punto de corte adicional. Concretamente, los autores eliminaron las células que contaban con un tamaño de la librería inferior a 500 o superior a 15.000 conteos. Asimismo, mantuvieron las células que expresaban entre 300 y 5.000 genes y aquellas con un porcentaje de conteos mitocondriales inferior al 20%. La identificación de agregados fue realizada con el paquete *doubletFinder*[89], descartando las células asignadas como dobletes. Atendiendo a los genes, los autores eliminaron aquellos expresados en menos de 10 células y todos los que no fueran codificantes de proteínas a excepción de *microRNAs y lncRNAs*[88].

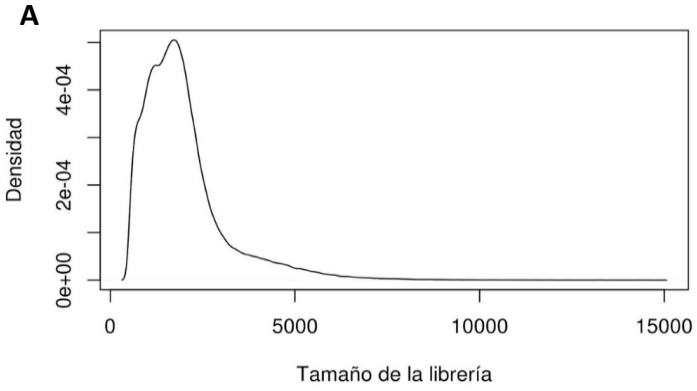


Resumen	
Mínimo	501
Primer cuartil	2066
Mediana	2771
Media	3170
Tercer cuartil	3698
Máximo	14941

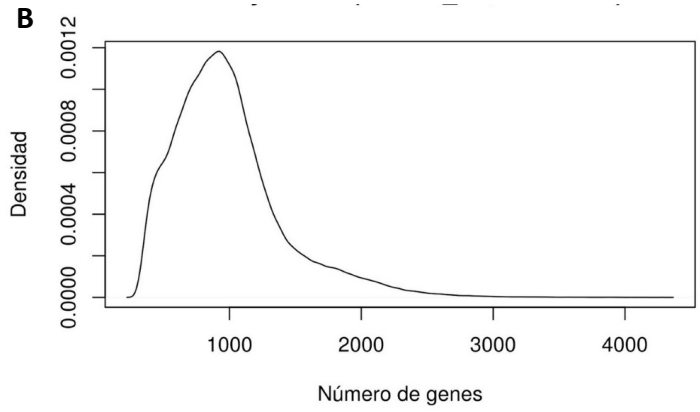


Resumen	
Mínimo	301
Primer cuartil	1038
Mediana	1295
Media	1373
Tercer cuartil	1602
Máximo	4257

Figura 21. Diagramas de densidad para el tamaño de las librerías (A) y el número de genes (B) del estudio GSE144744 cohorte 1. *Representación gráfica de la distribución del indicador evaluado (derecha) y los valores de sus medidas estadísticas descriptivas: valores mínimo, primer cuartil, mediana, media, tercer cuartil y máximo (izquierda).*



Resumen	
Mínimo	501
Primer cuartil	1207
Mediana	1734
Media	1977
Tercer cuartil	2319
Máximo	14879



Resumen	
Mínimo	301
Primer cuartil	699
Mediana	924
Media	989.3
Tercer cuartil	1170
Máximo	4288

Figura 22. Diagramas de densidad para el tamaño de las librerías (A) y el número de genes (B) del estudio GSE144744 cohorte 3. *Representación gráfica de la distribución del indicador evaluado (derecha) y los valores de sus medidas estadísticas descriptivas: valores mínimo, primer cuartil, mediana, media, tercer cuartil y máximo (izquierda).*

A su vez, a través del análisis exploratorio no se identificó ningún comportamiento atípico, por lo que se mantuvieron las células de todas las muestras procesadas (*Figuras 23 y 24 para las cohortes 1 y 3, respectivamente*).

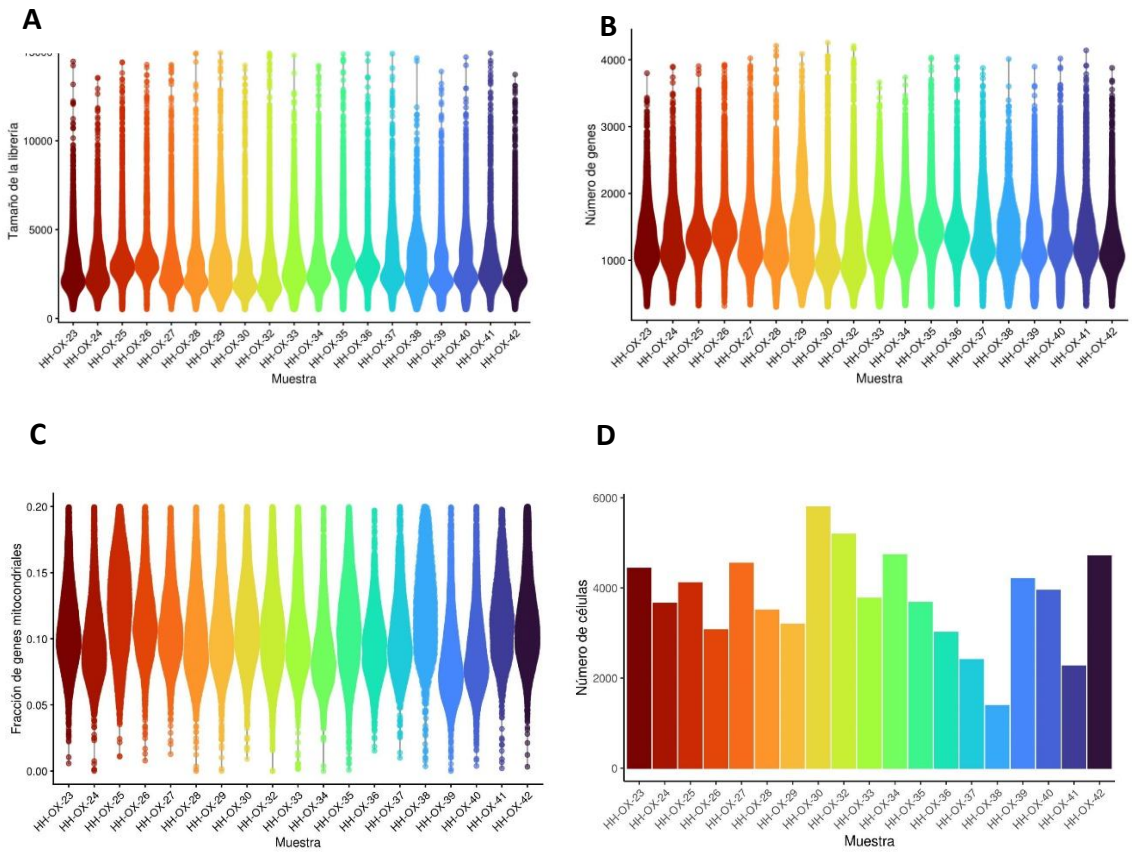


Figura 23. Descripción gráfica de la distribución de las células por muestra del estudio GSE144744 cohorte 1 en base al tamaño de la librería (A), el número de genes (B), y la proporción de conteos mitocondriales (C); así como del número de células evaluadas (D). *Representación de las células (puntos) de acuerdo a la muestra de procedencia (eje X) en función de los valores de los indicadores calculados (eje Y): tamaño de la librería (A), número de genes expresados (B) y fracción de conteos mitocondriales (C). Distribución del número de células obtenidas por muestra en formato de diagrama de barras (D). Cada punto representa una célula en los gráficos (A), (B) y (C).*

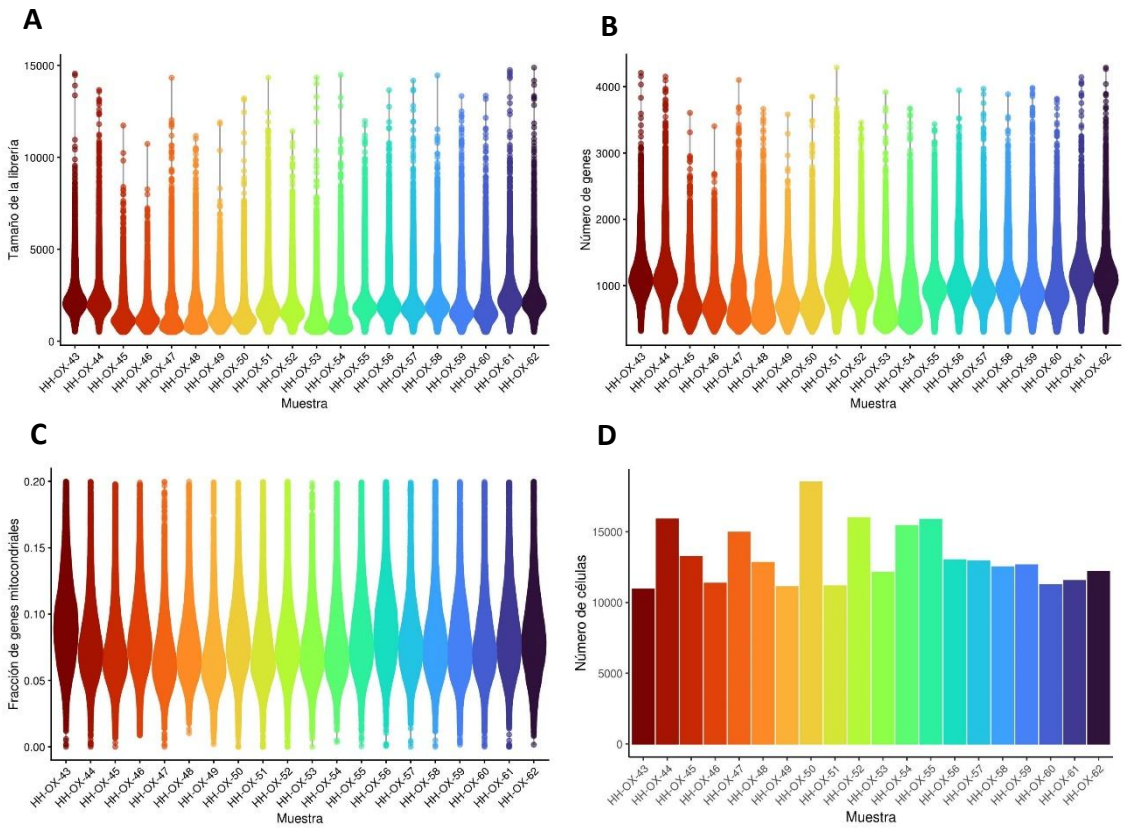


Figura 24. Descripción gráfica de la distribución de las células por muestra del estudio GSE144744 cohorte 3 en base al tamaño de la librería (A), el número de genes (B), y la proporción de conteos mitocondriales (C); así como del número de células evaluadas (D). *Representación de las células (puntos) de acuerdo a la muestra de procedencia (eje X) en función de los valores de los indicadores calculados (eje Y): tamaño de la librería (A), número de genes expresados (B) y fracción de conteos mitocondriales (C). Distribución del número de células obtenidas por muestra en formato de diagrama de barras (D). Cada punto representa una célula en los gráficos (A), (B) y (C).*

Por lo tanto, para ambas cohortes con muestras de sangre periférica se mantiene la totalidad de células y genes de los datos de partida descargados. Concretamente, **la cohorte 1 consta de 71.592 células y 15.354 genes** mientras que **la cohorte 3 dispone de 265.342 células y 15.354 genes**. La distribución del número de células de acuerdo a la condición y el sexo del individuo de procedencia puede visualizarse en la figura 25.

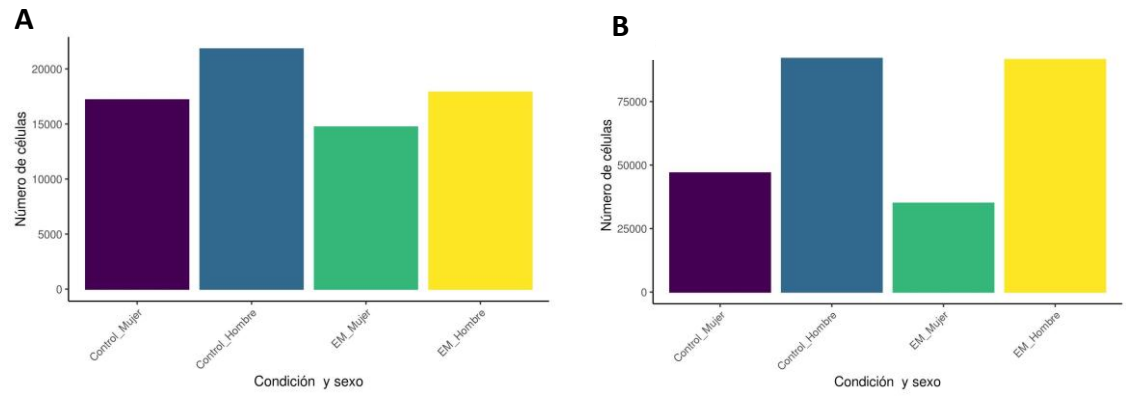


Figura 25. Número de células mantenidas en las cohortes 1 (A) y 3 (B) del estudio GSE144744 tras el control de calidad. *Distribución segregada por la condición y el sexo del total de células que permanecen en el análisis al cumplir los requerimientos de calidad establecidos. EM: esclerosis múltiple.*

5.2.2. Normalización

Para los tres estudios se calcularon los factores de escalado para cada célula por los métodos de tamaño de la librería y de deconvolución. De forma unánime se observa en la primera estrategia que el factor de escalado es completamente dependiente del número total de conteos. Sin embargo, el método de deconvolución permite calcular factores de escalado diferentes a partir de tamaños de librería muy similares. De este modo, se logra contemplar la heterogeneidad de los niveles de expresión característica de los abordajes de scRNA-seq y snRNA-seq (*Figura 26*). Los gráficos para las cohortes 1 y 3 del estudio GSE144744, así como la distribución de los factores de escalado por deconvolución en los 3 análisis puede consultarse en el anexo II (*Figuras II.22 - II.24, https://doi.org/10.5281/zenodo.5068587*).

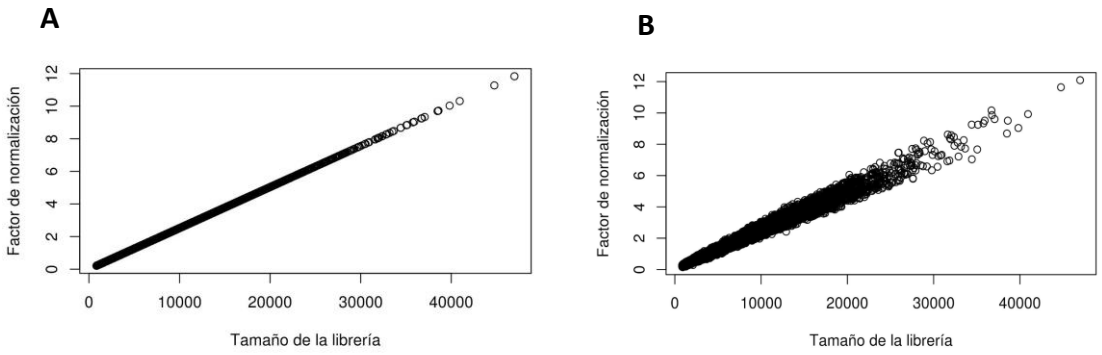


Figura 26. Factor de escalado por el método de tamaño de la librería (A) y de deconvolución (B) respecto del tamaño de la librería por célula del estudio Multiple sclerosis.

Ambos métodos no son totalmente independientes, ya que las dos metodologías están basadas en el total de conteos por célula. Con ello, al representar gráficamente los dos factores de escalado en un diagrama de puntos, se observa que entre ellos existe una correlación positiva *(Figura 27, resultados para las cohortes 1 y 3 de GSE144744 en la figura II.25 del anexo II; https://doi.org/10.5281/zenodo.5068587).*

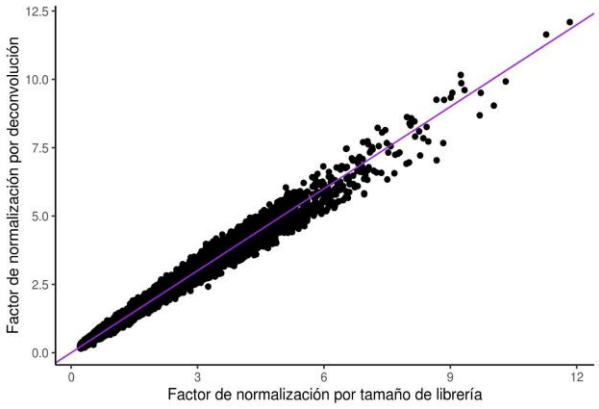
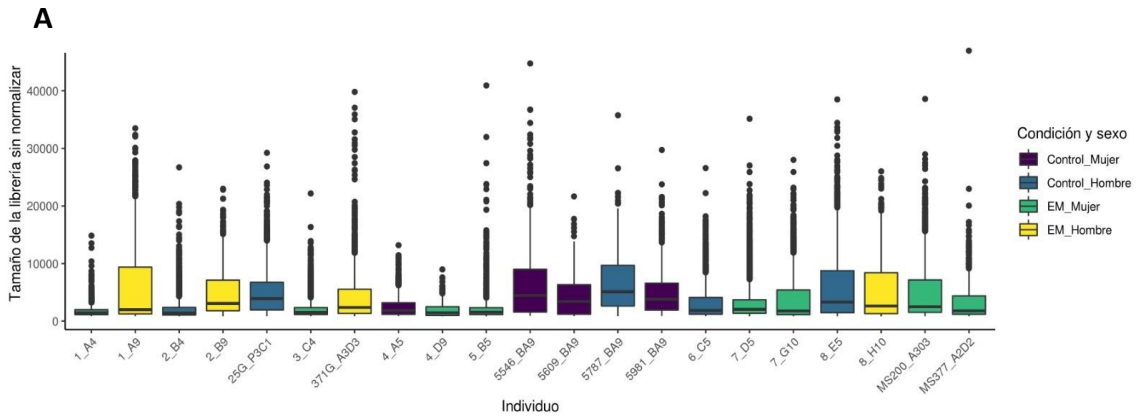


Figura 27. Comparación de los factores de escalado obtenidos en el estudio Multiple sclerosis. *Representación en un diagrama de puntos del factor de escalado por deconvolución (eje Y) frente al calculado por el tamaño de la librería (eje X) para cada célula. Ecuación de la línea continua morada: x = y.*

Tras calcular el factor de escalado por deconvolución se normalizaron las matrices de conteos y se transformaron a escala logarítmica. Como resultado, se puede observar el cambio en la distribución de los tamaños de la librería por muestra para cada estudio (*Figuras 28, 29 y 30*). Con todo ello, se ha logrado que los niveles de expresión sean comparables entre células, pudiendo evaluar su variabilidad biológica en los pasos posteriores.



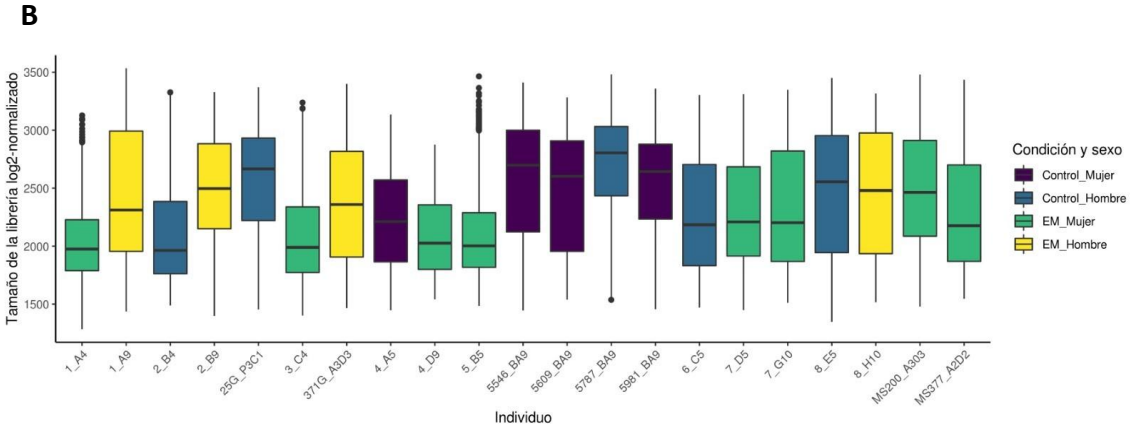


Figura 28. Distribución de los tamaños de las librerías sin normalizar (A) y normalizados con escala logarítmica (B) de las células por muestra para el estudio Multiple sclerosis. *Diagramas de cajas que referencian gráficamente el primer cuartil (base inferior de la caja), la mediana (franja que divide la caja) y el tercer cuartil (base superior de la caja) de la distribución de los datos. Representación de valores atípicos (inferiores o superiores a 1,5 veces el rango intercuartílico) en formato de puntos. Colores establecidos en base a la condición y al sexo del individuo: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

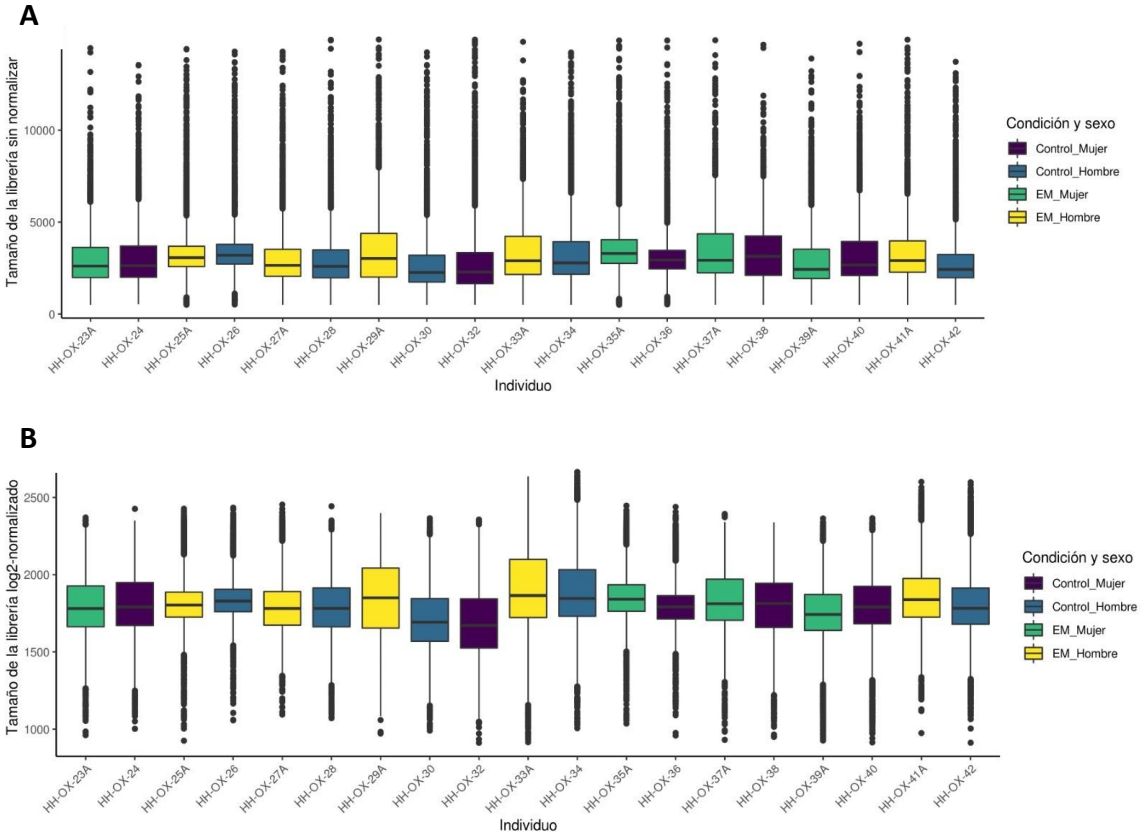


Figura 29. Distribución de los tamaños de las librerías sin normalizar (A) y normalizados con escala logarítmica (B) de las células por muestra para el estudio

GSE144744 cohorte 1. *Diagramas de cajas que referencian gráficamente el primer cuartil (base inferior de la caja), la mediana (franja que divide la caja) y el tercer cuartil (base superior de la caja) de la distribución de los datos. Representación de valores atípicos (inferiores o superiores a 1,5 veces el rango intercuartílico) en formato de puntos. Colores establecidos en base a la condición y al sexo del individuo: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

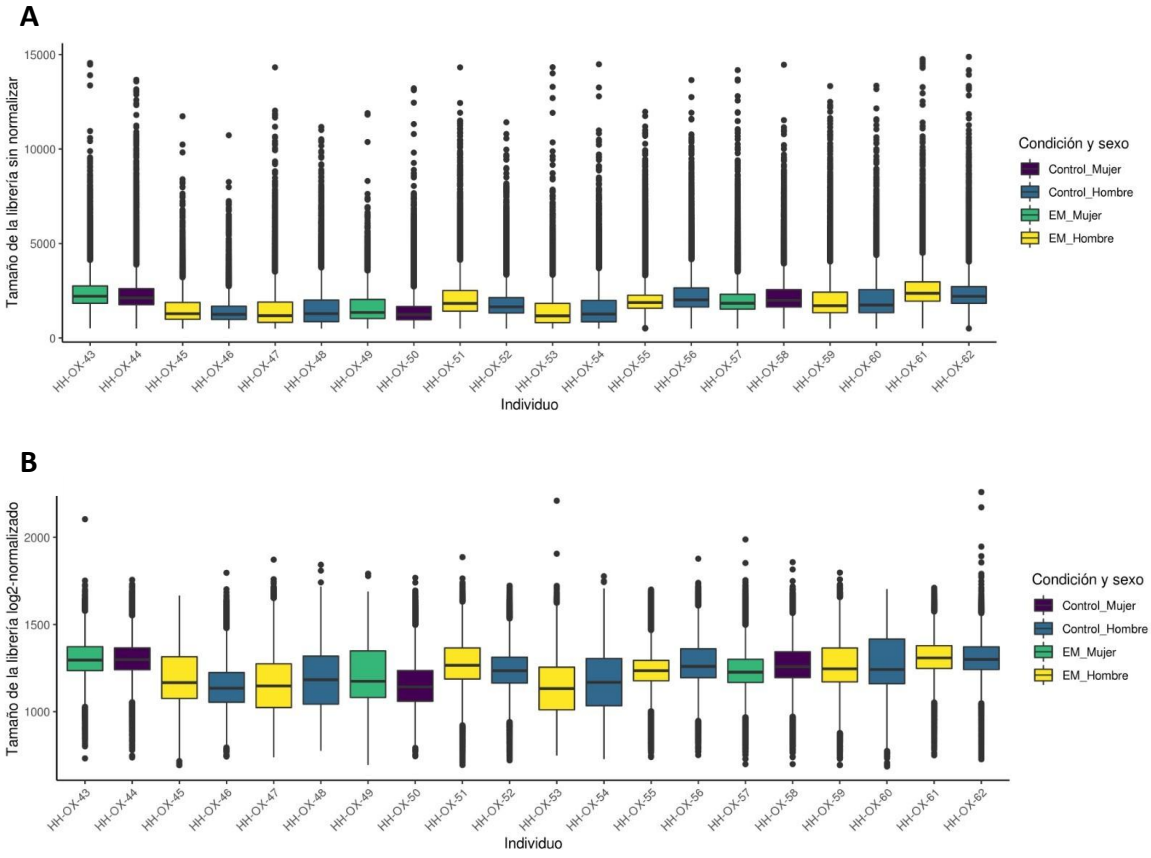


Figura 30. Distribución de los tamaños de las librerías sin normalizar (A) y normalizados con escala logarítmica (B) de las células por muestra para el estudio GSE144744 cohorte 3. *Diagramas de cajas que referencian gráficamente el primer cuartil (base inferior de la caja), la mediana (franja que divide la caja) y el tercer cuartil (base superior de la caja) de la distribución de los datos. Representación de valores atípicos (inferiores o superiores a 1,5 veces el rango intercuartílico) en formato de puntos. Colores establecidos en base a la condición y al sexo del individuo: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

5.2.3. Selección de genes altamente variables

Para determinar los genes que presentaron mayor variabilidad biológica el efecto *batch* fue bloqueado. Esta acción fue posible gracias a que los autores de los tres estudios incluyeron como metadato dicha información. Concretamente, para el estudio *Multiple*

sclerosis se diferencian dos efectos *batch:* el lote de captura de los núcleos únicos y el lote de secuenciación que, para su análisis, fueron combinados en una única variable. Por su parte, los datos de las dos cohortes del estudio GSE144744 incluían el lote de procesamiento de cada célula.

Tras obtener los resultados, se visualizaron las representaciones gráficas de la varianza total frente a los valores medios de expresión por lote de procesamiento. Sobre los diagramas de puntos se trazó la línea de tendencia correspondiente a la componente técnica de la varianza. Una muestra del resultado para cada estudio puede visualizarse en la figura 31, mientras que las gráficas pertenecientes al resto de lotes se adjuntan en el anexo II (*Figuras II.26 - II.28; https://doi.org/10.5281/zenodo.5068587*).

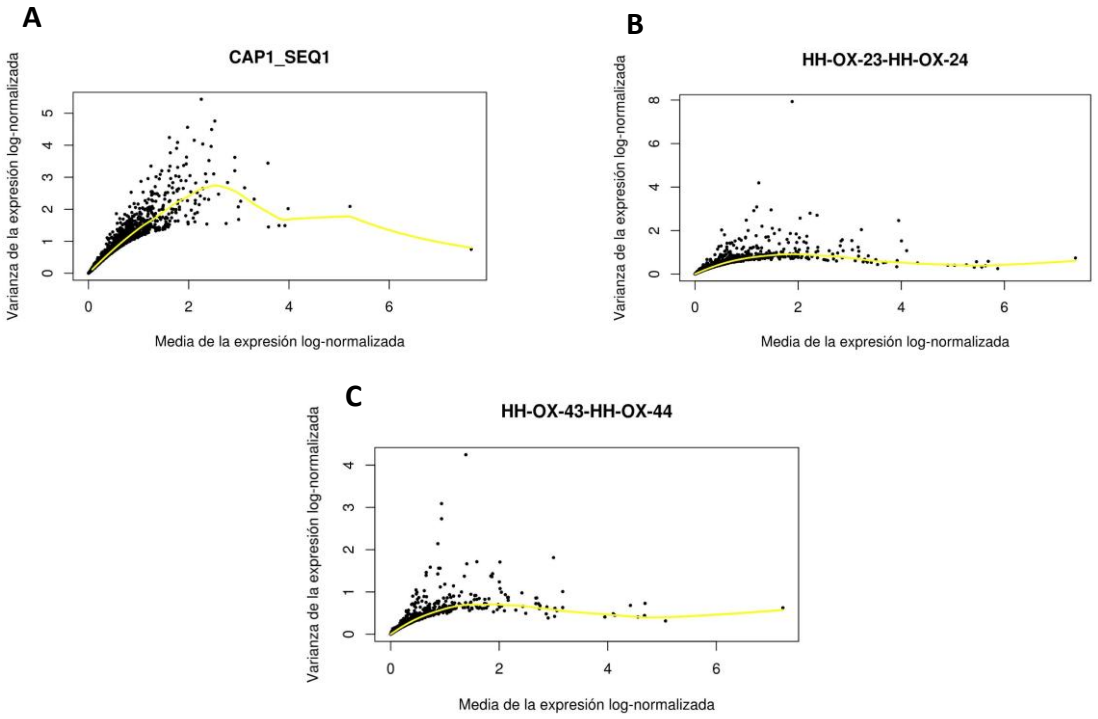


Figura 31. Representaciones gráficas de los valores de varianza frente al nivel medio de expresión normalizado a escala logarítmica de cada gen. *Visualización del resultado para el lote de procesamiento denominado CAP1_SEQ1 del estudio Multiple sclerosis (A), HH-OX-23-HH-OX-24 del estudio GSE144744 cohorte 1 (B) y HH-OX-43-HH-OX-44 del estudio GSE144744 cohorte 3 (C). Línea continua amarilla: modelización del componente técnico de la varianza.*

Para la mayoría de los lotes de procesamiento del estudio *Multiple sclerosis* se produce un sobreajuste al modelizar el componente técnico de la varianza *(Figura 31A y figura II.26 del anexo II; https://doi.org/10.5281/zenodo.5068587).* Este fenómeno se identifica cuando la línea de tendencia se dispone de forma errática generando un "camino" entre los puntos (genes) del gráfico, y ocurre cuando los genes con niveles mayores de expresión presentan valores de varianza muy dispersos. Para lograr un mejor ajuste, se repitió el proceso de modelización del componente técnico de la varianza desactivando los pesos de densidad considerados por la función *fitTrendVar (Figura 32, figura II.29 del anexo II; https://doi.org/10.5281/zenodo.5068587)*[63].

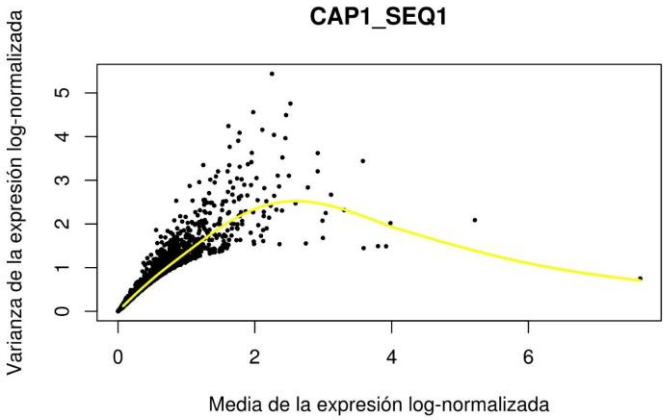


Figura 32. Representación gráfica de los valores de varianza frente al nivel medio de expresión normalizado a escala logarítmica de cada gen. *Visualización del resultado para el lote de procesamiento denominado CAP1_SEQ1 del estudio Multiple sclerosis. Línea continua amarilla: modelización del componente técnico de la varianza sin considerar los pesos de densidad.*

Como resultado de este proceso de selección se obtuvieron **2.093 genes para el estudio *Multiple sclerosis,* 1.825 para la cohorte 1 de GSE144744 y 1.768 para la cohorte 3 de GSE144744**.

5.2.4. Reducción de la dimensionalidad

La variabilidad biológica descrita por los genes altamente variables se ha resumido, para cada uno de los estudios, mediante la estrategia PCA. Tras aplicar el punto de corte por el método del codo, se seleccionaron 7 componentes principales en el estudio *Multiple sclerosis*, 4 en el estudio GSE144744 cohorte 1, y 3 en el estudio GSE144744 cohorte 3. Los gráficos de sedimentación y la representación gráfica en base a los dos primeros componentes principales de cada análisis pueden visualizarse en las figuras 33, 34 y 35, respectivamente. Por su parte, los mismos gráficos coloreados en función de la muestra de procedencia se encuentran adjuntos en el anexo II (*Figura II.30; https://doi.org/10.5281/zenodo.5068587*).

En conjunto, se puede observar en los tres estudios que las células no están separadas con claridad atendiendo a la condición, al sexo o al individuo de procedencia. Esta situación es esperable, ya que posteriormente se describe que la mayor fuente de variabilidad biológica está dirigida por el tipo celular al que pertenece cada célula.

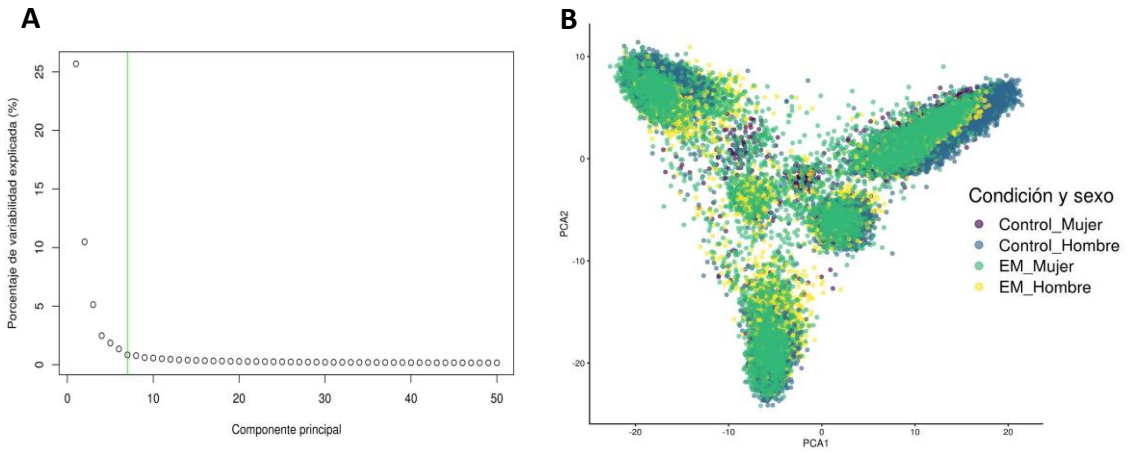


Figura 33. Gráfico de sedimentación (A) y representación de los dos primeros componentes principales (B) para el estudio Multiple sclerosis. *(A) Porcentaje de variabilidad explicada por cada componente principal. Línea horizontal verde: punto de corte establecido por el método del codo (componente principal número 7). (B) Diagrama de puntos de acuerdo a los valores del componente principal 1 (eje X) y el componente principal 2 (eje Y) de cada célula. Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

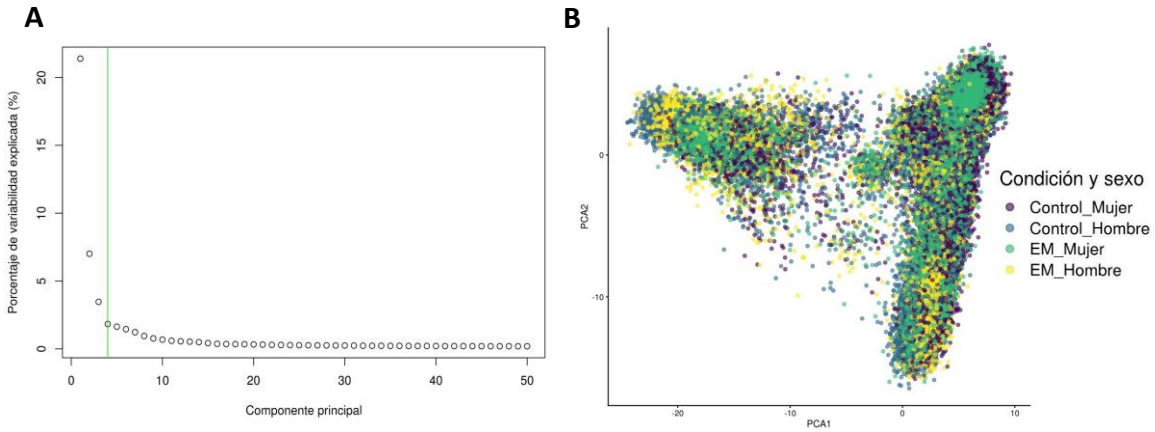


Figura 34. Gráfico de sedimentación (A) y representación de los dos primeros componentes principales (B) para el estudio GSE144744 cohorte 1. *(A) Porcentaje de variabilidad explicada por cada componente principal. Línea horizontal verde: punto de corte establecido por el método del codo (componente principal número 4). (B) Diagrama de puntos de acuerdo a los valores del componente principal 1 (eje X) y el componente principal 2 (eje Y) de cada célula. Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

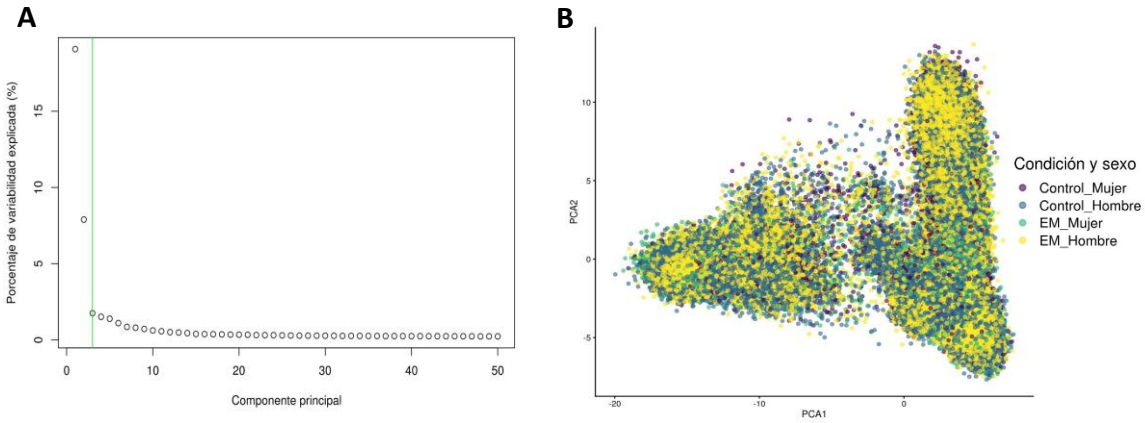


Figura 35. Gráfico de sedimentación (A) y representación de los dos primeros componentes principales (B) para el estudio GSE144744 cohorte 3. *(A) Porcentaje de variabilidad explicada por cada componente principal. Línea horizontal verde: punto de corte establecido por el método del codo (componente principal número 3). (B) Diagrama de puntos de acuerdo a los valores del componente principal 1 (eje X) y el componente principal 2 (eje Y) de cada célula. Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

Se realizaron diversas ejecuciones por estudio tanto para la metodología tSNE como UMAP con el objetivo de determinar qué número de vecinos deben ser evaluados para obtener las mejores representaciones gráficas. Un ejemplo de los diagramas de puntos resultantes puede visualizarse en el anexo II (*Figura II.31; https://doi.org/10.5281/zenodo.5068587*). Encontrar la mejor representación es un proceso que no está completamente acotado, donde el criterio seleccionado puede divergir en función de la perspectiva del investigador.

Atendiendo a las coordenadas calculadas por el método tSNE, se seleccionaron las visualizaciones generadas con un valor de semilla igual a 100 para los 3 estudios. Por el contrario, se especificó un número diferente de vecinos para cada caso: 40 para el estudio *Multiple sclerosis (Figura 35A),* 35 para el estudio GSE144744 cohorte 1 *(Figura 36A),* y 50 para el estudio GSE144744 cohorte 3 *(Figura 37A).*

Por su parte, las representaciones basadas en UMAP se obtuvieron en todos los casos estableciendo un valor de semilla de 100 y evaluando los 15 vecinos más próximos *(Figuras 35B, 36B y 37B).*

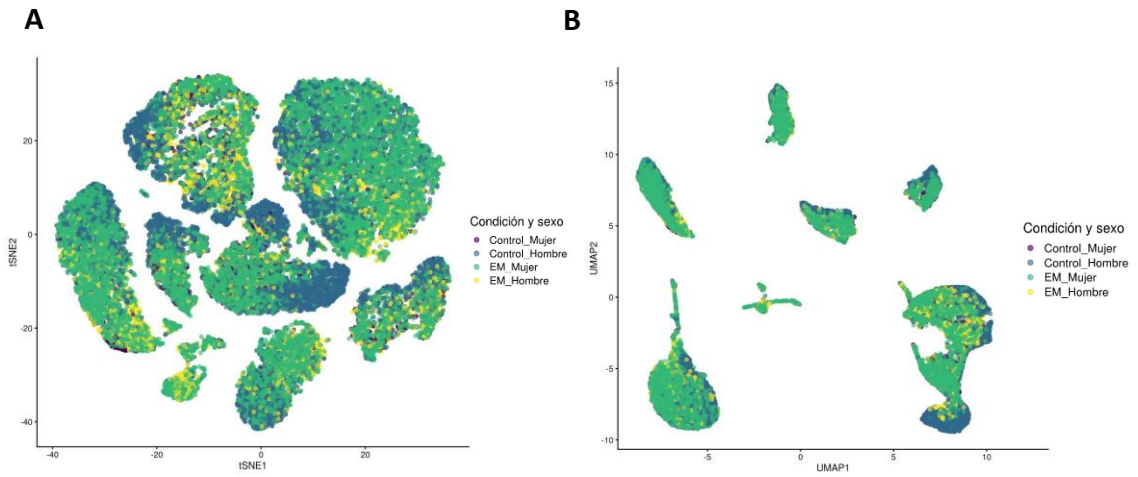


Figura 35. Diagramas de reducción de la dimensionalidad para el estudio Multiple sclerosis. *Coordenadas de cada célula calculadas con las estrategias (A) tSNE (semilla = 100, número de vecinos = 40) y (B) UMAP (semilla = 100, número de vecinos = 15). Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

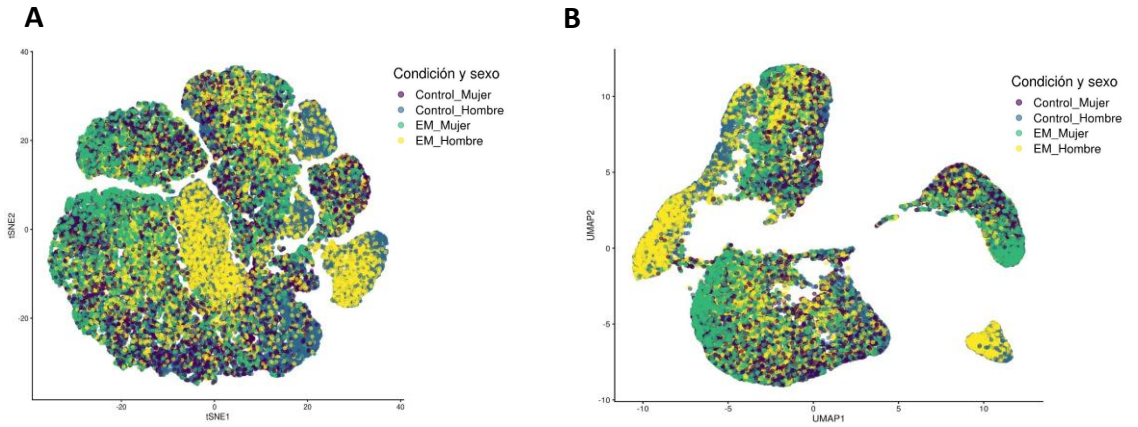


Figura 36. Diagramas de reducción de la dimensionalidad para el estudio GSE144744 cohorte 1. *Coordenadas de cada célula calculadas con las estrategias (A) tSNE (semilla = 100, número de vecinos = 35) y (B) UMAP (semilla = 100, número de vecinos = 15). Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

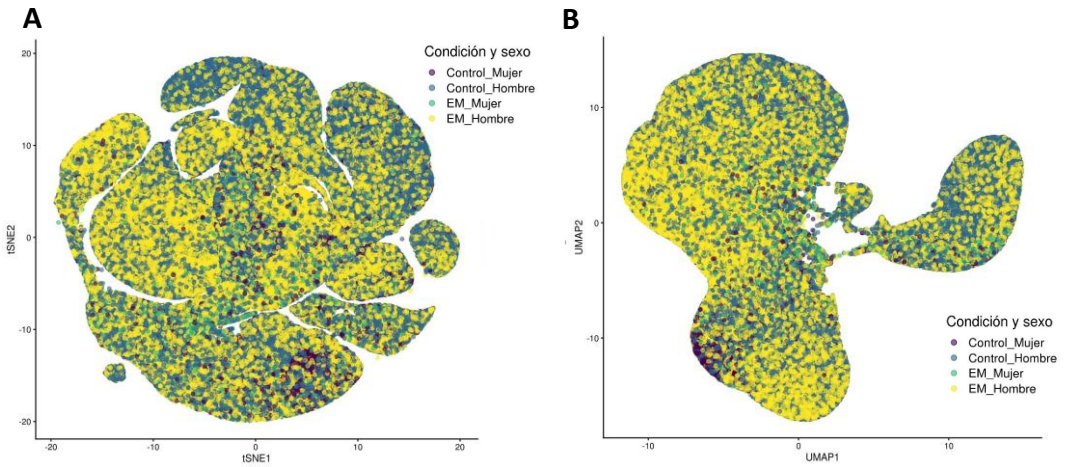


Figura 37. Diagramas de reducción de la dimensionalidad para el estudio GSE144744 cohorte 3. *Coordenadas de cada célula calculadas con las estrategias (A) tSNE (semilla = 100, número de vecinos = 50) y (B) UMAP (semilla = 100, número de vecinos = 15). Colores establecidos en función de la condición y el sexo del individuo de procedencia: morado (control-mujer), azul (control-hombre), verde (caso-mujer) y amarillo (caso-hombre). EM: esclerosis múltiple.*

5.2.5. Agrupamiento por identidad celular

Tras ejecutar el código descrito en el apartado de materiales y métodos de nombre homónimo, se obtuvieron 24 grupos en el estudio *Multiple sclerosis,* 20 grupos en la cohorte 1 de GSE144744 y 34 grupos en la cohorte 3 de GSE144744 *(Figuras 38A, 39A y 40A).*

En los resultados del estudio *Multiple sclerosis* y de la cohorte 1 de GSE144744 se observa que la mayoría de los grupos se encuentran definidos en un área concreta dispuestos de forma contigua. Esta situación queda reflejada en los indicadores de evaluación calculados. Por ejemplo, atendiendo a la pureza se puede observar cómo el grupo 8 del estudio *Multiple sclerosis* presenta la máxima pureza; mientras hay células del grupo 11 adyacentes a las del grupo 18 y viceversa *(Figura 38B).*

Por su parte, el agrupamiento de las células de la cohorte 3 de GSE144744 es mucho más difuso sobre las coordenadas tSNE. Esta situación también queda reflejada en el diagrama de pureza al observar cómo, en la mayoría de los casos, células pertenecientes a un grupo presentan como vecinas más próximas a las de otro.

Cabe destacar que *a priori* disponer de grupos más o menos definidos no presenta una connotación positiva ni negativa. Se asume que gran parte de los niveles de expresión de las células están gobernados por el tipo celular al que pertenezcan, situación que será evaluada en el siguiente apartado.

Por su parte, las representaciones gráficas sobre los dos primeros componentes principales y sobre las coordenadas UMAP de los grupos obtenidos se encuentran en el anexo II (*Figuras II.32 - II.34; https://doi.org/10.5281/zenodo.5068587*). Asimismo, en este anexo pueden visualizarse los gráficos que evalúan la modularidad y la estabilidad de cada estudio (*Figuras II.35 - II.37; https://doi.org/10.5281/zenodo.5068587*).

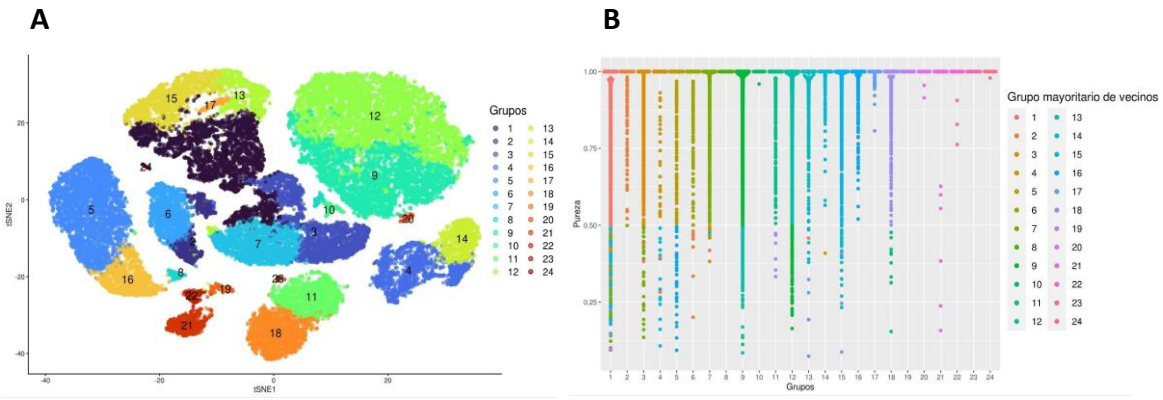


Figura 38. Agrupamiento por tipo celular (A) y pureza (B) del estudio Multiple sclerosis. *(A) Diagrama de puntos de las coordenadas tSNE obtenidas en el apartado de reducción de la dimensionalidad para cada célula. Colores establecidos en base al grupo de pertenencia por identidad celular (ver leyenda). (B) Disposición del grado de pureza por célula de cada grupo, coloreando cada célula en base al color del grupo mayoritario al que pertenecen sus vecinos (ver leyenda).*

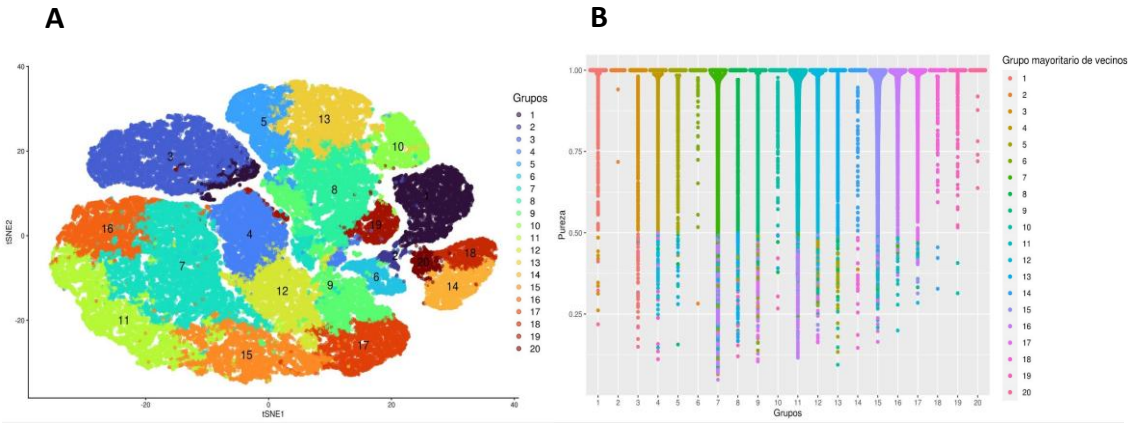


Figura 39. Agrupamiento por tipo celular (A) y pureza (B) del estudio cohorte 1 GSE144744. *(A) Diagrama de puntos de las coordenadas tSNE obtenidas en el apartado de reducción de la dimensionalidad para cada célula. Colores establecidos en base al grupo de pertenencia por identidad celular (ver leyenda). (B) Disposición del grado de pureza por célula de cada grupo, coloreando cada célula en base al color del grupo mayoritario al que pertenecen sus vecinos (ver leyenda).*

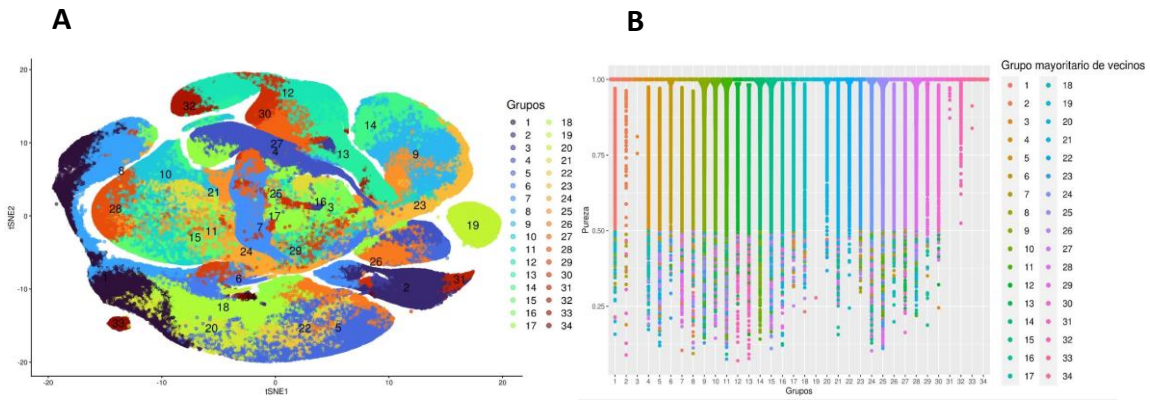


Figura 40. Agrupamiento por tipo celular (A) y pureza (B) del estudio cohorte 3 GSE144744*. (A) Diagrama de puntos de las coordenadas tSNE obtenidas en el apartado de reducción de la dimensionalidad para cada célula. Colores establecidos en base al grupo de pertenencia por identidad celular (ver leyenda). (B) Disposición del grado de pureza por célula de cada grupo, coloreando cada célula en base al color del grupo mayoritario al que pertenecen sus vecinos (ver leyenda).*

5.2.6. Anotación de los tipos celulares

La asignación del tipo celular al que pertenece cada célula se ha realizado con paquetes de R diferentes en función de si las células proceden de tejido nervioso o de PBMC. Por ello, los resultados se desglosan en dos subapartados independientes.

5.2.6.1. Anotación de los tipos celulares para el estudio *Multiple sclerosis*

Para el aislamiento de núcleos procedentes de tejido nervioso, los investigadores seccionaron muestras de corteza cerebral. Las muestras provenían mayoritariamente de la sustancia gris, aunque también contenían fracciones de sustancia blanca y meninges adyacentes. Este procedimiento experimental no ha sido realizado en la unidad de bioinformática donde se ha llevado el trabajo de fin de máster. Por ello, no se dispone de información detallada sobre los tipos celulares que se esperan encontrar, habiendo considerado que *a priori* pueden estar presentes todos los tipos celulares incluidos en el paquete *BRETIGEA.*

De este modo, se han evaluado los niveles de expresión de cada célula para asignarle uno de los seis tipos celulares principales de tejido nervioso: neuronas, astrocitos, células endoteliales, microglía, oligodendrocitos y precursores de oligodendrocitos. Los resultados obtenidos se representaron sobre los dos primeros componentes principales del análisis PCA *(Figura 41),* así como sobre las coordenadas obtenidas con las estrategias tSNE y UMAP (*Figura II.38 en el anexo II; https://doi.org/10.5281/zenodo.5068587*).

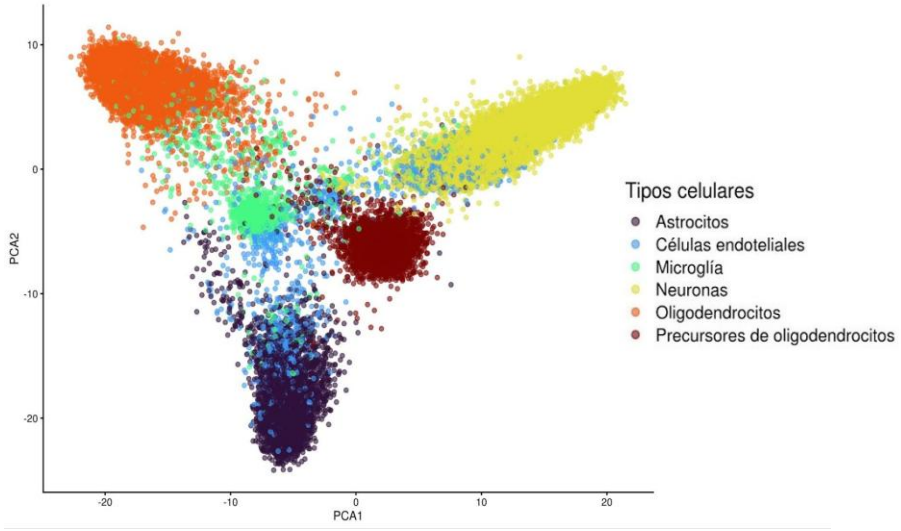


Figura 41. Representación de los dos primeros componentes principales para el estudio Multiple sclerosis. *Colores establecidos en base al tipo celular asignado para cada célula. Tipos celulares considerados: astrocitos, células endoteliales, microglía, neuronas, oligodendrocitos y precursores de oligodendrocitos (ver leyenda).*

En los 3 gráficos de reducción de la dimensionalidad se observa que la mayoría de las células se separa en base al tipo celular asignado; pudiendo visualizarse grupos definidos y separados. Por tanto, los pasos realizados anteriormente han permitido mantener la variación biológica que permite separar a las células en base a su tipo celular, descartando el ruido y variabilidad biológica no relevante.

Sin embargo, las células endoteliales se encuentran muy dispersas entre el resto de tipos celulares. Asimismo, cuando se modula el número de genes considerados en la función *brainCells,* las asignaciones para este tipo celular varían en gran medida. De hecho, al buscar genes marcadores en el siguiente paso del análisis para las células endoteliales, de los 19 resultantes no se consiguió identificar ninguno en la bibliografía (*Tabla 3*). Por todo ello, se decidió asumir que el estudio no presenta células endoteliales. Se volvió a ejecutar el procedimiento de asignación de tipos celulares considerando solamente 5 posibilidades: neuronas, astrocitos, microglía, oligodendrocitos y precursores de oligodendrocitos.

Tras volver a realizar la aproximación sin evaluar la posibilidad de presentar células endoteliales, se mantiene la separación de forma definida de los tipos celulares considerados (*Figura 42A, tSNE y UMAP en la figura II.39 en el anexo II; https://doi.org/10.5281/zenodo.5068587*). Además, cada tipo celular estaría conformado por varios de los grupos designados en agrupamiento por identidad celular (*Figura 42*); corroborando que esta categorización de las células está gobernada por una gran variabilidad biológica. Concretamente, los astrocitos se corresponderían con los grupos 5, 8, 16; la microglía con 20, 21 y 22; los oligodendrocitos con 9, 10, y 12; los precursores de oligodendrocitos con 11, 18 y 23; mientras que los grupos restantes representarían a las neuronas.

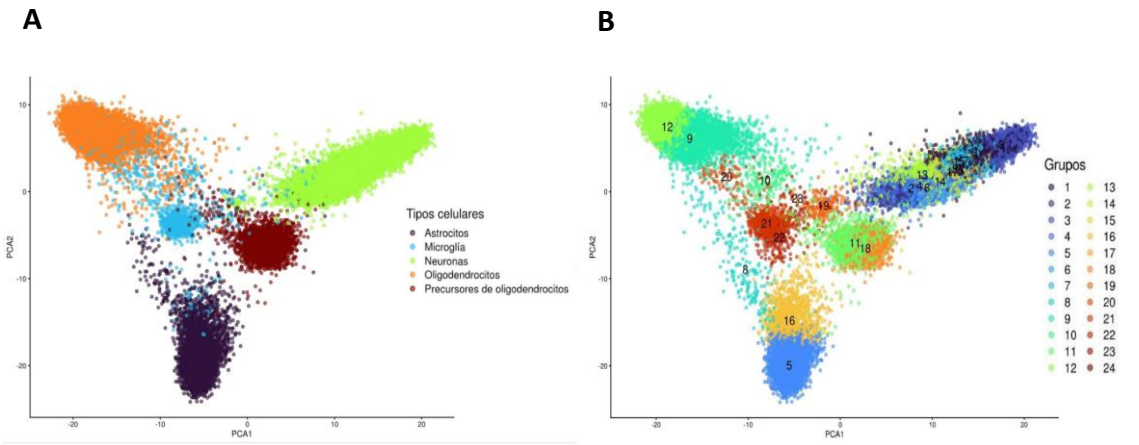


Figura 42. Representación de los dos primeros componentes principales para el estudio Multiple sclerosis. *Colores establecidos en base al tipo celular (A) o identidad celular (B) asignado para cada célula. Tipos celulares considerados: astrocitos, microglía, neuronas, oligodendrocitos y precursores de oligodendrocitos (ver leyenda).*

Como resultado final del proceso de anotación de tipos celulares se han identificado 5.321 núcleos de astrocitos, 1.589 de microglía, 16.676 de neuronas, 10.084 de oligodendrocitos y 3.446 de precursores de oligodendrocitos. Su distribución de acuerdo a la condición y al sexo del individuo de procedencia puede consultarse en la figura 43.

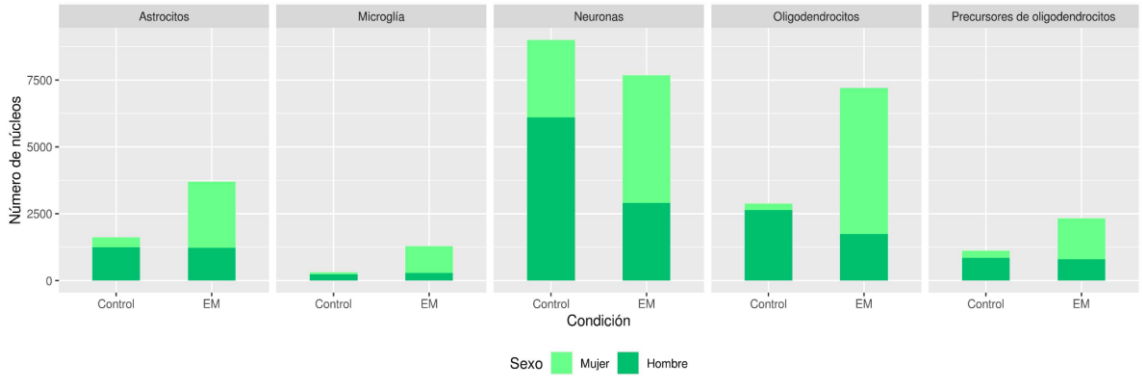


Figura 43. Número de núcleos asignados por tipo celular en el estudio Multiple sclerosis. *Distribución por condición (eje X) y sexo (color) de los núcleos en base a la anotación por tipo celular realizada en el trabajo. EM: esclerosis múltiple.*

5.2.6.2. Anotación de los tipos celulares para las dos cohortes del estudio GSE144744

Como se ha reflejado en el correspondiente apartado de materiales y métodos, tanto para la cohorte 1 como para la cohorte 3 se ha realizado el procesamiento con dos vertientes: considerando y sin considerar los agrupamientos por identidad celular

obtenidos en el paso anterior. En ambos casos, las células que la función *SingleR* no ha conseguido identificar con certeza han sido eliminadas. De nuevo, la asignación de tipos celulares se representa sobre los gráficos de reducción de la dimensionalidad para su evaluación *(Figuras 44 y 45 para las cohortes 1 y 3, respectivamente; gráficos tSNE y UMAP en las figuras II.40-II.43 en el anexo II; https://doi.org/10.5281/zenodo.5068587)*.

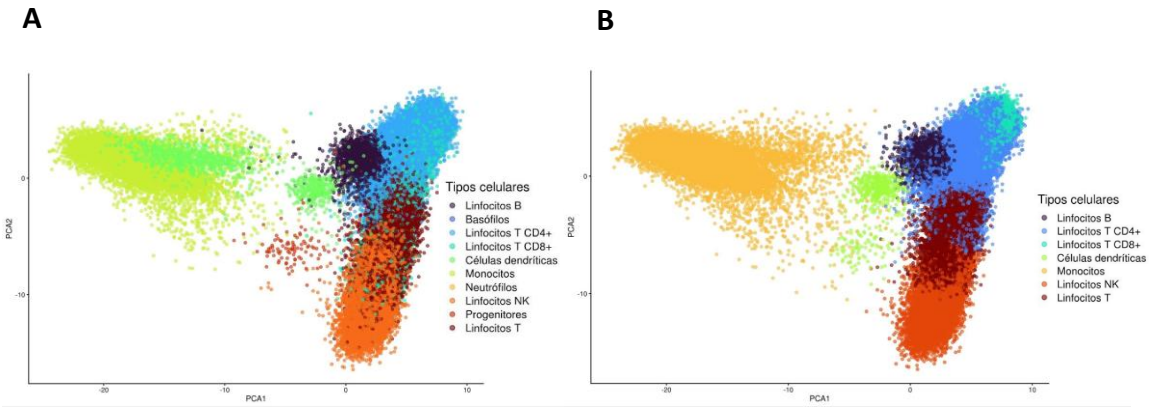


Figura 44. Representación de los dos primeros componentes principales para la cohorte 1 del estudio GSE144744. *Colores establecidos en base al tipo celular asignado para cada célula de forma individual (A) y atendiendo a los grupos por identidad celular (B). Tipos celulares considerados: linfocitos B, linfocitos T CD4+, linfocitos T CD8+, linfocitos NK, células dendríticas, monocitos, neutrófilos, basófilos y progenitores (ver leyenda).*

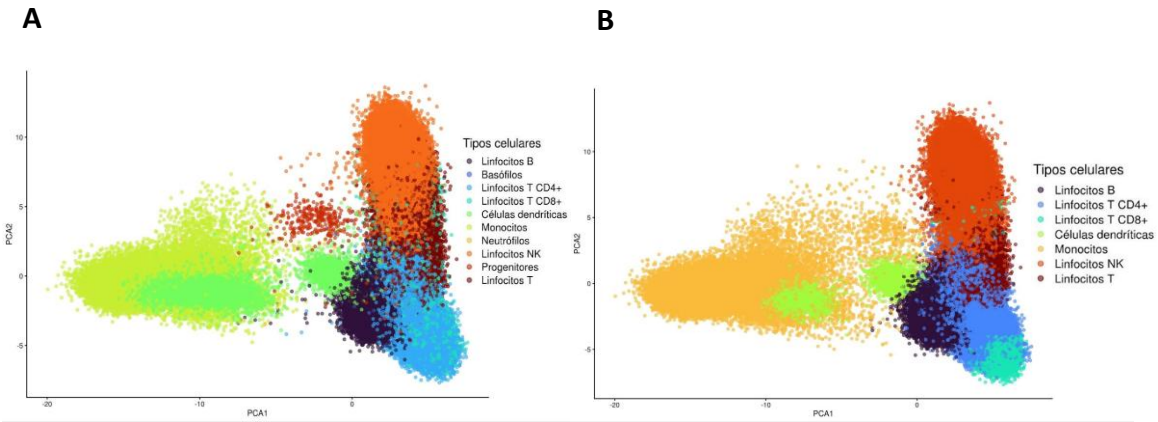


Figura 45. Representación de los dos primeros componentes principales para la cohorte 3 del estudio GSE144744. *Colores establecidos en base al tipo celular asignado para cada célula de forma individual (A) y atendiendo a los grupos por identidad celular (B). Tipos celulares considerados: linfocitos B, linfocitos T CD4+, linfocitos T CD8+, linfocitos T, linfocitos NK, células dendríticas, monocitos, neutrófilos, basófilos y progenitores (ver leyenda).*

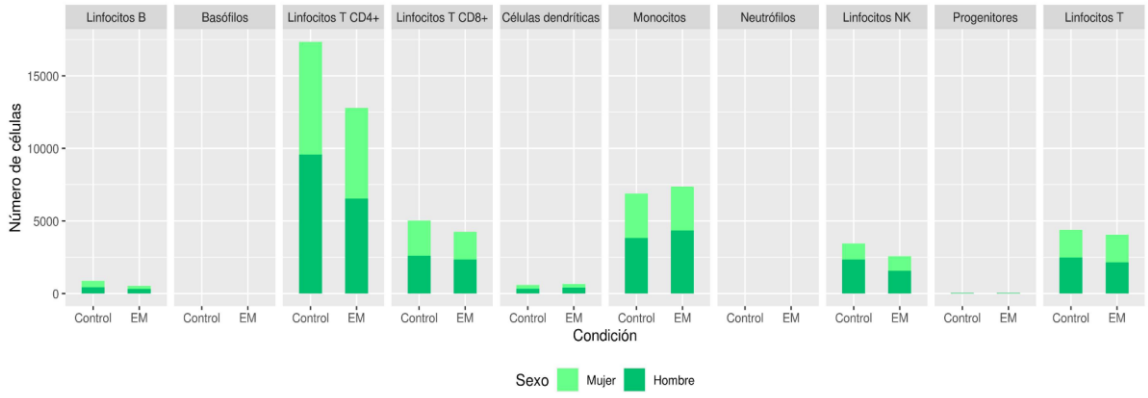
Tras analizar los resultados detalladamente, se decide seleccionar la **asignación de tipos celulares obtenida al evaluar las células de forma individual**. En primer lugar, esta estrategia presenta la ventaja de poder identificar las células polinuceladas (neutrófilos y basófilos) que no hayan sido eliminadas durante el proceso de obtención de PBMC. Estas células no se detectan atendiendo a los grupos de identidad celular porque no son suficientes para constituir un grupo independiente.

Seguidamente, es relevante destacar que se pretende identificar en las muestras de PBMC distintos tipos celulares con características muy similares. Concretamente, se evalúan diferentes clases de linfocitos (linfocitos B, linfocitos T CD4+, linfocitos T CD8+, linfocitos T y linfocitos NK); todos ellos procedentes de los progenitores linfoides. Asimismo, los monocitos tras ser estimulados pueden diferenciarse a células dendríticas[1]. Por todo ello, se considera plausible que los distintos tipos celulares no se dispongan como grupos independientes en los gráficos de reducción de la dimensionalidad *(Figuras 44 y 45).* Esta situación también se encuentra reflejada en los indicadores de los agrupamientos por identidad celular, al presentar resultados mucho más interconectados que en el estudio de tejido nervioso.

Por último, la anotación obtenida al evaluar las células de forma individual presenta mayor similitud a la designada por los autores del estudio en base a marcadores de superficie (*Figura II.44 en el anexo II; https://doi.org/10.5281/zenodo.5068587*).

Los resultados obtenidos para ambos estudios pueden visualizarse segregados por la condición y el sexo de los correspondientes individuos en la figura 46. Concretamente, en la cohorte 1 del estudio GSE144744 (*Figura 46A*) se han asignado 1.409 linfocitos B, 2 basófilos, 30.132 linfocitos T CD4+, 9.275 linfocitos T CD8+, 1.264 células dendríticas, 14.259 monocitos, 8 neutrófilos, 6.022 linfocitos NK, 114 progenitores y 8.444 linfocitos T.

Por otra parte, la cohorte 3 del estudio GSE144744 (*Figura 46B*) dispone de 6.323 linfocitos B, 10 basófilos, 95.763 linfocitos T CD4+, 31.434 linfocitos T CD8+, 4.475 células dendríticas, 54.842 monocitos, 33 neutrófilos, 34.096 linfocitos NK, 360 progenitores y 36.441 linfocitos T.



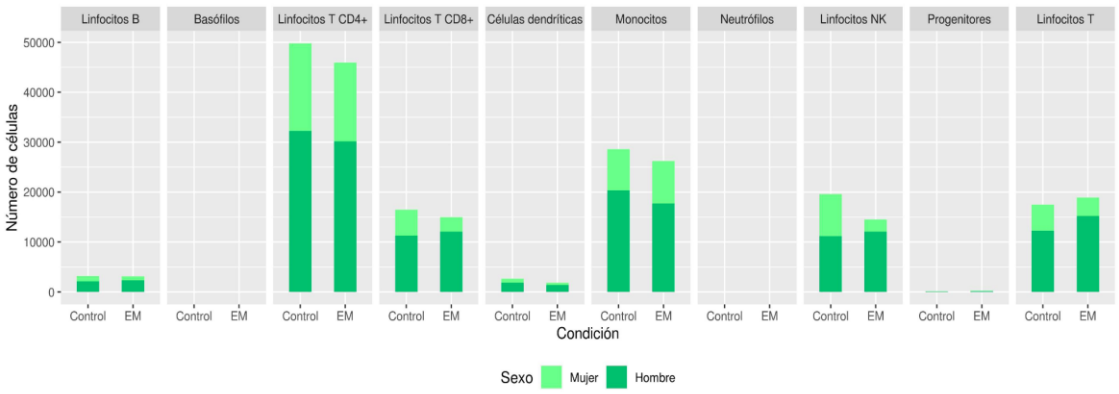


Figura 46. Número de células asignadas por tipo celular en las cohortes 1 (A) y 3 (B) del estudio GSE144744. *Distribución por condición (eje X) y sexo (color) de los núcleos en base a la anotación por tipo celular realizada en el trabajo. EM: esclerosis múltiple.*

5.2.7. Evaluación de los genes marcadores

Este apartado de resultados está centrado en los genes marcadores identificados por tipo celular, así como en la evidencia científica que se ha encontrado sobre ellos. Este procedimiento atribuye mayor robustez a la anotación de los tipos celulares realizada, al corroborar que los genes sobreexpresados que se han identificado en cada tipo celular ya han sido previamente descritos. En muchos casos se han obtenido una gran cantidad de genes marcadores, de los que se han seleccionado cinco como referencia para cada tipo celular. Por su parte, los genes marcadores por grupos de identidad celular pueden ser consultados en el anexo I (*Tablas I.2 - I.4; https://doi.org/10.5281/zenodo.5068587*).

Atendiendo a los resultados obtenidos para el estudio *Multiple sclerosis,* los genes marcadores resultantes considerando las células endoteliales se muestran en la tabla 3. En la búsqueda bibliográfica realizada no se encontró que ninguno de los 19 genes (*RBMS3, PTPRM, GAD2, AS2, KIT, CXCL14, AP1S2, SLC6A1, DGKD, AS1, SLC35F4, ZNF385D, EYA4, NHS, GRIN3A, GAPDH, KIAA1211, GAD1, UBAS43B*) tuviese un patrón de expresión que pudiese asociarse a dicho tipo celular.

TIPO CELULAR	GENES MARCADORES	EJEMPLOS DESCRITOS EN LA COMUNIDAD CIENTÍFICA
Astrocitos	483	*SOX9*[90], *ALDH1L1*[91], *AQP4*[91], *SLCA2*[91], *GJA1*[91]
Células endoteliales	19	-
Microglía	96	*CD74*[92], *CSFR1*[92], *C3*[92], *SYNDIG1*[92], *C1QB*[92]
Neuronas	454	*SLC17A7*[93], *STAT2*[93], *CCK*[93], *TMEM59*[91], *VSNL1*[91]
Oligodendrocitos	355	*PEX5*[91], *ANLN*[91], *CLDN11*[91], *ST18*[91], *TF*[91]
Precursores de oligodendrocitos	253	*PDGFRA*[91], *SOX10*[91], *TNR*[91], *EPN2*[91], *VCAN*[91]

Tabla 3. Genes marcadores para cada tipo celular en el estudio Multiple sclerosis considerando las células endoteliales. *Se establece para cada tipo celular (columna 1) el número de genes marcadores totales identificados (columna 2) y las referencias bibliográficas asociadas a cinco de ellos (columna 3).*

Por este motivo, junto con las razones adicionales expuestas en el subapartado 5.2.6.1, se volvió a realizar el procedimiento de anotación celular sin considerar las células endoteliales. Seguidamente, se establecieron de nuevo los genes marcadores que caracterizan cada tipo celular (*Tabla 4*). En este caso, se puede comprobar cómo los genes marcadores descritos aumentan para todos los tipos celulares. Asimismo, en todos los casos se han encontrado al menos 5 genes marcadores descritos en la bibliografía.

TIPO CELULAR	GENES MARCADORES	EJEMPLOS DESCRITOS EN LA COMUNIDAD CIENTÍFICA
Astrocitos	498	*SOX9*[90], *ALDH1L1*[91], *AQP4*[91], *SLCA2*[91], *GJA1*[91]
Microglía	98	*CD74*[92], *CSFR1*[92], *C3*[92], *SYNDIG1*[92], *C1QB*[92]
Neuronas	492	*SLC17A7*[93], *STAT2*[93], *CCK*[93], *TMEM59*[91], *VSNL1*[91]
Oligodendrocitos	362	*PEX5*[91], *ANLN*[91], *CLDN11*[91], *ST18*[91], *TF*[91]
Precursores de oligodendrocitos	264	*PDGFRA*[91], *SOX10*[91], *TNR*[91], *EPN2*[91], *VCAN*[91]

Tabla 4. Genes marcadores para cada tipo celular en el estudio Multiple sclerosis sin considerar las células endoteliales. *Se establece para cada tipo celular (columna 1) el número de genes marcadores totales identificados (columna 2) y las referencias bibliográficas asociadas a cinco de ellos (columna 3).*

Para la cohorte 1 del estudio GSE144744 no se obtiene ningún gen marcador por el procedimiento establecido (apartado de materiales y métodos 4.2.8.). En el caso de la cohorte 3, se obtiene un número muy limitado para algunos tipos celulares (*Tabla I.5 en el anexo I; https://doi.org/10.5281/zenodo.5068587*). Ambas situaciones eran previsibles al observar los gráficos de reducción de la dimensionalidad con la anotación por tipo celular *(Figuras 44A y 45A).* Al ser asignados evaluando las células de forma individual, los tipos celulares identificados en ambos estudios no se disponen de forma aislada unos de otros, por lo que sus patrones de expresión presentan mayor semejanza.

Por ello, se decide emplear una estrategia menos estricta para el cálculo de genes marcadores modificando los parámetros utilizados en la función *findMarkers.* Concretamente, se considera gen marcador aquel sobreexpresado en el tipo celular a evaluar y que presenta un p-valor significativo en al menos la mitad de los contrastes de expresión diferencial. De este modo, ya no es necesario que el gen presente una evidencia significativa de sobreexpresión contra cada uno de los tipos celulares restantes.

En ambos estudios se observa que con esta aproximación son identificadas combinaciones de genes marcadores que han sido previamente descritos por la comunidad científica *(Tablas 5 y 6)*. Para tratar de homogeneizar el análisis, se ha procurado que los genes evaluados en ambas cohortes sean los mismos. Con todo ello, los resultados obtenidos permiten aportar un mayor soporte a la asignación de tipos celulares realizada.

TIPO CELULAR	GENES MARCADORES	EJEMPLOS DESCRITOS EN LA COMUNIDAD CIENTÍFICA
Linfocitos B	498	*CD79A*[94], *CD79B*[94], *HLA-DRA*[91], *BLK*[94], *CD74*[94]
Linfocitos T CD4+	350	*ANK3*[94], *MXI1*[94], *GIMAP4*[95], *CD28*[91], *CCR4*[91]
Linfocitos T CD8+	323	*CD8A*[94], *CD8B*[94], *PRF1*[94] *KLRG1*[94], *CS27*[94]
Linfocitos T	270	*GATA3*[94], *CD2*[94], *CD3G*[94], *TCF7*[94], *ZAP70*[94]
Linfocitos NK	307	*NKG7*[91], *KLRD1*[91], *CD160*[96], *KLRF1*[96], *PRF1*[96]
Monocitos	554	*APOBEC3A*[91], *CD14*[91], *ADA2*[91], *NKFB1*[96], *CCR1*[97]
Células dendríticas	786	*IRF8*[98], *IRF4*[98], *ZEB2*[98], *ID2*[98], *KLF4*[98]

Neutrófilos	13	*FCN1*[99], *SAT1*[100], *S100A8*[100], *CD14*[100], *SC100A9*[100]
Progenitores	178	*DMTN*[101], *MPIG6B*[101], *ITGA2B*[101], *VCL*[101], *PF4*[101]
Basófilos	0	-

Tabla 5. Genes marcadores para cada tipo celular en el estudio GSE144744 cohorte 1. *Se establece para cada tipo celular (columna 1) el número de genes marcadores totales identificados (columna 2) y las referencias bibliográficas asociadas a cinco de ellos (columna 3).*

TIPO CELULAR	NÚMERO DE GENES MARCADORES	EJEMPLOS DESCRITOS EN LA COMUNIDAD CIENTÍFICA
Linfocitos B	674	*CD79A*[94], *CD79B*[94], *HLA-DRA* [91], *CD74*[94], *CD22*[94]
Linfocitos T CD4+	423	*ANK3*[94], *MXI1*[94], *GIMAP4*[95], *CD28*[91], *CD3G*[91]
Linfocitos T CD8+	339	*CD8A*[94], *CD8B*[94], *PRF1*[94] *KLRG1*[94], *CS27*[94]
Linfocitos T	312	*GATA3*[94], *CD2*[94], *CD3G* [94], *TCF7*[94], *ZAP70*[94]
Linfocitos NK	327	*NKG7*[91], *KLRD1*[91], *CD160*[96], *KLRF1*[96], *PRF1*[96]
Monocitos	332	*APOBEC3A*[91], *CD14*[91], *NKFB1*[97], *CD36*[97], *C1QA*[97]
Células dendríticas	735	*IRF8*[98], *ZEB2*[98], *ID2*[98], *KLF4*[98], *FCER1A*[98]
Neutrófilos	40	*FCN1*[99], *SAT1*[100], *S100A8*[100], *CD14*[100], *SC100A9*[100]
Progenitores	181	*MPIG6B*[101], *ITGA2B*[101], *VCL*[101], *PF4*[101], *HBA1*[101]
Basófilos	0	-

Tabla 6. Genes marcadores para cada tipo celular en el estudio GSE144744 cohorte 3. *Se establece para cada tipo celular (columna 1) el número de genes marcadores totales identificados (columna 2) y las referencias bibliográficas asociadas a cinco de ellos (columna 3).*

Por último, también se determinaron los genes marcadores para los tipos celulares asignados atendiendo a los grupos obtenidos por identidad celular (*Tabla I.6;*

https://doi.org/10.5281/zenodo.5068587). Para ambos estudios se observó el mismo patrón. Por una parte, se identificó un número más reducido de genes marcadores respecto a la asignación de tipo celular con células individuales. Adicionalmente, la mayoría de los genes consultados en la bibliografía no se encuentran sobreexpresados de forma significativa respecto al resto de tipos celulares.

5.2.8. Análisis bioestadísticos

Este subapartado está destinado a la exposición de los resultados obtenidos para la caracterización de la enfermedad de esclerosis múltiple con perspectiva de sexo.

5.2.8.1. Análisis de abundancia diferencial

Tras ejecutar el código desarrollado para aplicar las estrategias "método *orchestrating*" y "método *Schimer et al. 2019*" en cada estudio, se obtuvieron los resultados resumidos en la tabla 7.

En el estudio *Multiple sclerosis* se han evaluado tipos celulares de tejido nervioso. Ambos abordajes concuerdan en los resultados obtenidos para la comparación entre mujeres enfermas y sanas. Concretamente, se identifica una reducción en el número de neuronas, así como un aumento de células de microglía y oligodendrocitos en pacientes de EM respecto a la condición control. Además, con el "método *orchestrating*" también se detecta un mayor número de células de microglía en pacientes de EM hombres respecto a los controles respectivos. Por otra parte, el "método *Schimer et al. 2019*" determina que en la enfermedad de EM hay un mayor número de oligodendrocitos si los pacientes son mujeres.

Atendiendo a los estudios basados en muestras de PBMC, en la cohorte 1 de GSE144744 no se identifica con ninguno de los métodos resultados significativos para ninguna comparación. Por el contrario, para la cohorte 3 de GSE144744 ambos abordajes corroboran que en individuos que sufren EM hay un mayor número de linfocitos NK si el paciente es mujer. De nuevo, cada estrategia también presenta resultados particulares: con el "método *orchestrating*" se identifica un mayor número de linfocitos NK en mujeres enfermas respecto a mujeres sanas; mientras que con el "método *Shimer et al. 2019*" se establece en pacientes un menor número de células dendríticas en el respectivo contraste para hombres.

ESTUDIO	MÉTODO	CONTRASTE MUJER	CONTRASTE HOMBRE	CONTRASTE COMPLETO
Multiple sclerosis	*orchestrating*	Neuronas (-) Microglía (+) Oligodendrocitos (+)	Microglía (+)	-
	Schimer et al. 2019	Neuronas (-) Microglía (+) Oligodendrocitos (+)	-	Oligodendrocitos (+)
cohorte 1 GSE144744	*orchestrating*	-	-	-
	Schimer et al. 2019	-	-	-
cohorte 3 GSE144744	*orchestrating*	Linfocitos NK (-)	-	Linfocitos NK (-)
	Schimer et al. 2019	-	Células dendríticas (-)	Linfocitos NK (-)

Tabla 7. Resultados significativos para los análisis de abundancia diferencial. *Tipos celulares que varían en cantidad de forma significativa para los contrastes EM_Mujer – Control_Mujer (columna 3), EM_Hombre – Control_Hombre (columna 4), y (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre) (columna 5) en los estudios Multiple sclerosis (filas 1 y 2), y las cohortes 1 (filas 3 y 4) y 3 (filas 5 y 6) de GSE144744. Resultados obtenidos en la aplicación de los abordajes "método orchestrating" y "método Schimer et al. 2019". Símbolo entre paréntesis: signo del correspondiente logFC. EM: esclerosis múltiple; -: ausencia de tipos celulares.*

5.2.8.2. Análisis de expresión diferencial

En primer lugar, se realizó la **estrategia adaptada del análisis de RNA-seq** para los tres estudios. En todos los casos, se llevó a cabo un primer abordaje incluyendo exclusivamente en el modelo la variable conjunta de condición y sexo. En este abordaje se evaluaron de forma independiente cada uno de los tipos celulares. También se realizaron los contrastes de forma general; en los que se utilizaron la totalidad de los datos sin ser separados por tipo celular.

Como resultado, en el estudio *Multiple sclerosis* solamente se identificaron genes significativos en los contrastes entre hombres enfermos y sanos para los siguientes tipos celulares: neuronas, oligodendrocitos, y precursores de oligodendrocitos. Tanto en la

cohorte 1 como en la cohorte 3 del estudio GSE144744, referente a células del sistema inmunitario, no se identificó ningún gen significativo en ninguna de las comparaciones.

Por otra parte, en el análisis exploratorio se observó en todos los casos que las *pseudocélula*s no se separaban de acuerdo a la condición y al sexo del individuo de procedencia (*Figura II.45 en el anexo II; https://doi.org/10.5281/zenodo.5068587*). Esta situación, junto con los resultados del análisis de expresión diferencial, permitió plantear la siguiente hipótesis: existen otros aspectos descritos en los metadatos (por ejemplo, la edad) que enmascaran la variabilidad de interés. Con el objetivo de testar este planteamiento, se intentó añadir en el modelo diversas variables control. Sin embargo, la metodología implementada en las funciones de *edgeR* no permitió realizar esta consideración.

Atendiendo a la **estrategia específica para datos de scRNA-seq y snRNA-seq** se llevó a cabo un procedimiento muy similar. Sin embargo, en este caso no se desarrollaron los contrastes generales debido al alto coste computacional y temporal de los mismos. En primer lugar, se realizaron dos abordajes con los datos del estudio *Multiple sclerosis*, ya que estos presentan menores requerimientos de almacenamiento y procesamiento. La diferencia entre ambos radica en las variables consideradas en el modelo. El primero de ellos solamente incorpora la variable conjunta condición y sexo, y la variable escalada del número de genes para los cuales se ha detectado su expresión por célula. En la segunda aproximación se añaden las variables control consideradas en el análisis exploratorio: muestra de procedencia, estado de la lesión, edad, región cerebral afectada, lote de captura, lote de secuenciación y estado del ciclo celular.

Con ambos abordajes se obtuvieron genes diferencialmente expresados de forma significativa para todos los contrastes en todos los tipos celulares evaluados. Sin embargo, a nivel cuantitativo, el número de genes significativos siempre es inferior en la segunda aproximación respecto a la primera. (*Tabla 8 y tabla I.8 en el anexo I; https://doi.org/10.5281/zenodo.5068587*). Adicionalmente, los gráficos de análisis exploratorio reflejan que las principales fuentes de variabilidad enmascaran las diferencias debidas a la condición y al sexo de los individuos (*Figura II.46 en el anexo II; https://doi.org/10.5281/zenodo.5068587*). Con todo ello, se decide mantener los resultados de la segunda aproximación, considerando en el modelo las variables control.

TIPO CELULAR	DIRECCIÓN	COMPARACIÓN MUJER	COMPARACIÓN HOMBRE	COMPARACIÓN COMPLETA
Astrocitos	+	90	961	315
	-	271	597	427
	Total	361	1558	742
Microglía	+	118	129	174
	-	463	184	254
	Total	651	313	428
Neuronas	+	1039	3558	1886
	-	1884	2602	3017
	Total	2923	6160	4903
Oligodendrocitos	+	65	777	124
	-	144	466	162
	Total	209	1243	286
Precursores de oligodendrocitos	+	49	305	104
	-	92	171	112
	Total	141	476	216

Tabla 8. Número de genes significativos por comparación y tipo celular en el estudio Multiple sclerosis. *Dirección (columna 2): signo del valor logFC calculado (+: sobreexpresión en el primer término del contraste o infra expresión en el segundo término del contraste; -: sobreexpresión en el segundo término del contraste o infra expresión en el primer término del contraste; Total: número de genes significativos sin considerar el signo del valor logFC). Comparación mujer (columna 3): EM_Mujer – Control_Mujer. Comparación hombre (columna 4): EM_Hombre – Control_Hombre. Comparación completa (columna 5): (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). EM: esclerosis múltiple.*

Para homogeneizar el procesamiento de los tres estudios, los análisis de expresión diferencial para las cohortes 1 y 3 del estudio GSE144744 también se han realizado incluyendo en el modelo variables control. En ambos casos, las variables consideradas son: muestra de procedencia, efecto *batch*, edad, tratamientos previos y el ciclo celular. Los resultados del análisis exploratorio pueden consultarse en las figuras II.47 y II.48 del anexo II (https://doi.org/10.5281/zenodo.5068587); mientras que el número de genes diferencialmente expresados queda reflejado en las tablas 9 y 10 para la cohorte 1 y la cohorte 3, respectivamente.

TIPO CELULAR	DIRECCIÓN	COMPARACIÓN MUJER	COMPARACIÓN HOMBRE	COMPARACIÓN COMPLETA
Linfocitos B	+	43	20	7
	-	41	22	7
	Total	84	42	14
Linfocitos T CD4+	+	1052	497	876
	-	1281	608	813
	Total	2333	1105	1689
Linfocitos T CD8+	+	370	229	321
	-	364	280	209
	Total	734	509	630
Linfocitos T	+	179	143	142
	-	142	137	77
	Total	321	280	219
Linfocitos NK	+	296	73	226
	-	225	101	113
	Total	521	174	339
Monocitos	+	699	303	517
	-	805	432	500
	Total	1504	735	1017
Células dendríticas	+	100	19	48
	-	116	19	33
	Total	216	38	81

Tabla 9. Número de genes significativos por comparación y tipo celular en el estudio GSE144744 cohorte 1. *Dirección (columna 2): signo del valor logFC calculado (+: sobreexpresión en el primer término del contraste o infraexpresión en el segundo término del contraste; -: sobreexpresión en el segundo término del contraste o infraexpresión en el primer término del contraste; Total: número de genes significativos sin considerar el signo del valor logFC). Comparación mujer (columna 3): EM_Mujer – Control_Mujer. Comparación hombre (columna 4): EM_Hombre – Control_Hombre. Comparación completa (columna 5): (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). EM: esclerosis múltiple.*

TIPO CELULAR	DIRECCIÓN	COMPARACIÓN MUJER	COMPARACIÓN HOMBRE	COMPARACIÓN COMPLETA
Linfocitos B	+	162	134	125
	-	206	102	131
	Total	368	236	256
Linfocitos T CD4+	+	1253	1156	1291
	-	2638	1529	2234
	Total	3891	2685	3525
Linfocitos T CD8+	+	750	971	1030
	-	909	1166	1109
	Total	1659	2137	2139
Linfocitos T	+	1671	469	1092
	-	1649	509	1077
	Total	3320	978	2169
Linfocitos NK	+	904	479	643
	-	942	494	611
	Total	1846	973	1259
Monocitos	+	1642	1086	1469
	-	2600	1165	1906
	Total	4242	2251	3375
Células dendríticas	+	216	169	101
	-	177	228	107
	Total	393	397	208

Tabla 10. Número de genes significativos por comparación y tipo celular en el estudio GSE144744 cohorte 3. *Dirección (columna 2): signo del valor logFC calculado (+: sobreexpresión en el primer término del contraste o infraexpresión en el segundo término del contraste; -: sobreexpresión en el segundo término del contraste o infraexpresión en el primer término del contraste; Total: número de genes significativos sin considerar el signo del valor logFC). Comparación mujer (columna 3): EM_Mujer – Control_Mujer. Comparación hombre (columna 4): EM_Hombre – Control_Hombre. Comparación completa (columna 5): (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre). EM: esclerosis múltiple.*

5.3. Integración de los resultados del análisis primario

Tras obtener para cada estudio los listados de genes significativos en el análisis de expresión diferencial, se profundiza en sus resultados describiendo los aspectos comunes desde diversas perspectivas. Todos los abordajes planteados en este apartado se han realizado para los resultados obtenidos de la comparación (EM_Mujer – Control_Mujer) - (EM_Hombre – Control_Hombre) por ser de mayor interés.

En primer lugar, se ha determinado qué genes presentan un **patrón de expresión estadísticamente significativo, por estudio, independientemente del tipo celular**. En el caso del estudio *Multiple sclerosis,* de tejido nervioso, se han calculado las intersecciones entre los resultados obtenidos para los astrocitos, la microglía, las neuronas, los oligodendrocitos y los precursores de oligodendrocitos. Por su parte, los tipos celulares del sistema inmunitario analizados para las cohortes 1 y 3 de GSE144744 son los linfocitos B, los linfocitos T, los linfocitos T CD4+, los linfocitos T CD8+, los linfocitos NK, los monocitos y las células dendríticas. Los correspondientes diagramas de Venn que interconectan todos los tipos celulares pueden consultarse en el anexo II (*Figuras II.49 - II.51; https://doi.org/10.5281/zenodo.5068587*).

En el caso del estudio *Multiple sclerosis* (*Tabla 11*), ninguno de los tres genes significativos sobreexpresados en mujeres respecto a hombres se encuentra descrito en el repositorio *OpenTargets.* Sin embargo*,* se identifican 4 procesos biológicos sobrerrepresentados:

- *adherens junction organization* (GO:0034332)
- *homophilic cell adhesion via plasma membrane adhesion molecules* (GO:0007156)
- *cell-cell junction organization* (GO:0045216)
- *cell-cell adhesion via plasma-membrane adhesion molecules* (GO:0098742)

Por su parte, de los 11 genes sobreexpresados en hombres, 5 de ellos se encuentran asociados a la enfermedad en base a los datos de *OpenTargets*. En conjunto, este grupo de genes presenta sobrerrepresentada la función *sequestring of actin monomers* (GO:0042989).

Atendiendo a los resultados para las células del sistema inmunitario procedentes de sangre periférica (*Tabla 11*), en el estudio cohorte 1 GSE144744 solamente se encuentra sobreexpresado el gen *RASAL2* en mujeres. Al evaluar los resultados procedentes de la cohorte 3 GSE144744, el gen *EGR1* se identifica sobreexpresado en mujeres, y *FAM118A, HLA-A* y *HLA-DRB5* en hombres. Estos tres genes engloban conjuntamente a 118 procesos biológicos sobrerrepresentados respecto del total de genes analizados. Para facilitar su interpretación, se reduce el número de funciones eliminando la redundancia de términos mediante la herramienta web REVIGO (por sus siglas del inglés *REduce and VIsualize Gene Ontology*)[102]. En detalle, se han proporcionado a la plataforma los identificadores GO significativos y su p-valor ajustado. También se ha indicado que el organismo es *Homo sapiens,* mientras que el resto de parámetros se ha mantenido por defecto. Como

resultado, se mantienen 44 procesos biológicos (*Tabla I.8 en el anexo I; https://doi.org/10.5281/zenodo.5068587*), de los cuales se citan a continuación 5 a modo de ejemplo:

- *regulation of leukocyte proliferation* (GO:0070663)
- *antigen processing and presentation* (GO:0019882)
- *cytokine production involved in immune response* (GO:0002367)
- *response to interferon-gamma* (GO:0034341)
- *T cell mediated immunity* (GO:0002456)

ESTUDIO	MUESTRA	DIRECCIÓN	GENES COMUNES	GENES COMUNES
Multiple sclerosis	tejido nervioso	+	3	*CADM2, PCDH9, UBA6-AS1*
		-	11	*ACTB, BEX1, GAPDH, LINGO1, NDRG4, NRGN, TMSB10, TMSB4X, TUBA1B, TUBA4A, VSNL1*
Cohorte 1 GSE144744	sangre periférica	+	1	*RASAL2*
		-	0	
Cohorte 3 GSE144744	sangre periférica	+	1	*EGR1*
		-	3	*FAM118A, HLA-A, HLA-DRB5*

Tabla 11. Genes con expresión diferencial estadísticamente significativa para todos los tipos celulares de cada estudio. *Resultados desglosados por estudio (columna 1) indicando el origen de las muestras analizadas (columna 2) y por signo del valor logFC (columna 3, +: sobreexpresión en mujeres o infraexpresión en hombres; -: sobreexpresión en hombres o infraexpresión en mujeres); exponiendo el número (columna 4) y el identificador Gene Symbol (columna 5) de los genes con expresión significativa en todos los tipos celulares. Genes color verde: genes asociados a la enfermedad en la base de datos OpenTargets.*

Por último, dos de los estudios analizados de forma individual (cohortes 1 y 3 de GSE144744) contienen datos procedentes del mismo tipo de muestra (células del sistema inmunitario procedentes de sangre periférica). Esta situación permite identificar, para posteriormente caracterizar, **qué genes son comunes en los resultados de ambos estudios por tipo celular**. En detalle, se ha calculado la intersección entre los resultados

del análisis de expresión diferencial de las cohortes 1 y 3 de GSE144744 para los linfocitos T CD4+ y los linfocitos T CD8+. Estos tipos celulares se han seleccionado por su relevancia en los procesos de desmielinización y neuroinflamación de la EM[24]. De nuevo, los diagramas de Venn resultantes pueden consultarse en el anexo II (*Figuras II.52 y II.53; https://doi.org/10.5281/zenodo.5068587*).

Para todas las comparaciones se han identificado genes con expresión diferencial significativa en los dos estudios (*Tabla 12*). Asimismo, en todos los casos se dispone de genes que previamente se han asociado con la enfermedad de EM. Atendiendo su caracterización funcional, los genes sobreexpresados en los **linfocitos T CD4+** de mujeres presentan 14 procesos biológicos enriquecidos:

- *RNA splicing, via transesterification reactions with bulged adenosine as nucleophile* (GO:0000377)
- *mRNA splicing, via spliceosome* (GO:0000398)
- *RNA splicing, via transesterification reactions* (GO:0000375)
- *ribonucleoprotein complex assembly* (GO:0022618)
- *ribonucleoprotein complex subunit organization* (GO:0071826)
- *RNA splicing* (GO:0008380)
- *regulation of RNA splicing* (GO:0043484)
- *response to interferon-beta* (GO:0035456)
- *mRNA processing* (GO:0006397)
- *RNA localization* (GO:0006403)
- *mRNA transport* (GO:0051028)
- *V(D)J recombination* (GO:0033151)
- *establishment of RNA localization* (GO:0051236)
- *regulation of mRNA splicing, via spliceosome* (GO:0048024)

TIPO CELULAR	DIRECCIÓN	COHORTE 1 GSE144744	COHORTE 3 GSE144744	GENES COMUNES	GENES COMUNES ASOCIADOS
Linfocitos T CD4+	+	876	1291	138	47
	-	813	2234	207	56
Linfocitos T CD8+	+	321	1030	44	21
	-	209	1109	36	8

Tabla 12. Genes consenso entre los estudios cohorte 1 y cohorte 3 de GSE144744 con expresión diferencial estadísticamente significativa. *Resultados obtenidos para diferentes subtipos de linfocitos T (columna 1) divididos por el signo del valor logFC (columna 2, +: sobreexpresión en mujeres o infraexpresión en hombres; -: sobreexpresión en mujeres o infraexpresión en hombres). Se indican los genes significativos para la cohorte 1 (columna 3) y la cohorte 3 (columna 4) así como el número de genes presentes en la intersección (columna 5). Genes comunes asociados (columna 6): número de genes presentes en la intersección asociados a la enfermedad en la base de datos OpenTargets.*

Sin embargo, para los genes sobreexpresados en hombres no se ha identificado ninguna función sobrerrepresentada. Por su parte, los resultados obtenidos para **los linfocitos T CD8+** esclarecen que los genes presentes en la intersección sobreexpresados en mujeres están asociados a 5 procesos biológicos:

- *response to muscle stretch* (GO:0035994)
- *cellular response to chemical stress* (GO:0062197)
- *cellular response to oxidative stress* (GO:0034599)
- *response to oxidative stress* (GO:0006979)
- *response to virus* (GO:0009615)

De nuevo, los genes comunes sobreexpresados en hombres no presentan ninguna función que permita caracterizarlos de un modo estadísticamente significativo.

6. DISCUSIÓN

Cuando una persona accede al sistema de salud para realizar una consulta médica, los profesionales sanitarios deben evaluar sus manifestaciones y su historial clínico. Seguidamente, podrían requerir de diversas pruebas para confirmar un diagnóstico. Asimismo, se realizaría un seguimiento temporal para conocer la evolución de la enfermedad y, con el objetivo de mejorar su pronóstico, podría administrarse un tratamiento. Durante todo este proceso es esencial evaluar diversos biomarcadores, que proporcionen información relevante a los profesionales sanitarios para tomar las decisiones oportunas en el momento adecuado.

Sin embargo, este procedimiento no debe ser automatizado para cada enfermedad, sino que requiere considerar las características específicas de cada paciente. Esta concepción ha cobrado gran popularidad en los últimos años enmarcada bajo el término **medicina P4** (del inglés *Predictive, Preventive, Personalized and Participatory*)[103]. Para lograr este abordaje, son indispensables los estudios biomédicos que contemplen la variabilidad existente en la población afectada.

Concretamente, parte de las diferencias fisiopatológicas de multitud de enfermedades se establecen en base al **sexo** de los pacientes. Entre ellas, destacan el cáncer[104], las enfermedades neurodegenerativas[47], las enfermedades autoinmunitarias[46], y las enfermedades cardiovasculares[105]. Desafortunadamente, esta variabilidad no suele ser contemplada por la comunidad científica durante la investigación biomédica. En multitud de ocasiones la variable sexo no se recopila como metadato, mientras que en aquellos estudios que sí la presentan rara vez llega a ser evaluada. En último término, esta situación tiene consecuencias negativas en la salud de los pacientes; ya que la evaluación y el tratamiento de su enfermedad no se aborda con los biomarcadores más apropiados.

En el presente trabajo, se ha intentado mejorar este aspecto caracterizando a nivel molecular la **enfermedad de EM** con una perspectiva de sexo. Tal y como se ha descrito en el apartado de la introducción, la EM tiene un gran impacto en la sociedad, siendo la causa mayoritaria de discapacidad no asociada a accidentes en la población adulta joven[5]. Además, el conocimiento acerca de la misma es muy limitado. Actualmente su etiología es desconocida, y debido a la gran variabilidad de manifestaciones clínicas, su diagnóstico y pronóstico presenta gran complejidad. Asimismo, continúa siendo una enfermedad crónica, ya que los tratamientos actuales están destinados a paliar la severidad de los síntomas. Por todo ello, existe una gran **necesidad de investigación** sobre la enfermedad de EM.

Con este objetivo, se ha realizado un abordaje *in silico* utilizando datos depositados en repositorios públicos. Durante su procesamiento ha sido patente la falta de estandarización en la descripción, la nomenclatura, y el formato de los mismos. Por ello, además de incentivar que los datos generados en el proceso de investigación sean de **acceso público** (*Open science*), es de gran relevancia fomentar los **principios FAIR** (por su sigla en inglés *Findable, Accesible, Interoperable, Reusable*)[106]. Estas bases, establecidas por científicos de diferentes áreas, pretenden estandarizar el proceso para que los datos se depositen en los repositorios de forma homogénea. De este modo, otros

investigadores podrán encontrarlos con facilidad, para posteriormente ser reutilizados tras su descarga.

En concreto, los datos reutilizados proceden de las tecnologías scRNA-seq y snRNA-seq. El potencial de los datos generados a través de estas estrategias es inmenso; ya que permite obtener datos globales con un nivel de resolución de célula o núcleo único. Sin embargo, esta situación ocasiona que los **requerimientos computacionales** sean ingentes, debido al gran volumen de datos que es necesario almacenar y procesar. Concretamente, el análisis expuesto en este trabajo se ha realizado en la infraestructura computacional ubicada en el CIPF. Para poder ejecutar el código desarrollado, se han utilizado las colas del sistema *SLURM* que disponían de mayor capacidad de memoria y tiempo de ejecución. Incluso trabajando bajo estas condiciones, en ocasiones se ha necesitado reajustar las características de la ejecución de estos procesos, para asegurar la optimización del uso de recursos en el análisis de los datos de la cohorte 3 del estudio GSE144744.

También es relevante destacar la **falta de estandarización** en el análisis de este tipo de datos. En la actualidad existen numerosas herramientas bioinformáticas para su procesamiento. Disponer de una amplia cantidad de posibilidades presenta numerosas ventajas, y refleja el gran avance metodológico que se ha producido en los últimos años. Sin embargo, es necesario establecer guías que permitan abordar los análisis con robustez, al mismo tiempo que son consideradas las particularidades de cada estudio.

La implementación del código desarrollado se ha realizado ensalzando su **modularidad**[107]. De este modo, para cada paso del análisis se ha elaborado un *script* independiente; donde los resultados de cada procesamiento son los datos de entrada de los siguientes. Esta estrategia ha permitido testar diferentes abordajes para un mismo paso del análisis sin modificar la estructura del resto.

Como resultado del análisis bioinformático planteado, se ha determinado por el método *Schimer et al. 2019* que las **pacientes mujeres** de EM presentan un **mayor número de oligodendrocitos** respecto a los hombres afectados. Estudios en organismos modelo han establecido que, en estado de salud, este número es dependiente de las hormonas sexuales[108,109]. Debido a la relevancia de los oligodendrocitos en la EM, sería de gran interés profundizar en su investigación.

Por su parte, los dos métodos implementados al analizar los datos de la cohorte 3 del estudio GSE144744 establecen que el número de **linfocitos NK en sangre es mayor en hombres**. La implicación de este tipo celular en el transcurso de la enfermedad es controvertida; ya que hay estudios que reflejan tanto funciones protectoras como perjudiciales. Además, cuando se proporcionan tratamientos a los pacientes de EM, sus proporciones respecto al resto de tipos celulares suele modificarse[110]. En concreto, cuando se administra como tratamiento testosterona a pacientes hombres, la proporción de linfocitos NK aumenta[111]. Podría plantearse que este efecto hormonal estuviese involucrado en las diferencias encontradas en los individuos que sufren EM, ya que los hombres disponen de concentraciones mayores de testosterona que las mujeres.

En el trabajo no solamente se han analizado los datos procedentes de cada estudio individual, sino que se han **integrado** los resultados del análisis de expresión diferencial. Con ello, la información biológica extraída para la identificación de biomarcadores presenta robustez al proceder del **consenso** entre estudios y tipos celulares. A continuación, se disponen algunos de los genes y funciones biológicas que permiten caracterizar las diferencias de sexo en la enfermedad de EM.

En primer lugar, es relevante destacar que, tras realizar las intersecciones para los genes con expresión diferencial significativa entre todos los tipos celulares de un mismo estudio, la mayoría de los genes son específicos para cada tipo celular. Esta situación refleja la importancia de continuar investigando a nivel de célula y núcleo único; caracterizando de forma detallada las particularidades de cada tipo celular.

Atendiendo a la información consenso obtenida para el **sistema nervioso,** las mujeres que sufren EM presentan enriquecidas funciones relacionadas con la adhesión celular. Se ha descrito ampliamente que las células del sistema inmunitario de sangre periférica atraviesan la BHE para llegar hasta el tejido nervioso[5,6]. Conocer cómo las células del SNC interactúan entre ellas, así como con las células del sistema inmunitario infiltradas durante el transcurso de la enfermedad, aumentaría en gran medida el conocimiento sobre esta patología.

Por el contrario, los hombres presentan sobrerrepresentada la función *sequestring of actin monomers* (GO:0042989). Pese a que su interpretación biológica es complicada, se podría hipotetizar que podría estar relacionada con el sistema de "limpieza" de la actina extracelular (conocido en inglés como *extracellular actin scavenger system*). Este sistema está compuesto por un conjunto de proteínas encargadas de eliminar la actina del torrente sanguíneo procedente de tejidos dañados. Alteraciones en este sistema pueden ocasionar enfermedades como la hipogelsolinemia, la cual ha sido previamente descrita en pacientes de EM[112]. Además, entre las proteínas que componen al sistema se encuentra Dbp (por sus siglas en inglés *vitamin D-Binding Protein*). Tal y como indica su nombre, esta proteína es portadora de la vitamina D; micronutriente cuya deficiencia es uno de los principales factores de riesgo ambientales de la EM[113].

Es importante hacer énfasis en la sobreexpresión detectada en hombres respecto de mujeres del gen *LINGO1*, ya que la proteína codificada es un regulador negativo del proceso de mielinización de los axones neuronales. Por ello, se han propuesto antagonistas de esta proteína como tratamiento para favorecer el proceso de remielinización[114]. En base a los resultados obtenidos en este trabajo, sería esencial considerar la perspectiva de sexo en la evaluación de este tipo de compuestos farmacológicos.

Por otra parte, los resultados comunes para los tipos celulares del **sistema inmunitario** procedentes de sangre periférica también son de gran interés. En el caso de las mujeres, se identifica una sobreexpresión del gen *RASAL2* en el estudio cohorte 1 GSE144744. Su función principal es activar a la superfamilia de proteínas Ras, las cuales están involucradas en el proceso de autoinmunidad. En detalle, cuando se inhibe la expresión de algunas proteínas Ras en organismos modelo, estos sufren la enfermedad con sintomatología más leve. Además, se ha propuesto que la superfamilia Ras podría estar

implicada en la etiología de la EM[115]. En el caso de la cohorte 3 GSE144744 se ha identificado la sobreexpresión del gen *EGR1*; factor transcripcional que ya ha sido propuesto en la literatura científica como potencial biomarcador para esta patología[116].

En cuanto a los genes sobreexpresados en hombres destaca *HLA-A*, ya que el alelo *HLA-A*02* es uno de los factores de riesgo genéticos más consolidados[5,6]. Atendiendo a los procesos biológicos enriquecidos se encuentra la presentación de antígenos; función de interés central en las enfermedades autoinmunes y, en particular, en la EM[5,6]. Asimismo, se identifica la respuesta al interferón gamma. Esta citocina fue planteada como tratamiento para la EM hace más de tres décadas[117]. Pese a que se ha demostrado ampliamente su implicación en la enfermedad, su mecanismo molecular aún es desconocido, ya que presenta características tanto proinflamatorias como antiinflamatorias y actúa sobre multitud de tipos celulares[118].

Por último, los resultados obtenidos por tipo celular revelan diversas funciones biológicas sobrerrepresentadas en mujeres. En detalle, los linfocitos T CD4+ se caracterizan con numerosos eventos postranscripcionales como el procesamiento alternativo de las moléculas de ARNm. En concreto, se conoce que los pacientes de EM presentan variantes aberrantes de ARNm para genes que codifican proteínas implicadas en la señalización del sistema inmunitario[119].

Atendiendo a los linfocitos T CD8+, destacan los procesos celulares de respuesta a situaciones de estrés químico y oxidativo. Estos ambientes son comunes en las enfermedades neurodegenerativas y, en concreto, en la enfermedad de EM[120]. Por ello, caracterizar a nivel molecular cómo responden los tipos celulares ante las situaciones de estrés es esencial para abordar clínicamente la enfermedad. Cabe destacar que también se encuentra enriquecida en mujeres la función de respuesta a virus. Este resultado está en consonancia con los factores de riesgo ambientales de la EM; ya que uno de ellos es haber sufrido una infección vírica[5,6].

7. CONCLUSIONES

A continuación, se enumeran las conclusiones principales que pueden extraerse de este trabajo:

1. El resultado de la revisión sistemática ha puesto de manifiesto la necesidad de incluir suficiente tamaño muestral para poder analizar la perspectiva de sexo en los estudios biomédicos.

2. En base a la bibliografía revisada, este es el primer trabajo que evalúa las diferencias de sexo en la enfermedad de EM utilizando datos de scRNA-seq y snRNA-seq con un enfoque integrativo.

3. Se ha identificado un mayor número de oligodendrocitos (en el sistema nervioso) y un menor número de linfocitos NK (en sangre periférica) en mujeres enfermas respecto a hombres afectados.

4. Todos los tipos celulares evaluados presentan patrones de expresión diferencial en función del sexo del individuo que sufre la enfermedad.

5. En los tipos celulares de sistema nervioso, las mujeres presentan sobrerrepresentadas funciones involucradas en la adhesión celular. Por su parte, un proceso biológico relacionado con la actina se encuentra enriquecido en hombres. Asimismo, los hombres disponen de genes diferencialmente expresados descritos previamente en la bibliografía, como *LINGO1*.

6. Los linfocitos T CD4+ de las mujeres presentan sobrerrepresentadas funciones implicadas en el procesamiento de moléculas de ARNm, mientras que los linfocitos T CD8+ en procesos biológicos relacionados con la respuesta a estrés. No se han encontrado funciones significativas para estos tipos celulares en hombres, pero al evaluar el conjunto de células inmunitarias se ha identificado el proceso de presentación de antígenos (en base a los genes *HLA-A, HLA-DRB5*).

8. PERSPECTIVAS FUTURAS

El abordaje planteado es el inicio de una línea de investigación que puede extenderse mediante múltiples vertientes. En particular, el análisis podría revisarse con expertos de la enfermedad de EM. Asimismo, podrían utilizarse estrategias alternativas, como el metaanálisis, para la integración de los resultados del análisis primario. De hecho, se podría profundizar en la interpretación de los resultados obtenidos por tipo celular. Por ejemplo, evaluando otras perspectivas de la caracterización funcional como son las rutas de señalización.

Otra posibilidad a contemplar en el futuro es realizar de nuevo la revisión sistemática para ampliar el número de estudios evaluados. De este modo se podría considerar una mayor potencia estadística para describir las diferencias de sexo en la EM, tanto en el sistema nervioso como inmunitario. Incluso se podría alcanzar mayor nivel de detalle considerando de forma independiente cada subtipo de la enfermedad.

Por último, es importante considerar el refinamiento del código implementado para optimizar los requerimientos de memoria y del tiempo de ejecución utilizados.

9. AGRADECIMIENTOS

En primer lugar, me gustaría darle las gracias a Francisco García, director de la unidad de Bioinformática y Bioestadística del CIPF, por brindarme la oportunidad de realizar este trabajo.

Junto a él, a todos los componentes de la unidad, especialmente a Zoraida Andreu, María de la Iglesia, José Francisco Català y Adolfo López. Muchas gracias por ayudarme técnicamente siempre que lo he necesitado, por todas las horas dedicadas a este trabajo, y, sobre todo, por vuestra calidad humana.

Por supuesto, no me olvido de Almudena. Amiga, gracias por ser una gran compañera de viaje en esta aventura.

10. BIBLIOGRAFÍA

1) Delves, P. Roitt - Inmunología: fundamentos. 12º ed. *Editorial Médica Panamericana* (2014).

2) Theofilopoulos, A. N., Kono, D. H. & Baccala, R. The Multiple Pathways to Autoimmunity. *Nat. Immunol.* **18**, 716–724 (2017).

3) Gutierrez-Arcelus, M., Rich, S. S. & Raychaudhuri, S. Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. *Nat. Rev. Genet.* **17**, 160–174 (2016).

4) Aguilar, B., *et al*. Libro blanco de la esclerosis múltiple en España 2020. Resumen ejecutivo. https://www.conlaem.es/sites/default/files/Libro-Blanco-EM-2020_resumen-ejecutivo.pdf (visitada el 20/05/2021).

5) Dobson, R. & Giovannoni, G. Multiple sclerosis - a review. *Eur. J. Neurol.* **26**, 27–40 (2019).

6) Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B. & Ciccarelli, O. Multiple sclerosis. *The Lancet* **391**, 1622–1636 (2018).

7) Antontseva, E., Bondar, N., Reshetnikov, V. & Merkulova, T. The Effects of Chronic Stress on Brain Myelination in Humans and in Various Rodent Models. *Neuroscience* **441**, 226–238 (2020).

8) Brownlee, W. J., Hardy, T. A., Fazekas, F. & Miller, D. H. Diagnosis of multiple sclerosis: progress and challenges. *Lancet Lond. Engl.* **389**, 1336–1346 (2017).

9) The Multiple Sclerosis International Federation. Atlas of MS, 3rd Edition (September 2020). https://www.msif.org/wp-content/uploads/2020/10/Atlas-Epidemiology-report-Sept-2020-Final-ES.pdf (visitada el 21/05/2021).

10) Perez-Carmona, N., Fernandez-Jover, E. & Sempere, A. P. Epidemiology of multiple sclerosis in Spain. *Rev. Neurol.* **69**, 32–38 (2019).

11) Olsson, T., Barcellos, L. F. & Alfredsson, L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nat. Rev. Neurol.* **13**, 25–36 (2017).

12) Waubant, E. *et al.* Environmental and genetic risk factors for MS: an integrated review. *Ann. Clin. Transl. Neurol.* **6**, 1905–1922 (2019).

13) Voskuhl, R. R., Sawalha, A. H. & Itoh, Y. Sex chromosome contributions to sex differences in multiple sclerosis susceptibility and progression. *Mult. Scler. J.* **24**, 22–31 (2018).

14) Alfredsson, L. & Olsson, T. Lifestyle and Environmental Factors in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **9**, a028944 (2019).

15) Lublin, F. D. *et al.* Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology* **83**, 278–286 (2014).

16) McGinley, M. P., Goldschmidt, C. H. & Rae-Grant, A. D. Diagnosis and Treatment of Multiple Sclerosis: A Review. *JAMA* **325**, 765–779 (2021).

17) Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018).

18) Derfuss, T. *et al.* Advances in oral immunomodulating therapies in relapsing multiple sclerosis. *Lancet Neurol.* **19**, 336–347 (2020).

19) Mansilla, M. J. *et al.* Paving the way towards an effective treatment for multiple sclerosis: advances in cell therapy. *Cell. Mol. Immunol.* **18**, 1353–1374 (2021).

20) DeLuca, J., Chiaravalloti, N. D. & Sandroff, B. M. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. *Nat. Rev. Neurol.* **16**, 319–332 (2020).

21) Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., LaMantia, A. S. & White, L. Neuroscience. 5th ed. *Sinauer Assoc*. (2012).

22) Bar-Or, A. & Li, R. Cellular immunology of relapsing multiple sclerosis: interactions, checks, and balances. *Lancet Neurol.* **20**, 470–483 (2021).

23) Wang, J. *et al.* HLA-DR15 Molecules Jointly Shape an Autoreactive T Cell Repertoire in Multiple Sclerosis. *Cell* **183**, 1264-1281.e20 (2020).

24) Dendrou, C. A., Fugger, L. & Friese, M. A. Immunopathology of multiple sclerosis. *Nat. Rev. Immunol.* **15**, 545–558 (2015).

25) Cortini, A., Bembich, S., Marson, L., Cocco, E. & Edomi, P. Identification of novel non-myelin biomarkers in multiple sclerosis using an improved phage-display approach. *PloS One* **14**, e0226162 (2019).

26) Baecher-Allan, C., Kaskow, B. J. & Weiner, H. L. Multiple Sclerosis: Mechanisms and Immunotherapy. *Neuron* **97**, 742–768 (2018).

27) Matthews, P. M. Chronic inflammation in multiple sclerosis — seeing what was always there. *Nat. Rev. Neurol.* **15**, 582–593 (2019).

28) Neves, S. P. das, Sousa, J. C., Sousa, N., Cerqueira, J. J. & Marques, F. Altered astrocytic function in experimental neuroinflammation and multiple sclerosis. *Glia* **69**, 1341–1368 (2021).

29) Raja, K. *et al.* A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. *Int. J. Genomics* **2017**, 6213474 (2017).

30) Tebani, A., Afonso, C., Marret, S. & Bekri, S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. *Int. J. Mol. Sci.* **17**, E1555 (2016).

31) Schneider, M. V. & Orchard, S. Omics Technologies, Data and Bioinformatics Principles. in *Bioinformatics for Omics Data: Methods and Protocols* **719**, 3–30 (2011).

32) Bedia, C. Chapter Two - Experimental Approaches in Omic Sciences. in *Comprehensive Analytical Chemistry* **82**, 13–36 (2018).

33) Nelson, D. L. & Cox, M. M. Lehninger – Principios de bioquímica. 6º ed. *Omega* (2015).

34) Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

35) Virkud, Y. V., Kelly, R. S., Wood, C. & Lasky-Su, J. A. The nuts and bolts of omics for the clinical allergist. *Ann. Allergy Asthma Immunol. Off. Publ. Am. Coll. Allergy Asthma Immunol.* **123**, 558–563 (2019).

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

37) Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

38) Adilijiang, A. *et al.* Next Generation Sequencing-Based Transcriptome Predicts Bevacizumab Efficacy in Combination with Temozolomide in Glioblastoma. *Mol. Basel Switz.* **24**, E3046 (2019).

39) Cha, J. & Lee, I. Single-cell network biology for resolving cellular heterogeneity in human diseases. *Exp. Mol. Med.* **52**, 1798–1808 (2020).

40) Wen, L. & Tang, F. Boosting the power of single-cell analysis. *Nat. Biotechnol.* **36**, 408–409 (2018).

41) Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).

42) Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).

43) Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

44) Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PloS One* **13**, e0209648 (2018).

45) Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

46) Ortona, E. *et al.* Sex-based differences in autoimmune diseases. *Ann. Ist. Super. Sanita* **52**, 205–212 (2016).

47) Yanguas-Casás, N. Sex differences in neurodegenerative diseases. *SM J Neurol Disord Stroke* **1014**, 3 (2017).

48) Zagni, E., Simoni, L. & Colombo, D. Sex and Gender Differences in Central Nervous System-Related Disorders. *Neurosci. J.* **2016**, 2827090 (2016).

49) Golden, L. C. & Voskuhl, R. The importance of studying sex differences in disease: The example of multiple sclerosis. *J. Neurosci. Res.* **95**, 633–643 (2017).

50) Voskuhl, R. R. The Effect Of Sex on Multiple Sclerosis Risk And Disease Progression. *Mult. Scler. Houndmills Basingstoke Engl.* **26**, 554–560 (2020).

51) Roeder, H. J. & Leira, E. C. Effects of the Menstrual Cycle on Neurological Disorders. *Curr. Neurol. Neurosci. Rep.* **21**, 34 (2021).

52) Gilli, F., DiSano, K. D. & Pachner, A. R. SeXX Matters in Multiple Sclerosis. *Front. Neurol.* **11**, 616 (2020).

53) R core team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2020).

54) Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M. & Altman, D. G. Epidemiology and Reporting Characteristics of Systematic Reviews. *PLOS Med.* **4**, e78 (2007).

55) Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, T. P. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Med.* **6**, e1000097 (2009).

56) Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

57) Athar, A. *et al.* ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).

58) Speir, M. L. *et al.* UCSC Cell Browser: Visualize Your Single-Cell Data. *bioRxiv* 2020.10.30.361162 (2020).

59) Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

60) Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

61) McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinforma. Oxf. Engl.* **33**, 1179–1186 (2017).

62) Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).

63) Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).

64) Ballesteros-González, J. A. Single cell RNA-sequencing analysis in Alzheimer's disease. *Master's thesis, Universitat Oberta de Catalunya* (2020).

65) Hicks, S. Analysis of single cell RNA-seq data: 2018 BioInfoSummer Workshop. https://www.stephaniehicks.com/2018-bioinfosummer-scrnaseq/ (visitada el 2/06/2021).

66) Germain, P. scDblFinder: scDblFinder. R package version 1.6.0, https://github.com/plger/scDblFinder (2021).

67) Bruford, E. A. *et al.* Guidelines for Human Gene Nomenclature. *Nat. Genet.* **52**, 754–758 (2020).

68) Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods San Diego Calif* **85**, 54–61 (2015).

69) L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

70) Blighe, K. & Lun, A. PCAtools: PCAtools: Everything Principal Components Analysis. R package version 2.4.0, https://github.com/kevinblighe/PCAtools (2021).

71) Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

72) CSC. Single cell RNA-seq data analysis with R. https://github.com/NBISweden/excelerate-scRNAseq (2019).

73) McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

74) Csardi, G. & Nepusz, T. The igraph software package for complex network research. 10.

75) Lun, A. bluster: Clustering Algorithms for Bioconductor. R package version 1.2.1. (2021).

76) McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.* **8**, 8868 (2018).

77) Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

78) Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

79) Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).

80) Schirmer, L. *et al.* Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).

81) R core team. R: A language and environment for statistical computing: stats package. *R Foundation for Statistical Computing, Vienna, Austria* (2020).

82) Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

83) Dusa, A. venn: Draw Venn Diagrams. R package version 1.10. https://CRAN.R-proyect.org/package=venn (2021).

84) Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

85) Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580 (2004).

86) Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).

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

88) Kaufmann, M. *et al.* Identifying CNS-colonizing T cells as potential therapeutic targets to prevent progression of multiple sclerosis. *Med* **2**, 296-312.e8 (2021).

89) McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4 (2019).

90) Sun, W. *et al.* SOX9 Is an Astrocyte-Specific Nuclear Marker in the Adult Brain Outside the Neurogenic Regions. *J. Neurosci.* **37**, 4493–4507 (2017).

91) Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database J. Biol. Databases Curation* **2019**, baz046 (2019).

92) Gerrits, E., Heng, Y., Boddeke, E. W. G. M. & Eggen, B. J. L. Transcriptional profiling of microglia; current state of the art and future perspectives. *Glia* **68**, 740–755 (2020).

93) Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).

94) Palmer, C., Diehn, M., Alizadeh, A. A. & Brown, P. O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 115 (2006).

95) Kanduri, K. *et al.* Identification of global regulators of T-helper cell lineage specification. *Genome Med.* **7**, 122 (2015).

96) Crinier, A. *et al.* High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity* **49**, 971-986.e5 (2018).

97) Gren, S. T. *et al.* A Single-Cell Gene-Expression Profile Reveals Inter-Cellular Heterogeneity within Human Monocyte Subsets. *PLOS ONE* **10**, e0144351 (2015).

98) Collin, M. & Bigley, V. Human dendritic cell subsets: an update. *Immunology* **154**, 3–20 (2018).

99) Rørvig, S. *et al.* Ficolin-1 is present in a highly mobilizable subset of human neutrophil granules and associates with the cell surface after stimulation with fMLP. *J. Leukoc. Biol.* **86**, 1439–1449 (2009).

100) Shaath, H., Vishnubalaji, R., Elkord, E. & Alajez, N. M. Single-Cell Transcriptome Analysis Highlights a Role for Neutrophils and Inflammatory Macrophages in the Pathogenesis of Severe COVID-19. *Cells* **9**, E2374 (2020).

101) Zhu, Y. *et al.* Characterization and generation of human definitive multipotent hematopoietic stem/progenitor cells. *Cell Discov.* **6**, 1–18 (2020).

102) Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* **6**, e21800 (2011).

103) Flores, M., Glusman, G., Brogaard, K., Price, N. D. & Hood, L. P4 medicine: how systems medicine will transform the healthcare sector and society. *Pers. Med.* **10**, 565–576 (2013).

104) Cook, M. B., McGlynn, K. A., Devesa, S. S., Freedman, N. D. & Anderson, W. F. Sex disparities in cancer mortality and survival. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **20**, 1629–1637 (2011).

105) Spence, J. D. & Pilote, L. Importance of sex and gender in atherosclerosis and cardiovascular disease. *Atherosclerosis* **241**, 208–210 (2015).

106) Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

107) Kutmon, M., van Iersel, M., Kelder, T. & Evelo, C. The Importance of Modularity in Bioinformatics Tools. *Nat. Preced.* (2011).

108) Swamydas, M., Bessert, D. & Skoff, R. Sexual Dimorphism of Oligodendrocytes is Mediated by Differential Regulation of Signaling Pathways. *J. Neurosci. Res.* **87**, 3306–3319 (2009).

109) Cerghet, M. *et al.* Proliferation and death of oligodendrocytes and myelin proteins are differentially regulated in male and female rodents. *J. Neurosci. Off. J. Soc. Neurosci.* **26**, 1439–1447 (2006).

110) Mimpen, M., Smolders, J., Hupperts, R. & Damoiseaux, J. Natural killer cells in multiple sclerosis: A review. *Immunol. Lett.* **222**, 1–11 (2020).

111) Gold, S. M., Chalifoux, S., Giesser, B. S. & Voskuhl, R. R. Immune modulation and increased neurotrophic factor production in multiple sclerosis patients treated with testosterone. *J. Neuroinflammation* **5**, 32 (2008).

112) Kułakowska, A. *et al.* Hypogelsolinemia, a disorder of the extracellular actin scavenger system, in patients with multiple sclerosis. *BMC Neurol.* **10**, 107 (2010).

113) Gauzzi, M. C. Vitamin D-binding protein and multiple sclerosis: Evidence, controversies, and needs. *Mult. Scler. Houndmills Basingstoke Engl.* **24**, 1526–1535 (2018).

114) Rudick, R. A., Mi, S. & Sandrock, A. W. LINGO-1 antagonists as therapy for multiple sclerosis: in vitro and in vivo evidence. *Expert Opin. Biol. Ther.* **8**, 1561–1570 (2008).

115) Messina, S. Small GTPase RAS in multiple sclerosis - exploring the role of RAS GTPase in the etiology of multiple sclerosis. *Small GTPases* **11**, 312–319 (2020).

116) Islam, T. *et al.* Detection of multiple sclerosis using blood and brain cells transcript profiles: Insights from comprehensive bioinformatics approach. *Inform. Med. Unlocked* **16**, 100201 (2019).

117) Panitch, H. S., Hirsch, R. L., Schindler, J. & Johnson, K. P. Treatment of multiple sclerosis with gamma interferon: exacerbations associated with activation of the immune system. *Neurology* **37**, 1097–1102 (1987).

118) Arellano, G., Ottum, P. A., Reyes, L. I., Burgos, P. I. & Naves, R. Stage-Specific Role of Interferon-Gamma in Experimental Autoimmune Encephalomyelitis and Multiple Sclerosis. *Front. Immunol.* **6**, 492 (2015).

119) Hecker, M. *et al.* Aberrant expression of alternative splicing variants in multiple sclerosis - A systematic review. *Autoimmun. Rev.* **18**, 721–732 (2019).

120) Ohl, K., Tenbrock, K. & Kipp, M. Oxidative stress in multiple sclerosis: Central and peripheral mode of action. *Exp. Neurol.* **277**, 58–67 (2016).