VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO FIN DE MÁSTER

CARACTERIZACIÓN DE LAS DIFERENCIAS DE SEXO EN LA ENFERMEDAD DE ALZHEIMER, MEDIANTE ESTRATEGIAS BASADAS EN EL ANÁLISIS MASIVO DE DATOS DE CÉLULA ÚNICA

AUTORA:

ALMUDENA NEVA ALEJO

TUTORES:

MARÍA PASCUAL MORA
FRANCISCO GARCÍA GARCÍA
MARÍA DE LA IGLESIA VAYÁ

SEPTIEMBRE, 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 FIN DE MÁSTER

CARACTERIZACIÓN DE LAS DIFERENCIAS DE SEXO EN LA ENFERMEDAD DE ALZHEIMER, MEDIANTE ESTRATEGIAS BASADAS EN EL ANÁLISIS MASIVO DE DATOS DE CÉLULA ÚNICA

AUTORA:

ALMUDENA NEVA ALEJO

TUTORES:

MARÍA PASCUAL MORA
FRANCISCO GARCÍA GARCÍA
MARÍA DE LA IGLESIA VAYÁ

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 Enfermedad de Alzheimer (EA) es la principal causa de demencia. Los síntomas clínicos de la EA empiezan con deterioros de la memoria a corto plazo y progresivamente se van mermando importantes funciones cognitivas y conductuales, conllevando finalmente a la muerte del paciente. El creciente número de pacientes, así como la falta de tratamientos eficientes y de métodos de diagnóstico tempranos y precisos, resalta el interés de la investigación en este campo. Al haber sido detectadas diferencias de sexo en la prevalencia, gravedad y progresión de la EA y de muchas otras enfermedades, la perspectiva de sexo en la investigación biomédica puede contribuir a una mejor y más rápida comprensión de las enfermedades, siendo un paso fundamental hacia la medicina personalizada. En este estudio se han explorado las diferencias moleculares relacionadas con el sexo en la EA mediante un análisis integrativo de estudios de transcriptómica de núcleo único de muestras postmortem de cortex cerebral. En concreto, se ha centrado en las diferencias de sexo en microglía y astrocitos por su importante papel en la neuroinflamación en la EA. Con este propósito se llevó a cabo una revisión sistemática siguiendo las directrices PRISMA (en inglés, Preferred Reporting Items for Systematic Reviews and Meta-Analyses) que permitió la selección de 3 estudios. A continuación, se realizaron el análisis de expresión diferencial por tipo celular y el análisis de abundancia celular diferencial, en función del sexo y la condición experimental, para cada estudio por separado. Finalmente, se integraron los resultados y se interpretaron biológicamente. Los resultados obtenidos no muestran diferencias de sexo en la composición celular de pacientes con EA. Sin embargo, este trabajo ha permitido la caracterización de diferencias transcriptómicas en microglía y astrocitos específicas y no específicas de sexo en la EA. Se detectaron 6 genes diferencialmente expresados en función del sexo en microglía y 59 en astrocitos. También se detectó 1 gen diferencialmente expresado no específico de sexo en microglía y 20 en astrocitos. De los genes detectados, algunos confirman las conclusiones de otros estudios, mientras que otros proporcionan nuevas líneas de investigación que podrían ayudar a comprender esta compleja enfermedad, así como mejorar el tratamiento y diagnóstico de la misma.

Palabras clave: Enfermedad de Alzheimer, transcriptómica de célula única, transcriptómica de núcleo único, perspectiva de sexo, microglía y astrocitos.

Abstract

Alzheimer's disease (AD) is the leading cause of dementia. The clinical symptoms of AD begin with short-term memory impairment and gradually important cognitive and conductual functions are impaired, finally driving to death. The growing number of patients, in addition to the lack of efficient treatments and precise and early diagnostic methods, highlights the interest of the investigation in this field. Since sex differences in prevalence, severity and progression of AD and many other diseases have been reported, sex perspective in biomedical research may lead for a better and faster understanding of diseases, being a fundamental step towards personalized medicine. We have explored the sex-related molecular differences in AD by an integrative analysis of single nucleus transcriptomics studies from postmortem cortical brain samples. In particular, we have focused on sex differences in astrocytes and microglia for its important role in neuroinflammation in AD. For this purpose, we have performed a systematic review following the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) that afforded the selection of three studies. Then, we conducted a differential gene expression analysis by cell type and a differential cell abundance analysis, by sex and experimental condition for each study individually. Finally, we integrated all the results and interpreted them biologically. The results don't show sex differences in cellular composition in AD patients. However, this study has allowed the characterization of sex-specific anf non-sex-specific transcriptomic differences in microglia and astrocytes in AD. The analysis detected 6 sex-specific differentially expressed genes in microglia and 59 in astrocytes. Also, 1 non-sex-specific differentially expressed gene was detected in microglia and 20 in astrocytes. Some detected genes confirm the conclusions of others studies, while others provide new research lines that could help to understand this complex disease, as well as improve its treatment and diagnosis.

Keywords: Alzheimer's disease, single cell transcriptomics, single nucleus transcriptomics, sex perspective, microglia and astrocytes.

Índice

Índice de figuras

Índice de tablas

Glosario de abreviaturas

Aβ: Péptidos β-amiloides

ADNc: ADN complementario

APP: Amyloid Precursor Protein, proteína precursora amiloidea

ARN: Ácido ribonucleico

ARNm: Ácido ribonucleico mensajero

CIPF: Centro de Investigación Príncipe Felipe

CPU: Central Processing Unit, unidad central de procesamiento

DCL: Deterioro cognitivo leve

EA: Enfermedad de Alzheimer (AD, en inglés)

EOAD: Early-onset Alzheimer's disease, Alzheimer familiar

FDR: False Discovery Rate, tasa de falsos rechazados

GEO: Gene Expression Omnibus, recopilación de expresiones génicas

GO: Gene Ontology, ontología génica

LOAD: Late-onset Alzheimer's disease, Alzheimer esporádico

logFC: Fold Change logarithm, logaritmo de la tasa de cambio

MRI: Magnetic Resonance Imaging, imagen por resonancia magnética

NB GLM: Negative Binomial generalized linear model, modelo lineal generalizado binomial negativo

NTFs: Neurofibrillary tangles, marañas neurofibrilares

OPCs: Oligodendrocyte precursor cells, células precursoras de oligodendrocitos

PCA: Principal Component Analysis, análisis de componentes principales

PET: Positron Emission Tomography, tomografía por emisión de positrones

PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses, ítems preferentes a informar en las revisiones sistemáticas y los metaanálisis

RAM: Random Access Memory, memoria de acceso aleatorio

RNA-seq: Ribonucleic acid sequencing, secuenciación de ácido ribonucleico

QL: Quasi-likelihood, cuasiverosimilitud

scRNA-seq: Single cell Ribonucleic acid Sequencing, secuenciación de ácido ribonucleico de célula única

SLURM: Simple Linux Utility for Resource Management, utilidad básica de Linux para la gestión de recursos

SNC: Sistema nervioso central

snRNA-seq: Single nucleus Ribonucleic acid Sequencing, secuenciación de ácido ribonucleico de núcleo único

TMM: Trimmed Mean of M-values, media ajustada de M-valores

UMI: Unique Molecular Identifier, identificador molecular único

1. Introducción

1.1 Importancia de la perspectiva de sexo en biomedicina

Tradicionalmente en numerosos estudios biomédicos no se ha considerado la **perspectiva de sexo**, sin embargo, son muchas las enfermedades en las que se han observado diferencias de sexo en la epidemiología, patofisiología, manifestaciones clínicas, progresión de la enfermedad y respuesta a tratamiento. Estas diferencias de sexo pueden estar asociadas a varios aspectos, como son las **diferencias genéticas** existentes entre mujeres y hombres. El sexo biológico viene determinado por los cromosomas sexuales: XX en mujeres y XY en hombres. El cromosoma Y contiene genes específicos del sexo masculino y genes que presentan diferencias funcionales sutiles respecto a sus homólogos en el cromosoma X. Por otra parte, en las células femeninas se inactiva aleatoriamente uno de los dos cromosomas X, para compensar los niveles de expresión de los genes de dicho cromosoma con respecto a los de las células masculinas. Sin embargo, algunos de los genes del cromosoma X no se ven afectados por esta inactivación. Como consecuencia, se producen diferencias en la expresión génica entre sexos, ya que las mujeres presentarían activas dos copias del mismo gen mientras que los hombres solo una. Uno de los genes que posiblemente origina mayores diferencias entre hombres y mujeres es el gen SRY, específico del cromosoma Y, responsable del desarrollo de los testículos, y por tanto, de la producción de testosterona testicular. Esta hormona produce durante el desarrollo cambios epigenéticos que alteran la expresión génica y la estructura de tejidos en muchos órganos. Por último, cabe destacar la influencias de las **diferencias hormonales** asociadas al sexo. Todas estas diferencias remarcan la importancia de considerar el sexo en la investigación biomédica, para asegurar su inclusión en el diagnóstico y el tratamiento de enfermedades, siendo este un paso fundamental hacia la **medicina personalizada**[1].

1.2 Enfermedades neurodegenerativas

Las enfermedades neurodegenerativas son una causa común y creciente de mortalidad y morbilidad a nivel mundial, particularmente en la población envejecida. Pueden definirse como trastornos caracterizados por la pérdida progresiva de neuronas asociada a la deposición de proteínas en el cerebro y órganos periféricos, lo que se traduce en la pérdida de funciones cognitivas y motoras[2–4]. Estas enfermedades son muy heterogéneas, además el solapamiento y combinación de las mismas es frecuente, por lo que su diagnóstico es complejo, siendo la evaluación neuropatológica en la autopsia, el diagnóstico más fiable en la actualidad[3,5].

Se han detectado diferencias de sexo en la incidencia, gravedad y progresión de muchas de estas enfermedades. El desarrollo del cerebro, así como la estructura del cerebro adulto, su función y bioquímica es distinto en hombres y mujeres. El estudio del **dimorfismo sexual del cerebro** es clave para entender la importancia del sexo en la progresión y el tratamiento de las enfermedades neurodegenerativas. Concretamente, en este trabajo nos hemos centrado en las diferencias de sexo en la **enfermedad de Alzheimer** (EA), la enfermedad neurodegenerativa más común y devastadora[4,6].

1.3 Enfermedad de Alzheimer

La enfermedad de Alzheimer (EA) es la causa más común de demencia, de hecho se denomina demencia senil de tipo Alzheimer. Se manifiesta en sus inicios como un deterioro de la memoria a corto plazo y progresivamente va provocando el mermado de importantes funciones cognitivas y conductuales hasta que finalmente causa la muerte del paciente. Se estima que en 2050 más de 50 millones de personas la padecerán. A pesar de los esfuerzos empleados en descifrar los enclaves de esta enfermedad desde que Alois Alzheimer describió el primer caso en 1907, sigue **sin** haber un **tratamiento efectivo** ni un **método de diagnóstico preciso** para la enfermedad[6–9]. Al igual que en otras enfermedades se han detectado diferencias de sexo en la prevalencia, gravedad y progresión de la EA. El estudio de la perspectiva de sexo en la EA podría contribuir a la mejora y avance del tratamiento y diagnóstico de esta enfermedad, ya que, considerar las aportaciones de las diferencias de sexo en las enfermedades representa un eslabón importante hacia la **medicina personalizada**. En primer lugar, **el número de mujeres con EA es mayor** que el de hombres, probablemente por su mayor esperanza de vida, ya que el envejecimiento es el principal factor de riesgo de esta enfermedad. Por otra parte, **en hombres, la progresión de la enfermedad suele ser más rápida y agresiva**, ocasionando una muerte más temprana[6,10].

1.3.1 Patología

Las principales características de la EA son la presencia de depósitos extracelulares de **péptidos β-amiloides** (Aβ), conocidos como **placas amiloides** y la acumulación intracelular de agregados de **proteína tau hiperfosforilada** (**marañas neurofibrilares**, NTFs)[11], que conllevan a la neurodegeneración y a la atrofia cerebral (Figura 1A). Las regiones principalmente afectadas son el lóbulo temporal y parietal, partes de la corteza frontal y la circunvolución cingulada. Existen muchas hipótesis (hipótesis colinérgica, hipótesis de la cascada amiloide, hipótesis sináptica, hipótesis tau, etc.), que tratan de explicar la etiología de esta enfermedad, que todavía es un tema de intenso debate. La hipótesis predominante de la EA, **hipótesis de la cascada amiloide**, propone que la acumulación de péptidos β-amiloides es el proceso patológico central, siendo la hiperfosforilación de la proteína tau, la inflamación, el estrés oxidativo y la excitotoxicidad, procesos secundarios. Los péptidos β-amiloides se producen en la proteolisis de la proteína transmembrana precursora amiloidea (**APP**) por la acción de las enzimas β y γ secretasa (**ruta amiloidogénica**) y son secretados al líquido intersticial (Figura 1B). En los individuos sanos el exceso de Aβ es eliminado del cerebro, pero en los enfermos estos péptidos se agregan formando oligómeros neurotóxicos, fibrillas y placas amiloides. Esta hipótesis se ve reforzada por la genética, ya que las mutaciones asociadas actualmente a la EA están involucradas en la generación, procesamiento o eliminación del β-amiloide[7,8,12,13].

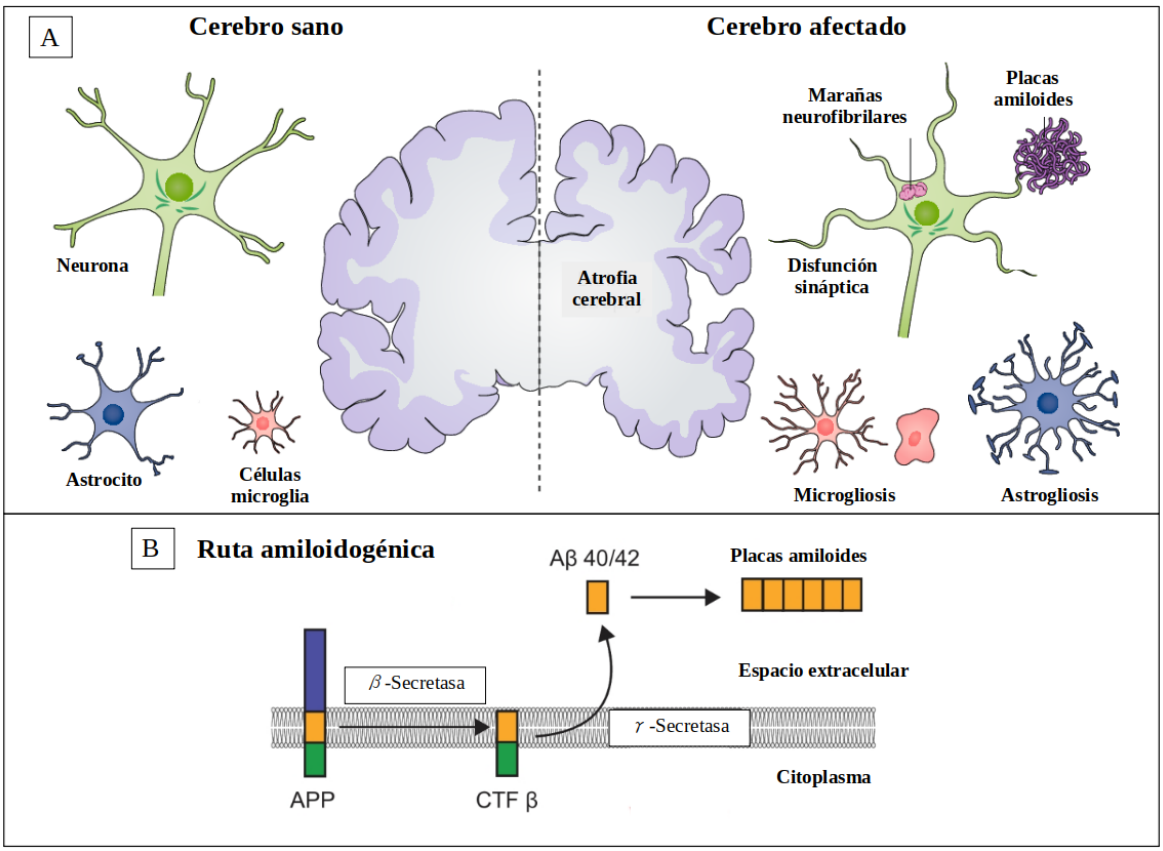

Figura 1: Patología de la EA. A) Principales características de la enfermedad de Alzheimer. La enfermedad de Alzheimer se caracteriza por la acumulación de placas amiloides extracelulares y marañas neurofibrilares intracelulares y por la neuroinflamación causada por una respuesta inmune prolongada. Todo esto conlleva a la neurodegeneración y a la atrofia cerebral. B) Ruta amiloidogénica de proteolisis de la proteína precursora amiloidea. Participan dos enzimas, la β-secretasa que genera el producto secretable sAPPβ y la γ-secretasa que genera el producto secretable β-amiloide. No se sabe mucho sobre la regulación de esta ruta, pero es evidente su gran importancia en el desarrollo del Alzheimer[7]. Imagen creada a partir de las imágenes obtenidas en Zhou, Y et al.(2018)[14] y Congdon, E (2018)[15].

Sin embargo, aunque la acumulación de Aβ sea necesaria para el diagnóstico de la EA, no es suficiente, se requiere también la acumulación de tau. Tau es una proteína soluble asociada a microtúbulo cuya función es la estabilización de los microtúbulos neuronales. En la EA se hiperfosforila, se desprende del microtúbulo y se agrega originando NTFs. La acumulación de los NTFs en las neuronas conlleva a la pérdida de la función y a la muerte celular[7,8,12,13].

Por último, cabe destacar el papel de la neuroinflamación en la EA, mediada por la activación de las células gliales, concretamente las células de la **microglía** y los **astrocitos**, y la liberación de citoquinas. La microglía son los macrófagos del sistema nervioso central (SNC), encargados de la detección de patógenos y de la fagocitosis de los mismos, así como de restos celulares y células en degeneración. Además, participan en la presentación de antígenos a las células T, coordinando así el sistema inmune innato y adaptativo. Los astrocitos son los subtipos de células gliales más abundantes del SNC, y llevan a cabo varias funciones involucradas en la homeostasis de iones, la transmisión de neurotransmisores, la secreción de factores de crecimiento y la regulación del estrés oxidativo. Además, participan en el mantenimiento y permeabilidad de la barrera

hematoencefálica. La respuesta inflamatoria mediada por las células gliales que se observa en la EA no se sabe si es consecuencia o causa de la neurodegeneración, pero cada vez existen más evidencias que indican que es uno de los principales contribuyentes de los procesos neurodegenerativos y los déficits cognitivos de la EA. Además, se ha visto que es un arma de doble filo, desempeñando tanto funciones beneficiosas como perjudiciales[16]. El estudio de la microglía y los astrocitos puede ayudar a dilucidar el papel de la neuroinflamación en la EA y al descubrimiento de nuevas estrategias terapéuticas.

1.3.2 Genética de la enfermedad

Podemos clasificar la EA en dos categorías, Alzheimer esporádico o de inicio tardío (LOAD) y Alzheimer familiar o de inicio temprano (EOAD). La gran mayoría de casos de Alzheimer pertenecen a la primera categoría, en la que la genética es un factor influyente pero no determinante, ya que depende también de distintos factores ambientales. El Alzheimer de tipo esporádico suele aparecer en edades más tardías y su progresión es más lenta. En este tipo de Alzheimer, a nivel genético, el mayor factor de riesgo es ser portador de la variante **APOE-e4**[7,8,17]. La frecuencia de la variante APOE-e4 no difiere por sexo, sin embargo, el **riesgo** de **EA en mujeres** de entre 65 y 75 años portadoras de la misma es 4 veces **mayor** que en hombres[1]. El gen APOE codifica para la apolipoproteina E (APOE) que está implicada en el transporte y metabolismo de lípidos[7]. Se ha demostrado que APOE afecta a la EA debido a su función inmunomoduladora. Esta función de APOE está ligada al gen **TREM2**, expresado por la microglía en el SNC, ya que se ha demostrado que TREM2 es un receptor de APOE, de manera que la lipoproteína APOE y su unión al receptor de la microglía TREM2 es clave para desencadenar el proceso de fagocitosis microglial. TREM2 constituye el segundo factor de riesgo de la LOAD. Su variante más común R47H incrementa de dos a tres veces el riesgo de padecer la enfermedad[18].

Respecto al Alzheimer familiar, aunque a veces está causado por una mutación genética autosómica dominante, es considerada una enfermedad poligénica. Los genes cuyas mutaciones se han asociado a esta enfermedad son **APP**, **PSEN1** y **PSEN2**, todos relacionados con la ruta amiloidogénica del procesamiento de APP[7,8,17].

1.3.3 Epidemiología y factores de riesgo

El aumento de la esperanza de vida de las personas ha provocado un incremento significativo de la población envejecida que supera los 65 años y por tanto, de las enfermedades asociadas a una edad avanzada, como es el caso de la EA. La incidencia de esta enfermedad aumenta casi exponencialmente con la edad hasta alcanzar los 85 años. Se estima que en 2050, solo en Estados Unidos, habrá 13.8 millones de personas diagnosticadas con demencia por Alzheimer. Esto refleja la importancia de la investigación en este campo, ya que el número de casos seguirá creciendo si no avanzamos en el conocimiento, diagnóstico, tratamiento y **prevención** de la enfermedad[17].

Respecto a los **factores de riesgo** de EA, el principal es la **edad**, aunque también se conocen enfermedades y factores asociados al estilo de vida, que influyen en el desarrollo de la enfermedad, y que podrían ser útiles para trabajar en la prevención de la misma. Por una parte, se ha descrito que el ejercicio físico regular, una dieta saludable baja en azúcares y grasas saturadas y las actividades mentales son factores importantes para disminuir el riesgo de la EA y retrasar la aparición o progresión de la enfermedad. Por otra parte, la

obesidad, la diabetes tipo 2, los factores de riesgo vascular y el estrés, pueden aumentar el riesgo de EA e influir en el desarrollo de la misma[17].

1.3.4 Diagnóstico y tratamiento

En cuanto al diagnóstico de la EA, en la actualidad es un **diagnóstico de probabilidad**, ya que el diagnóstico preciso y definitivo solo se obtiene en la autopsia del paciente[9]. Se basa principalmente en una entrevista clínica y en la evaluación de las capacidades cognitivas. Aunque también se realiza un examen físico y se evalúa la salud neurológica mediante pruebas para analizar reflejos, tono muscular y fuerza, equilibrio, coordinación, sentido de la vista y audición; acompañado de pruebas complementarias, como análisis de sangre o imágenes médicas, para descartar que la causa de la demencia sea otra enfermedad[8].

La EA podemos dividirla en tres fases: preclínica, deterioro cognitivo leve (DCL) y demencia. El diagnóstico en las fases iniciales es difícil por la ausencia de síntomas y en fases más avanzadas por la posible confusión de los síntomas con los provocados por otras enfermedades. Además, aunque todos los pacientes con EA han pasado por una fase de DCL no todos los pacientes con DCL desarrollan EA[9,19].

El objetivo perseguido sería el diagnóstico en etapas tempranas para prevenir sus irreversibles e incontrolables consecuencias. De ahí la importancia de los **biomarcadores** y las **técnicas de imagen** para el diagnóstico[9,19].

Comúnmente se utilizan tres biomarcadores del líquido cerebroespinal para el diagnóstico de EA. La baja concentración de Aβ y la alta concentración de proteína tau total o de proteína tau fosforilada en el líquido cefalorraquídeo pueden ser indicativos de la enfermedad. Sin embargo, al ser la extracción del líquido cefalorraquídeo un procedimiento muy invasivo, se está estudiando activamente la presencia de biomarcadores en sangre[19].

Debido a la variabilidad en las medidas de estos marcadores, se suelen combinar con métodos de diagnóstico por imagen para aumentar la eficiencia del diagnóstico. La **imagen por resonancia magnética** (MRI) se puede utilizar para estudiar los grandes cambios anatómicos y de conectividad (Figura 2), y la **tomografía por emisión de positrones** (PET) puede servir como modalidad de imagen molecular para rastrear la propagación del Aβ y la proteína tau (Figura 3) y descartar tumores, accidentes cerebrovasculares y lesiones cerebrales[19].

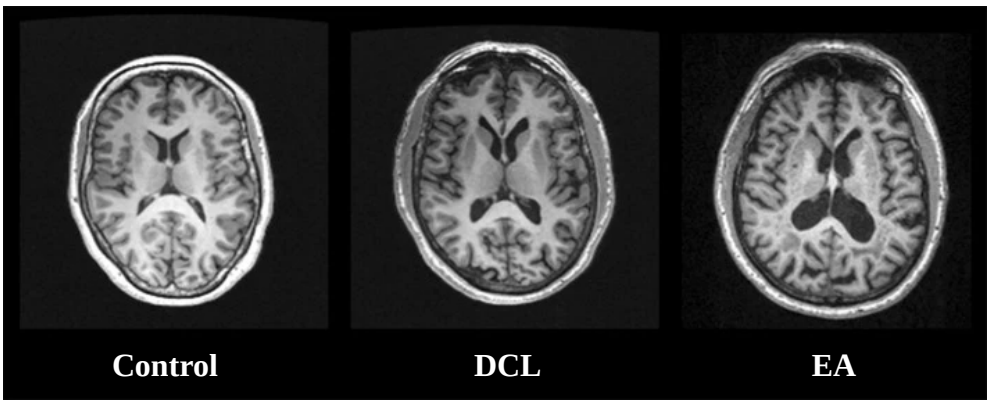

Figura 2: Comparación de MRI de control, paciente con DCL y paciente con Alzheimer. Se puede observar una disminución del volumen de materia gris en DCL y EA. Imagen modificada de Chandra, A. et al. (2019)[20].

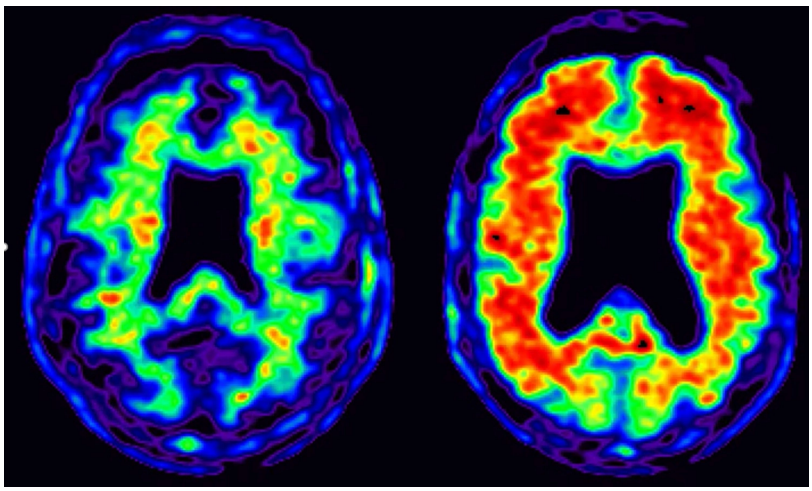

Figura 3: PET de control (izquierda) y paciente de Alzheimer (derecha). Los colores cálidos indican alta acumulación de Aβ. Imagen obtenida de Lane, C. A et al. (2018)[8].

Por último, se pueden llevar a cabo pruebas genéticas, que generalmente, se restringen a personas con antecedentes familiares de EA de aparición temprana.

Respecto al tratamiento, en la actualidad no existen **tratamientos** que modifiquen la EA, solo existen fármacos, como los inhibidores de la colinesterasa y la memantina, destinados al alivio de los síntomas, o tratamientos no farmacológicos, como la actividad física o el entrenamiento cognitivo, pero **no cambian la fisiopatología ni alteran el curso de la enfermedad**[1,19].

1.4 Transcriptómica

Aunque el genoma de todas las células de un organismo es prácticamente el mismo, con el estudio del transcriptoma o **transcriptómica** podemos conocer la información del genoma que se está transcribiendo en cada **tejido o tipo celular** en un **momento** y bajo una **condición** fisiológica o patológica determinada. La transcriptómica es fundamental para entender la estructura, la función y la regulación de la expresión de los genes, así como las alteraciones de los procesos biológicos que provocan las enfermedades humanas, pudiendo utilizar este conocimiento en el diagnóstico y tratamiento de las mismas[21,22]. Las dos técnicas principales utilizadas en la actualidad en este campo son los **microarrays**, que cuantifican un subconjunto de secuencias predeterminadas, y la **secuenciación de ARN** (RNA-seq, en inglés) que pretende capturar y cuantificar todas las secuencias[23]. Para este tipo de estudios se suelen utilizar muestras que contienen miles de células, basándose en la asunción de que las células de un tejido son homogéneas y obteniendo la expresión promedio del conjunto de células analizado[21,24].

1.4.1 Transcriptómica de célula única

La **secuenciación de ARN de célula única** (scRNA-seq, en inglés) es una técnica muy potente capaz de revelar la heterogeneidad celular enmascarada en las medidas del nivel de expresión medio de las poblaciones de células, obtenidas con los abordajes de secuenciación de ARN en masa (bulk RNA-seq, en inglés). Esta técnica genera grandes conjuntos de datos, lo cual supone un **desafío** respecto al análisis **computacional** e interpretación de los mismos. Otra barrera que encontramos en este tipo de tecnología es la

falta de estandarización en el análisis de los datos, debido a que es un campo muy novedoso[25–27].

Existen varios métodos de secuenciación de scRNA-seq, sin embargo, se ha centrado la atención en el método *10xGenomics chromium*[26], por ser el abordaje utilizado en los tres estudios inicialmente seleccionados para este trabajo. *10xGenomics chromium* es un **método basado en gotas**, muy usado en la actualidad por su alto rendimiento y su bajo coste. El procedimiento experimental puede resumirse en los siguientes pasos: 1) Extracción de la muestra del órgano de interés. 2) Disociación del tejido. 3) Aislamiento de células individuales y construcción de librerías. 4) Secuenciación. 5) Análisis bioinformático de los datos (Figura 4).

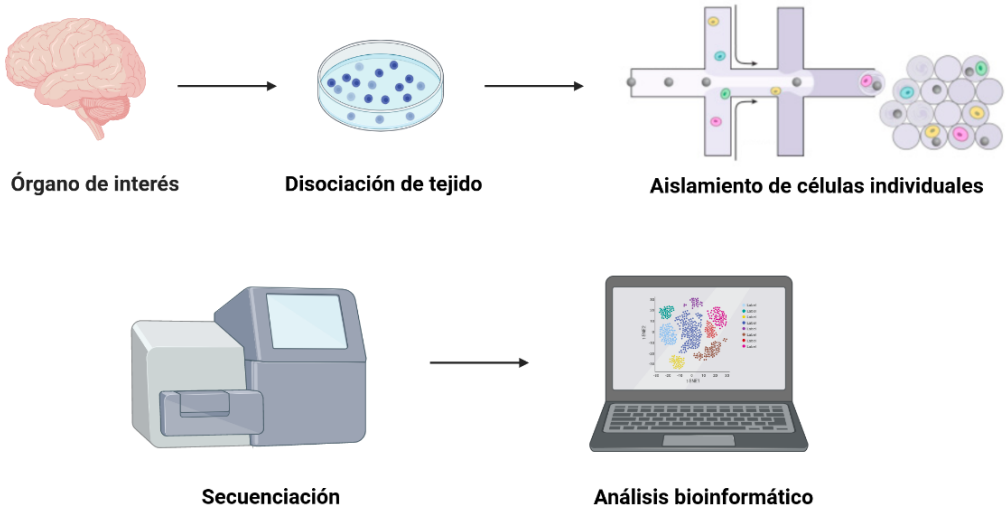

Figura 4: Procedimiento de los experimentos de transcriptómica de célula única. 1) Extracción de la muestra del órgano de interés. 2) Disociación del tejido. 3) Aislamiento individual de las células y construcción de las librerías. 5) Secuenciación y 6) Análisis bioinformático de los datos. Imagen creada con BioRender.com utilizando una imagen de Macosko, E. Z. et al (2015)[28].

Para aislar las células individuales se utiliza un chip de microfluidos por el que se pasa un flujo de células, un flujo de perlas y un flujo de aceite, generando gotas por emulsión. Aunque lo ideal sería que cada gota contuviese una perla y una célula, las proporciones de perlas y células se ajustan para que las mayoría de las gotas contengan una o ninguna célula, intentando impedir que dos células sean capturadas en la misma gota, generando lo que se conoce como un **doblete** (Figura 5)[29–31].

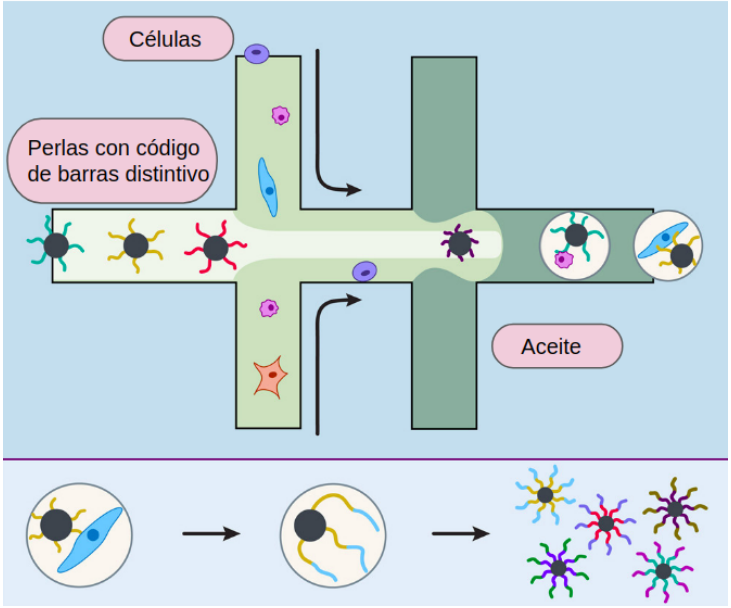

Figura 5: Método de secuenciación de scRNA-seq basado en gotas. Se pasa un flujo de células y un flujo de perlas con cebadores por un chip de microfluidos, generando gotas por emulsión al incorporar un flujo de aceite. Imagen modificada de Macosko, E. Z. et al (2015)[28].

Cada perla contiene cebadores con identificadores moleculares únicos (**UMI**, en inglés) y un código de barras (***barcode***, en inglés) específico de cada perla. Dentro de las gotas se produce la lisis celular y la captura de los ARN por los cebadores. A continuación, se llevan a cabo los pasos de transformación a ADNc, amplificación, construcción de la librería con los extremos 3' y por último, la secuenciación[30,31].

Tras la secuenciación podemos saber las moléculas que pertenecen a cada célula por su *barcode* y diferenciar, gracias a los UMIs, entre las copias obtenidas por la amplificación del ADNc o de la librería, procedentes del mismo ARN, y las lecturas procedentes de distintas moléculas de ARN transcritas a partir del mismo gen, las cuales nos informarán sobre el nivel de expresión del mismo. A partir de los datos de secuenciación, se construye una matriz cuyas filas y columnas corresponden a genes y células respectivamente, que contiene el número de conteos o transcritos por gen en cada célula (Figura 6)[27,30].

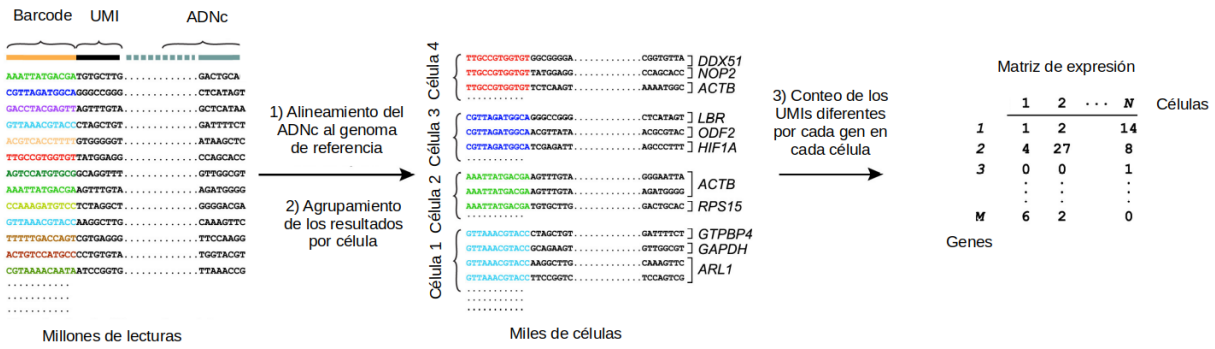

Figura 6: Generación de la matriz de conteos. Las lecturas obtenidas en la secuenciación se alinean contra un genoma de referencia para identificar a qué genes pertenecen. Los barcodes de las lecturas permiten asignarlas a una célula. Por último, se crea la matriz en la que se refleja el número de UMIs por gen en cada célula. Imagen modificada de Macosko, E. Z. et al (2015)[28].

Por último, una característica importante a tener en cuenta en los datos de transcriptómica de célula única es el fenómeno de "**dropout**", que ocurre debido a la pequeña cantidad de ARN de partida que contiene una única célula y a la ineficiencia de captura del los ARN. Esta situación provoca que un gen que tenga una expresión baja o moderada en una célula no sea detectado en una célula del mismo tipo. Como consecuencia de este fenómeno se obtienen muchos ceros en la matriz de conteos[25].

1.4.2 Transcriptómica de núcleo único

La transcriptómica de núcleo único (snRNA-seq, en inglés) es una estrategia muy similar a la de célula única, pero utilizando núcleos en lugar de las células completas. Se utiliza con tejidos que **no pueden ser disociados** fácilmente en una suspensión celular, como el cerebro, el músculo esquelético o el tejido adiposo y en muestras congeladas. Las principales ventajas son que no es necesario proteger la integridad celular durante la preparación de las muestras y que minimiza los cambios en la expresión génica provocados por la disociación del tejido. Sin embargo, aunque los transcriptomas de la célula completa y del núcleo son muy similares, existen algunas diferencias[29,32].

Grindberg, R. et al. describieron que aproximadamente el 98% de los transcritos estaban igualmente representados en núcleos y célula completa. Estos resultados junto con los de otros estudios[33,34] confirmaron que el **uso del núcleo** para el estudio del transcriptoma **no introduce grandes alteraciones en las medidas de expresión génica**. No obstante, en el núcleo encontraron una sobrerrepresentación de transcritos relacionados con la regulación de la transcripción (GO:0006355) y la regulación de los procesos metabólicos del ARN (GO:0051252), lo cual concuerda con el hecho de que el núcleo es la fuente de ARN. El estudio del transcriptoma nuclear también puede ser útil para investigar los transcritos que se acumulan selectivamente en el núcleo[35].

Debido a la **heterogeneidad** celular presente en el cerebro y la **dificultad** que presenta la **disociación** de las células de las muestras tomadas del mismo, que además, suelen estar congeladas, los abordajes de transcriptómica de núcleo único son fundamentales para el estudio de este órgano y de las enfermedades que afectan al mismo. Todos los *datasets* analizados en este trabajo utilizan esta estrategia. Los principales tipos celulares que se pueden encontrar en el cerebro son las neuronas, las células gliales (microglía, oligodendrocitos y astrocitos) y las células endoteliales. Sin embargo, este trabajo se ha centrado en la microglía y los astrocitos por su papel en la neuroinflamación en la EA[36]. Además, como la neuroinflamación es un tema candente en la actualidad, muchos de los estudios publicados en repositorios están orientados a estos tipos celulares, por lo que si nos centramos en estos tipos celulares podremos hacer un análisis integrativo con mayor número de estudios.

A partir de ahora y en el resto del manuscrito, se emplearán indistintamente las palabras "células" o "núcleos" para referirnos a los núcleos utilizados en los abordajes de snRNA-seq.

2. Objetivos

El objetivo principal de este trabajo es la identificación de las bases moleculares que afectan de forma diferencial a hombres y mujeres en la enfermedad de Alzheimer mediante un abordaje *in silico*. Para ello, se han propuesto cuatro objetivos específicos:

1) Revisar bibliografía sobre la Enfermedad de Alzheimer y el análisis de datos de transcriptómica de célula única.

2) Seleccionar mediante una revisión sistemática estudios de Alzheimer con datos de transcriptómica de célula única y de núcleo único que incluyan información de sexo disponibles en repositorios públicos.

3) Analizar bioinformáticamente cada uno de los estudios seleccionados.

4) Comparar e integrar los resultados de los estudios individuales.

3. Materiales y métodos

La metodología desarrollada en este trabajo puede resumirse en tres pasos (Figura 7). En primer lugar, se llevó a cabo una revisión sistemática que permitió la selección de los estudios cuyos datos se descargaron posteriormente para ser analizados. A continuación, se analizó cada estudio de forma individual con el objetivo final de detectar si existen diferencias entre hombres y mujeres con la EA tanto a nivel de expresión génica en los tipos celulares analizados, como a nivel de número de células de cada tipo celular. Por último, se compararon los resultados obtenidos en los tres estudios. Se llevó a cabo la intersección de los genes significativos expresados diferencialmente en cada estudio. Los genes comunes fueron caracterizados funcionalmente. Para ello se comprobó en distintas bases de datos si estos genes estaban asociados previamente a la EA y se identificaron los procesos biológicos de la Gene Ontology (GO) asociados a cada gen. Si el número de genes era muy elevado para algún grupo, se realizó un análisis de sobrerrepresentación y se estudió la red formada por las asociaciones entre las proteínas codificadas por estos genes, en el caso de que fuesen codificantes, descritas en distintas bases de datos.

Para el análisis individual de los estudios se desarrolló una *pipeline* adecuada al análisis de datos de transcriptómica de núcleo único, cuyos pasos, resumidos en la Figura 7, serán explicados con más detalle en los siguientes apartados.

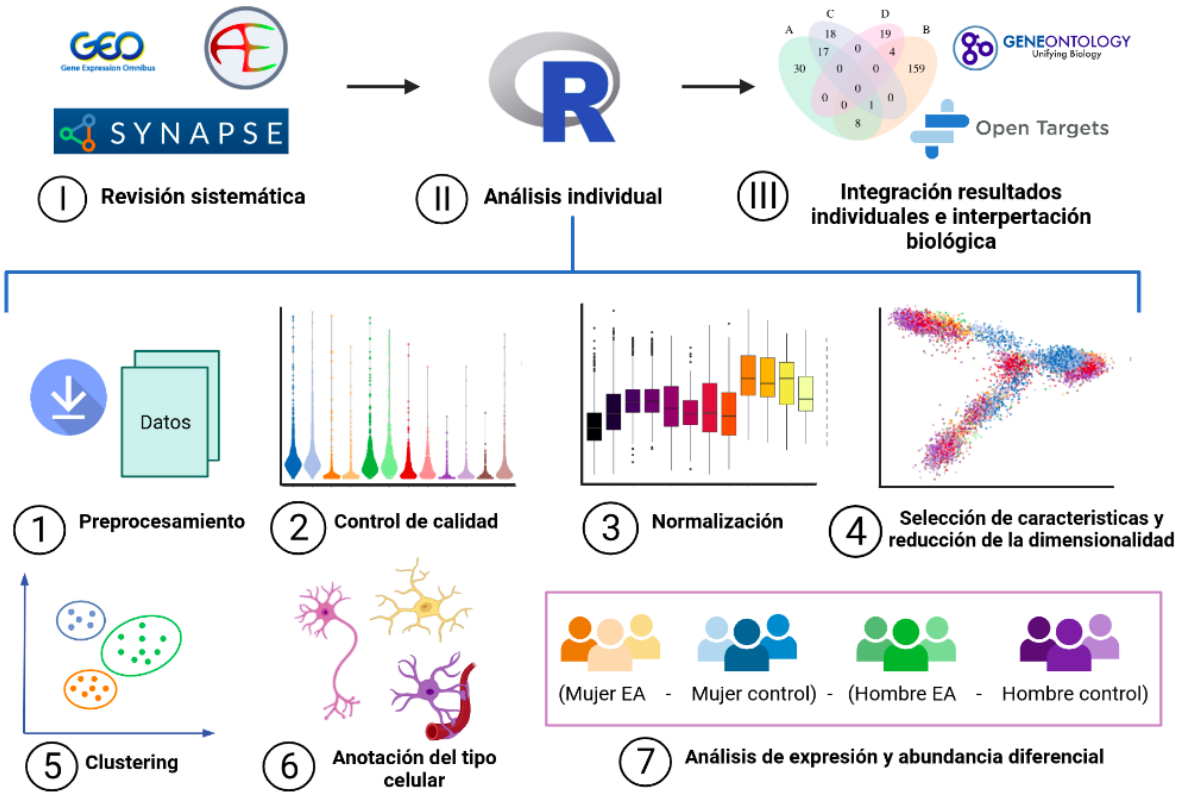

Figura 7: Metodología. Consta de tres pasos: I) Revisión sistemática para buscar los datos a analizar en los repositorios públicos. II) Análisis individual de los estudios seleccionados. Los pasos llevados a cabo son el preprocesamiento para construir la matriz de conteos a partir de los datos de secuenciación, el control de calidad para eliminar las células de baja calidad, la normalización para hacer los conteos comparables, la selección de características y reducción de dimensionalidad para reducir y compactar los datos, el clustering para agrupar las células por similitud de perfil de expresión génica, la anotación del tipo celular y por último, los análisis de expresión y abundancia diferencial para detectar diferencias entre hombres y mujeres con EA. III) Combinación e integración de los resultados individuales e interpretación biológica de los mismos.

El análisis bioinformático asociado a este trabajo se ha llevado a cabo en lenguaje de programación R[37]. El Software y los paquetes utilizados en el análisis, así como el código de los programas generados están disponibles en el repositorio **ZENODO**[38], en los anexos I y II respectivamente del siguiente enlace (https://zenodo.org/record/5457264). Este repositorio está enfocado para compartir datos, programas u otro material adicional derivados de la actividad científica, fomentando así la ciencia de libre acceso.

El coste computacional de este tipo de abordajes es muy alto y precisa una gran capacidad de memoria para almacenar grandes cantidades de datos biológicos. Para reducir el tiempo de cómputo se utilizó el paquete BiocParallel[39] para paralelizar aquellas funciones que lo permitiesen, y se aumentó el número de núcleos (o cores, en inglés) utilizados para ejecutar los programas en función de las dimensiones de la matriz de partida. En un ordenador personal, el número de núcleos que se podría usar sería entre 1 y 4 y la memoria también podría ser un problema. Por ello, ha sido necesario el uso del Cluster del Centro de Investigación Príncipe Felipe (CIPF) que actualmente cuenta con 44 nodos, 600 CPUs y una memoria RAM de 11 TeraBytes. Además, presenta un sistema de almacenamiento de archivos distribuido llamado Lustre de 1 PetaByte de capacidad. Todas

estas características han asegurado la viabilidad del análisis y del almacenamiento de datos biomédicos del presente trabajo.

3.1 Revisión sistemática y selección de estudios

En febrero de 2021 se llevó a cabo una búsqueda y revisión sistemática, siguiendo las directrices de la declaración PRISMA[40], de estudios de transcriptómica de célula única o de núcleo único de *Homo sapiens*, centrados en la enfermedad de Alzheimer, en los repositorios GEO[41] y ArrayExpress[42], y en la herramienta de búsqueda de Google, utilizando como palabra clave "Alzheimer" y aplicando como filtro un tamaño muestral mínimo de 12 sujetos. A continuación, se examinaron los estudios con más detalle y se aplicaron los criterios de exclusión de la Tabla 1 para seleccionar los estudios incluidos en este trabajo.

Tabla 1: Criterios de inclusión y exclusión utilizados en la revisión sistemática.

Criterios de inclusión	Criterios de exclusión
- Repositorios GEO y ArrayExpress y la herramienta de búsqueda de Google. - Organismo: *Homo sapiens.* - Palabra clave: "Alzheimer". - Tipo de estudio: Transcriptómica de célula única o núcleo único. - Tamaño muestral mínimo: 12 sujetos	- No centrado en Alzheimer. - Abordaje distinto a transcriptómica de célula única o núcleo único. - Diseño experimental distinto al buscado, se necesitan muestras de pacientes y controles. - Tamaño muestral insuficiente (al menos 3 muestras por grupo experimental) - Carencia de información de sexo o ausencia de representación de ambos sexos tanto en controles como en pacientes.

3.2 Análisis individual de los estudios

Los datos obtenidos mediante transcriptómica de célula única pueden ser analizados de distintas formas según la información que se quiera obtener de los mismos y los objetivos del estudio. En este caso se desarrolló una *pipeline* para analizar los datos de single nucleus RNA-seq procedentes de la técnica de secuenciación *10xGenomics,* con el objetivo final de llevar a cabo un análisis de expresión diferencial por tipo celular y un análisis de abundancia diferencial. Para ello se tomaron de referencia, el manual online "Orchestrating Single-Cell Analysis with Bioconductor"[29], el artículo "Current best practices in single-cell RNA-seq analysis: a tutorial"[27], los artículos originales de los estudios incluidos en este análisis y otros estudios similares, y scripts almacenados en GitHub.

El análisis de datos obtenidos por snRNA-seq es muy similar al que se realiza con los datos de transcriptómica de célula única. La mayor diferencia reside en la construcción de la matriz de conteos, ya que en el núcleo hay presentes ARNm no maduros por lo que hay que tener en cuenta las regiones intergénicas en la anotación de los genes. Por otra parte, la supresión del citoplasma está asociada a la pérdida de transcritos mitocondriales,

siendo su presencia indicativa de un aislamiento incompleto del núcleo o de contaminación ambiental. Además, si se lleva a cabo la anotación celular basada en perfiles de expresión procedentes de células de referencia, hay que tener en cuenta que la mayoría de referencias están construidas a partir de datos de bulk RNA-seq o de single Cell RNA-seq[29].

3.2.1 Preprocesamiento

Consiste en la construcción de las matrices de conteos a partir de los datos crudos generados por los equipos de secuenciación[27,29]. Normalmente los datos obtenidos tras la secuenciación por el método *10xGenomics Chromium* usando las plataformas de Illumina, son analizados con el Software *10xGenomics Cell Ranger*, el cual permite alinear las lecturas y generar las matrices de conteos.

En este trabajo se partió directamente de las matrices de conteos o de los ficheros de salida de sofware *10xGenomics Cell Ranger*, al estar disponibles en los repositorios públicos. Las matrices de conteos y los metadatos asociados se importaron en el entorno de R en un objeto de la clase **SingleCellExperiment** del paquete **SingleCellExperiment**[29]. Otra alternativa sería el uso del objeto de la clase Seurat[43] del paquete Seurat. Sin embargo, se optó por SingleCellExperiment por pertenecer al proyecto Bioconductor[44], el cual presenta la ventaja de la interoperatividad entre sus paquetes[29]. Para generar el objeto SingleCellExperiment a partir de los ficheros de salida del sofware *10xGenomics Cell Ranger* se utilizó la función **read10xCounts()** del paquete **DropletUtils**[45].

La clase SingleCellExperiment (Figura 8) es una estructura de datos que permite almacenar la matriz de conteos y los metadatos de las células y de los genes, y manipularlos de una forma sincronizada. Además, permite ir guardando los resultados propios del análisis de datos de transcriptómica de single cell, como son las métricas del control de calidad, los conteos normalizados o los resultados de la reducción de la dimensionalidad. Esto permite que el propio objeto sirva como registro del análisis que se ha llevado a cabo[29].

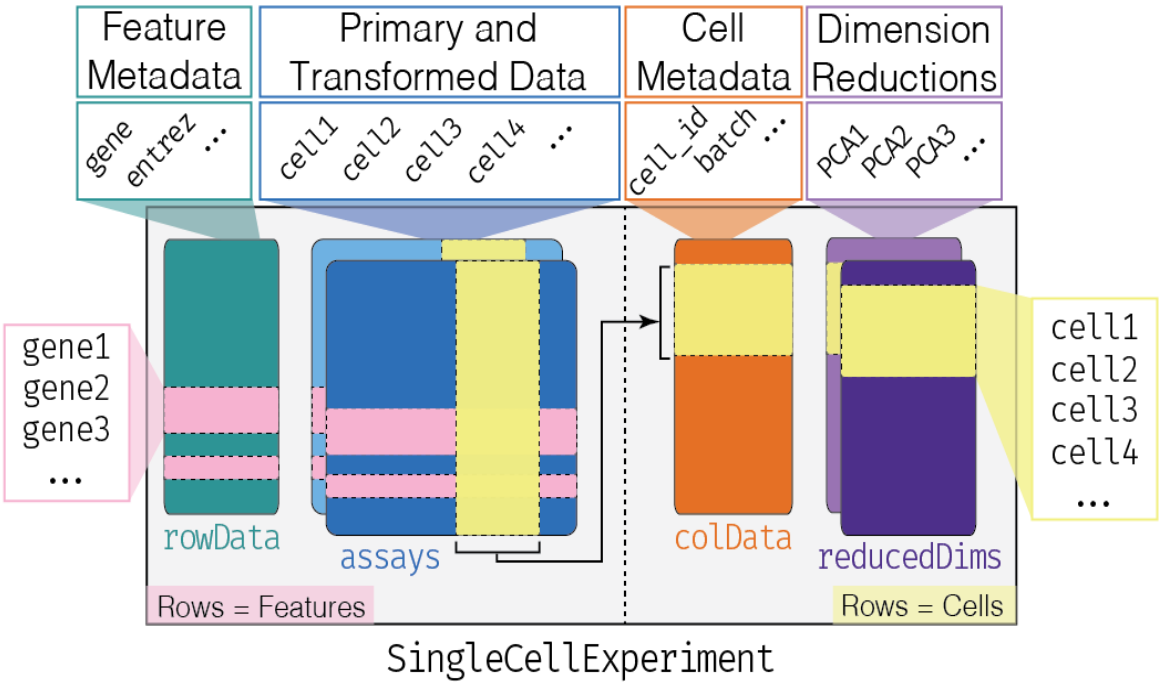

Figura 8: Esquema de la estructura de datos del objeto SingleCellExperiment del paquete con el mismo nombre. Figura obtenida del manual de Bioconductor[29].

Por último y para facilitar el posterior estudio integrativo, se crearon objetos con una estructura de información similar para cada uno de los estudios. Se les asignó el mismo nombre a las variables de las células que designan el sujeto al que pertenecen, el sexo y la condición, y se creó un a variable conjunta que aúna el sexo y la condición. En el caso de los genes se utilizaron los identificadores "gene symbol" como identificadores primarios.

3.2.2 Control de calidad

Este paso se llevó a cabo para eliminar los núcleos de baja calidad, originados por algún problema asociado a la preparación de las muestras o por algún fallo en la preparación de la librería (transformación a ADNc o amplificación por PCR ineficientes). El principal problema de este paso es la asunción de que las métricas de calidad son impulsadas por factores técnicos y son independientes del estado biológico de cada célula. Esta asunción es problemática en poblaciones celulares heterogéneas, pudiendo provocar la eliminación de tipos celulares[29].

Para identificar estos núcleos o células de baja calidad se suelen utilizar 3 covariables: **número de conteos por células**, **número de genes por célula** y el **porcentaje de genes mitocondriales**. Se suelen eliminar células que presentan número de conteos y número de genes detectados bajos y alto porcentaje de genes mitocondriales, ya que estos datos podrían proceder de células a las que se le ha roto la membrana y han perdido el citoplasma. Por otra parte, si el número de conteos y genes detectados son altos, puede ser que estemos ante un caso de doblete, por lo que también conviene eliminarlas. En el caso del abordaje de transcriptómica de núcleo único no se deberían detectar genes mitocondriales, pudiendo estar asociada su presencia a un aislamiento incompleto del núcleo o contaminación ambiental. Por último, si el número de conteos y el número de características de la célula son bajos, puede ser indicativo de una captura y amplificación del ADNc ineficientes[27,29].

Para llevar a cabo el control de calidad de los núcleos se utilizó la función **addPerCellQC()** del paquete **scater**[46], que calcula las tres métricas de calidad descritas anteriormente y las añade al objeto SingleCellExperiment. Por otra parte, como el método usado para la secuenciación en los tres estudios es un procedimiento basado en gotas, se tuvo en cuenta la posibilidad de que los conteos que tenemos asociados a un núcleo, en realidad provengan de una gota vacía o de una gota con dos células (dobletes). En este caso se decidió no evaluar ni filtrar la presencia de gotas vacías porque el software Cell Ranger usado por los autores de los artículos originales lo lleva a cabo automáticamente. En cambio, sí se utilizó la función **scDblFinder()** del paquete **scDblFinder**[47] para detectar y eliminar dobletes. Esta función identifica dobletes, generando dobletes artificiales basándose en *clusters* y evaluando su prevalencia entre los vecinos. Es importante paralelizar esta función, ya que lleva a cabo procesos computacionalmente muy costosos[29].

Una vez se tienen las métricas de calidad se establecen filtros para eliminar los núcleos de baja calidad. En estos casos existen dos opciones, que el analista de datos establezca sus **propios umbrales**, lo cual requiere mucha experiencia, o usar **umbrales adaptativos**. La segunda opción consiste en asumir que la mayoría de células son de alta calidad e identificar valores atípicos usando la desviación media absoluta[29]. Para ello se puede utilizar la función **isOutlier()** del paquete **scuttle**[46].

Como a priori no se puede saber si la calidad de los datos es suficiente, se llevó a cabo un doble abordaje de forma paralela para poder decidir, tras evaluar la distribución de las covariables en el *clustering* y en la visualización de los datos, los filtros más adecuados. Este paso es muy importante en muestras como las que estamos analizando, ya que contienen tipos celulares bastante heterogéneos, pudiendo provocar la confusión de un tipo celular con células de baja calidad[27]. Por una parte, se filtró cada estudio con los umbrales establecidos por la función isOutlier(), y por otra parte, se aplicaron filtros comunes para todos los estudios, más permisivos y basados en el conocimiento adquirido tras consultar la metodología seguida en los artículos originales, así como los resultados obtenidos con la función isOutlier().

Los filtros aplicados en el primer abordaje se reflejan en la Tabla 2, son bastante restrictivos sobre todo respecto al porcentaje de genes mitocondriales, lo cual concuerda con la supuesta ausencia de genes mitocondriales en experimentos de snRNA-seq. En el segundo abordaje se eliminaron aquellos núcleos en los que el número de características fuese menor a 200 o el porcentaje de genes mitocondriales fuera mayor al 10%. En ambos abordajes se eliminaron los núcleos identificados como dobletes por scDblFinder().

Tabla 2: Filtros establecidos por la función isOutlier() para cada uno de los estudios.

Estudio	Número total de conteos	Número de características	Porcentaje de genes mitocondriales
GSE138852	<227,97	<201,09	>1,27
GSE157827	<118,39	<151,57	>4,04
GSE160936	<429,84	<497,71	>2,52

En el control de calidad también se suele realizar a nivel de transcritos, ya que las matrices suelen contener demasiados genes, que pueden ser reducidos si eliminamos aquellos que estén expresados en pocas células y que por tanto no sean relevantes para explicar la heterogeneidad celular[27]. Para ello se usó el criterio establecido por el artículo del *dataset* GSE138852[48]. Se eliminaron aquellos genes que no tenían 2 o más transcritos en al menos 10 células. Para ello se utilizó la función **nexprs()** del paquete **scater**[46].

3.2.3 Normalización

En el procedimiento experimental se introducen diferencias técnicas entre células que hacen que los conteos obtenidos no sean comparables. En este paso del análisis se pretende corregir los conteos de la matriz para que las diferencias de expresión en las comparaciones entre células se deban exclusivamente a factores biológicos. En este trabajo se llevó a cabo una normalización de escalado, por ser la más simple y las más usada entre las estrategias de normalización. Esta estrategia consiste en dividir los conteos entre un factor de escalado específico de cada célula[27,29].

La elección del método de normalización depende del análisis posterior que se vaya a hacer de los datos. Si el objetivo del estudio fuese identificar clusters, así como los principales genes marcadores de los mismos, nos bastaría con aplicar el método de bulk RNA-seq basado en el tamaño de la librería. En este método, para el cálculo del factor de

escalado, se asume que el tamaño de la librería de todas las células del análisis debería ser el mismo, debido a que la mayoría de los genes no están expresados diferencialmente entre células, y los pocos que sí, se compensan, ya que algunos están sobreexpresados y otros infraexpresados. Sin embargo, en los abordajes de célula única, encontramos poblaciones celulares heterogéneas, que no tienen por qué tener un tamaño de librería similar. Además, hay que tener en cuenta el fenómeno de "dropout", característico de transcriptómica de célula única, que provoca la obtención de matrices de conteos con alta prevalencia de 0 y números bajos. Por estas razones, es más conveniente utilizar un método de normalización más complejo específico de transcriptómica de célula única, sobre todo si se quieren analizar las diferencias de expresión entre distintas condiciones, como es el caso de este trabajo[27,29].

En este caso se utilizó el método de deconvolución (Figura 9), que consiste en agrupar los conteos de muchas células para incrementar el número de conteos, reduciendo así la problemática incidencia de 0 en la matriz, y poder estimar de forma más precisa el factor de escalado. Los grupos de células se realizan de forma solapada para poder después inferir los factores de escalado específicos de cada célula mediante un sistema de ecuaciones lineales[29,49].

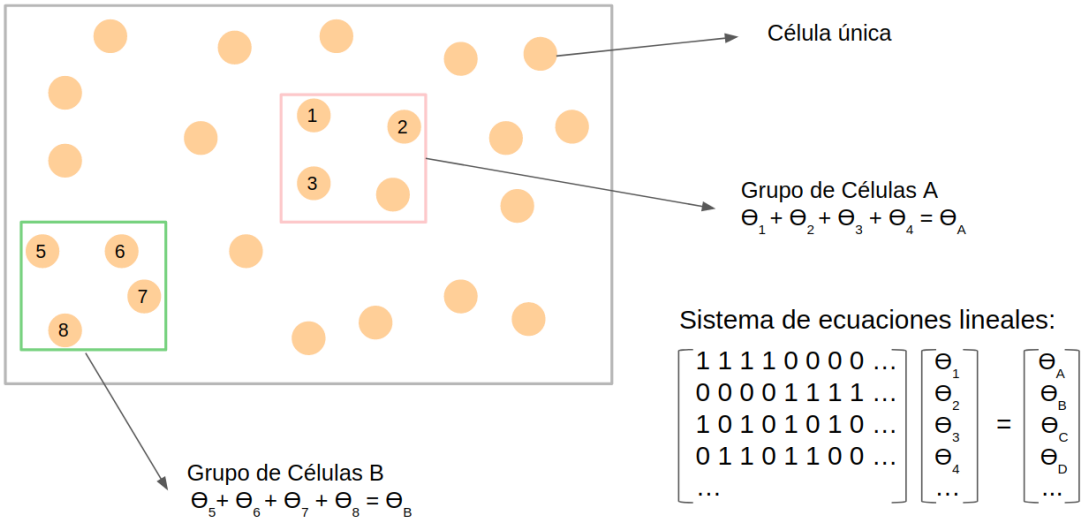

Figura 9: Esquema del método de normalización de deconvolución, que consiste en agrupar las células y sumar sus conteos para estimar el factor de escalado. Los factores de escalado específicos de cada células se infieren después por deconvolución, mediante un sistema de ecuaciones lineales. Figura propia basada en Lun, Aaron T. L. et al[49].

Para llevar a cabo la normalización, se usó en primer lugar la función **quickCluster()** del paquete **scran**[50], que agrupa células similares en función de sus perfiles de expresión. A continuación, se utilizó la función **computeSumFactors()** del paquete **scran**, que calcula los factores escalado específicos de cada célula mediante el método de deconvolución. Por último, se usó la función **logNormCounts()** del paquete scater[46], que lleva a cabo la normalización y el logaritmo en base 2 de los valores de expresión de cada célula. Esta transformación logarítmica es útil porque las distancias entre valores transformados logarítmicamente representan el logFC, medida habitualmente utilizada para detectar cambios en la expresión. Además, reduce la asimetría de los datos aproximándolos a una distribución normal, asumida por muchos de los métodos en el análisis posterior[27,29].

3.2.4 Selección de características y reducción de la dimensionalidad

Este paso consiste en reducir y compactar las matrices de alta dimensionalidad obtenidas en este tipo de abordajes. En primer lugar, se eliminan parte de los genes contenidos en la matriz para reducir el ruido y el coste computacional del análisis posterior. La estrategia más simple para el paso de selección de características sería seleccionar los genes más variables basándonos en la expresión del mismo en la población de núcleos analizada[27,29].

Para ello se utilizaron dos funciones del paquete **scran**. En primer lugar, **modelGeneVar()** que modeliza la varianza por gen a partir del logaritmo de los valores de expresión normalizados, descomponiéndola en varianza debida a componentes técnicos y a componentes biológicos. Se basa en la tendencia media-varianza ajustada. En segundo lugar, **getTopHVGs()** para seleccionar los genes más variables. En este caso se seleccionaron 3000 genes[29].

Estos genes solo se utilizaron para la reducción de dimensionalidad, el *clustering* y la anotación celular. Una vez anotados los tipos celulares se recuperaron todos los genes para llevar a cabo el análisis de expresión diferencial.

A pesar de haber reducido el número de genes a 3000, aún se puede reducir más la matriz de expresión. En esta matriz cada gen representa una dimensión de los datos, sin embargo, estas dimensiones se pueden reducir, ya que muchos genes están correlacionados al participar en el mismo proceso biológico. La reducción de la dimensionalidad tiene dos objetivos, resumir y visualizar los datos. En este trabajo se optó por el método de reducción de la dimensionalidad denominado “Análisis de componentes principales”, PCA. Es un método lineal que resume la dimensionalidad de los datos en sus N componentes principales, los cuales capturan los factores dominantes responsables de la heterogeneidad. N puede ser determinado por el método del codo. La ventaja del PCA respecto a otros métodos de reducción de la dimensionalidad es que al ser lineal, las distancias son comparables en todo el espacio dimensional reducido[27,29].

Para llevar a cabo el PCA se usó la función **runPCA()** del paquete **scater** sobre los valores de expresión normalizados y en escala logarítmica, la cual calcula las 50 primeras componentes principales y las almacena en el objeto SingleCellExperiment. A continuación, se utilizó la función **findElbowPoint()** del paquete **PCAtools**[51] para determinar N por el método del codo. Por último, se usó la función **plotReducedDim()** del paquete **scater** para realizar distintas representaciones.

3.2.5 Clustering

Este paso consiste en agrupar las células por sus perfiles de expresión con el objetivo de resumir los datos para una exploración e interpretación más fácil de los mismos. En este caso para generar los *clusters* se usó un método de detección de comunidades. Los métodos de detección de comunidades son métodos basados en grafos. Estos grafos son construidos usando un abordaje *K-Nearest Neighbors*, K vecinos más próximos. En estos grafos cada nodo es una célula que está conectada a las K células más parecidas, el peso de cada enlace depende de la similitud existente entre las células que conecta. Por último, se aplica un algoritmo para identificar comunidades de células que están más interconectadas entre sí que con el resto de células[27,29].

En este estudio se utilizó la función **buildSNNGraph()** de paquete **scran** para la construcción del grafo utilizando los datos dimensionalmente reducidos por el método PCA y considerando K = 20 vecinos para construirlo. Este parámetro controla la resolución de los *clusters*, a mayor número de vecinos menor número de *clusters* se obtendrán. Para determinar este parámetro se hicieron varias pruebas, no obstante no existe un número de *clusters* correcto o verdadero, solo depende cómo se quieran explorar los datos. Para la detección de comunidades se utilizó el algoritmo de walktrap con la función **cluster_walktrap()** del paquete **igraph**[52].

3.2.6 Anotación del tipo celular

Para la identificación y anotación del tipo celular se utilizó el paquete **BRETIGEA**[53], el cual permite identificar los seis tipos celulares principales que se pueden encontrar en el cerebro (neuronas, astrocitos, oligodendrocitos, microglía, células precursoras de oligodendrocitos (OPCs, en inglés) y células endoteliales), utilizando un conjunto de genes marcadores curados específicos de cada tipo celular. En principio este paquete fue creado para estimar la proporción de células de cada tipo en *datasets* de bulk RNA-seq, pero en la actualidad también se usa en *datasets* de scRNA-seq para la identificación del tipo celular. En concreto se utilizó la función **brainCells()**, que analiza en la matriz de conteos los niveles de expresión de los genes marcadores específicos de cada tipo celular y asigna a cada célula una puntuación para cada tipo celular. Finalmente, se anota cada célula con el tipo celular para el que tenga la puntuación más alta[48].

3.2.7 Análisis de expresión diferencial

Una vez anotados los tipos celulares se recuperó la matriz de conteos sin normalizar con todos los genes que fueron filtrados en el paso de selección de características y se llevó a cabo un análisis de expresión diferencial por tipo celular. En este paso se pretende conocer los genes diferencialmente expresados en la EA en la microglía y los astrocitos. Para ello se plantearon 3 contrastes:

- **(Mujer EA – Mujer Control)**, para identificar qué genes se expresan diferencialmente en mujeres con EA respecto a las mujeres sanas, al que nos referiremos en el resto del trabajo como contraste mujer.
- **(Hombre EA – Hombre Control)**, para identificar qué genes se expresan diferencialmente en hombres con EA respecto a los hombres sanos, al que nos referiremos en el resto del trabajo como contraste hombre.
- **((Mujer EA – Mujer Control) - (Hombre EA – Hombre Control))**, para detectar diferencias de sexo en la EA, al que nos referiremos en el resto del trabajo como contraste de diferencias de sexo.

Para ello, se realizó un doble abordaje, utilizando dos de los paquetes más usados actualmente para llevar a cabo análisis de expresión diferencial con datos de scRNA-seq.

En primer lugar, **edgeR**[54], paquete específico para el análisis de bulk RNA-seq, que fue diseñado para analizar la varianza genética de pocas muestras. Los datos de single cell sin embargo, contienen la información de muchas muestras o células y presentan unas peculiaridades específicas como la alta cantidad de 0 en los conteos provocado por el fenómeno de dropout o la alta variabilidad entre células[27]. Para poder utilizar este método

propio de bulk RNA-seq con datos de scRNAseq, se suman los conteos de las células de cada tipo celular de un mismo sujeto, obteniendo una pseudo-célula de cada tipo celular por cada sujeto. A continuación, se lleva a cabo el análisis de expresión diferencial para cada tipo celular por separado utilizando el método quasi-likelihood (QL) del paquete edgeR, que utiliza un modelo lineal generalizado binomial negativo (NB GLM)[29].

Para el abordaje con edgeR, en primer lugar se usó la función **aggregateAcrossCells()** del paquete **scuttle**[46] para construir el *pseudo-bulk dataset*. A continuación, se filtraron los genes con baja expresión para reducir el coste computacional y mejorar la precisión de la modelización con la función **filterByExpr()** del paquete **edgeR**. Después, se llevó a cabo un paso de normalización por el método TMM (Trimmed Mean of M-values), típico de bulk RNA-seq, con la función **calcNormFactors()** de **edgeR**. En este caso no se necesitaría el método de deconvolución porque se han sumado los conteos haciendo el *dataset* similar a uno de bulk RNA-seq. A continuación, se creó la matriz de diseño con **model.matrix()** del paquete **stats,** incluyendo la variable que combina el sexo y la condición en el modelo y se definió el contraste correspondiente con la función **makeContrasts()** del paquete **limma**[55]. Por último, se emplearon tres funciones del paquete edgeR: **estimateDisp();** para estimar la dispersión binomial negativa para cada tipo celular, **glmQLFit()**; para estimar la dispersión QL y **glmQLFTest()**; para evaluar las diferencias en la expresión génica entre los grupos definidos en el contraste correspondiente, considerando significativos aquellos genes en los que el logFC de la comparación correspondiente es distinto de cero con un FDR del 5%. El signo del logFC indica si el gen está sobreexpresado (+) o infraexpresado (-) en el contraste correspondiente.

Como el número de células de cada tipo celular en cada sujeto era muy distinto, se llevó a cabo una modificación del abordaje anterior que consistió en seleccionar en todos los sujetos el mismo número de células. El sujeto con menor número de células estableció el número de células que se seleccionaron aleatoriamente en el resto de sujetos. Con este abordaje se pretendía que el número de células no introdujese un sesgo en los resultados.

En segundo lugar, se utilizó **MAST**[56], paquete específico para el análisis de scRNA-seq, por lo que tiene en cuenta las peculiaridades de este tipo de datos. Este paquete trabaja con los conteos de las células individuales y utiliza un modelo de obstáculos (hurdle model, en inglés), un modelo lineal generalizado de dos partes que considera el fenómeno de “dropout”[27]. Este modelo está formado a su vez por dos modelos, un modelo de regresión logística para modelizar si existe o no expresión del gen y un modelo gaussiano para modelizar el nivel de expresión cuando corresponda. Para llevar a cabo este abordaje se utilizaron los conteos sin normalizar en escala logarítmica y se aplicaron dos funciones del paquete MAST. En primer lugar, **zlm()** para ajustar el modelo de obstáculos para cada gen. Además de la matriz de conteos, a esta función se le proporcionó la variable que combina el sexo y la condición experimental, y el número escalado de genes. En segundo lugar, la función **lrTest()** para el cálculo de los estadísticos, a la que se le proporcionó el contraste correspondiente. Por último, se corrigieron los p-valores por el método de Benjamini-Hochberg y se consideraron significativos aquellos genes en los que el logFC de la comparación correspondiente es distinto de cero con un p-valor ajustado menor a 0.05. De nuevo, el signo del logFC indica si el gen está sobreexpresado (+) o infraexpresado (-) en el contraste correspondiente.

3.2.8 Análisis de abundancia celular

En este tipo de análisis se pretende conocer si hay algún cambio en la composición celular del tejido entre dos grupos. Se llevaron a cabo los mismos contrastes que en el análisis de expresión diferencial: el contraste mujer, el contraste hombre y el contraste de diferencias de sexo. Para ello, se usó el método QL del paquete edgeR, que utiliza un modelo NB GLM, apropiado para tratar datos de conteos muy dispersos y con pocas réplicas, siendo los conteos en este caso células por tipo celular en vez de lecturas por gen. El abordaje es el mismo que el utilizado en el análisis de expresión diferencial con edgeR. Solo que en este caso se cuantifica el número de células de cada tipo celular, en este caso neuronas, células de la microglía, astrocitos, oligodendrocitos, OPCs y células endoteliales, en cada sujeto y no es necesario el paso de normalización. Una vez evaluadas las diferencias de abundancia entre los grupos definidos en el contraste correspondiente, se consideraron significativos aquellos tipos celulares en los que el logFC de la comparación correspondiente es distinto de cero con un FDR del 5%[29].

3.3 Comparación e integración de los resultados individuales

3.3.1 Intersecciones de los resultados individuales

Una vez analizados los estudios por separado, se evaluó la señal común detectada en los tres estudios, con el objetivo de identificar los resultados de mayor robustez y consenso. Tras el análisis de expresión diferencial se han obtenido 6 listas de genes por cada tipo celular de cada estudio, correspondientes a los genes sobreexpresados o infraexpresados en cada uno de los tres contrastes.

En primer lugar se integraron los resultados por tipo celular, contraste y signo del logFC. Se realizaron las intersecciones de cada una de las listas en los tres estudios, obteniendo un total de 12 listas de genes comunes, 6 para microglía y 6 para astrocitos. A continuación, se exploraron los resultados comunes entre los distintos contrastes y entre los distintos tipos celulares. Las distintas combinaciones pueden aportar información interesante. Por ejemplo, si hacemos la intersección entre los genes con logFC positivo del contraste de mujer y del contraste de hombre podemos detectar genes sobreexpresados en Alzheimer en ambos sexos. En cada una de las intersecciones realizadas se generó una lista de genes y un diagrama de Venn.

3.3.2 Caracterización funcional

En primer lugar, se analizó si los genes comunes detectados presentaban alguna asociación descrita previamente a EA. Para ello se descargaron de las bases de datos OPEN TARGETS[57] y NCBI[58] las listas de genes asociados a EA y se determinó cuántos y cuáles de los genes detectados en este estudio estaban presentes en estas listas.

A continuación, se realizó una búsqueda de los procesos biológicos de la GO asociados a los genes comunes del tercer contraste, que es el que nos permite detectar diferencias de sexo en la EA. Si el número de genes de la lista era reducido se buscaron los términos asociados a cada gen en la base de datos GENECARDS[59]. Si el número de genes de la lista era elevado se utilizó la herramienta Panther[60] para llevar a cabo un análisis de sobrerrepresentación y se utilizó la herramienta web STRING[61] para identificar las

relaciones descritas entre las proteínas codificadas por los distintos genes, en el caso de ser codificantes.

Por último, se utilizó la herramienta STRING para la caracterización funcional de los resultados de los contrastes individuales en hombres y mujeres comunes en ambos sexos.

4. Resultados

4.1 Revisión sistemática y selección de estudios

La revisión sistemática se llevó a cabo aplicando los criterios de inclusión y exclusión detallados en el apartado de “Materiales y métodos” (ver Tabla 1). De los 40 estudios identificados inicialmente, se eliminaron 11 por no estar centrados en la EA, 12 por no utilizar abordajes de transcriptómica de célula única o núcleo único y 10 por no encajar con el diseño experimental buscado. Finalmente, se tuvo que modificar el criterio de exclusión por tamaño muestral por la baja cantidad de estudios que pasaban este filtro. El filtro se cambió de un mínimo de tres muestras por grupo experimental a dos. De los 7 estudios restantes se seleccionaron 3 para este trabajo, por ser de libre acceso y haber usado el mismo método de secuenciación basado en gotas *10xGenomics* (Figura 10, Tabla 3).

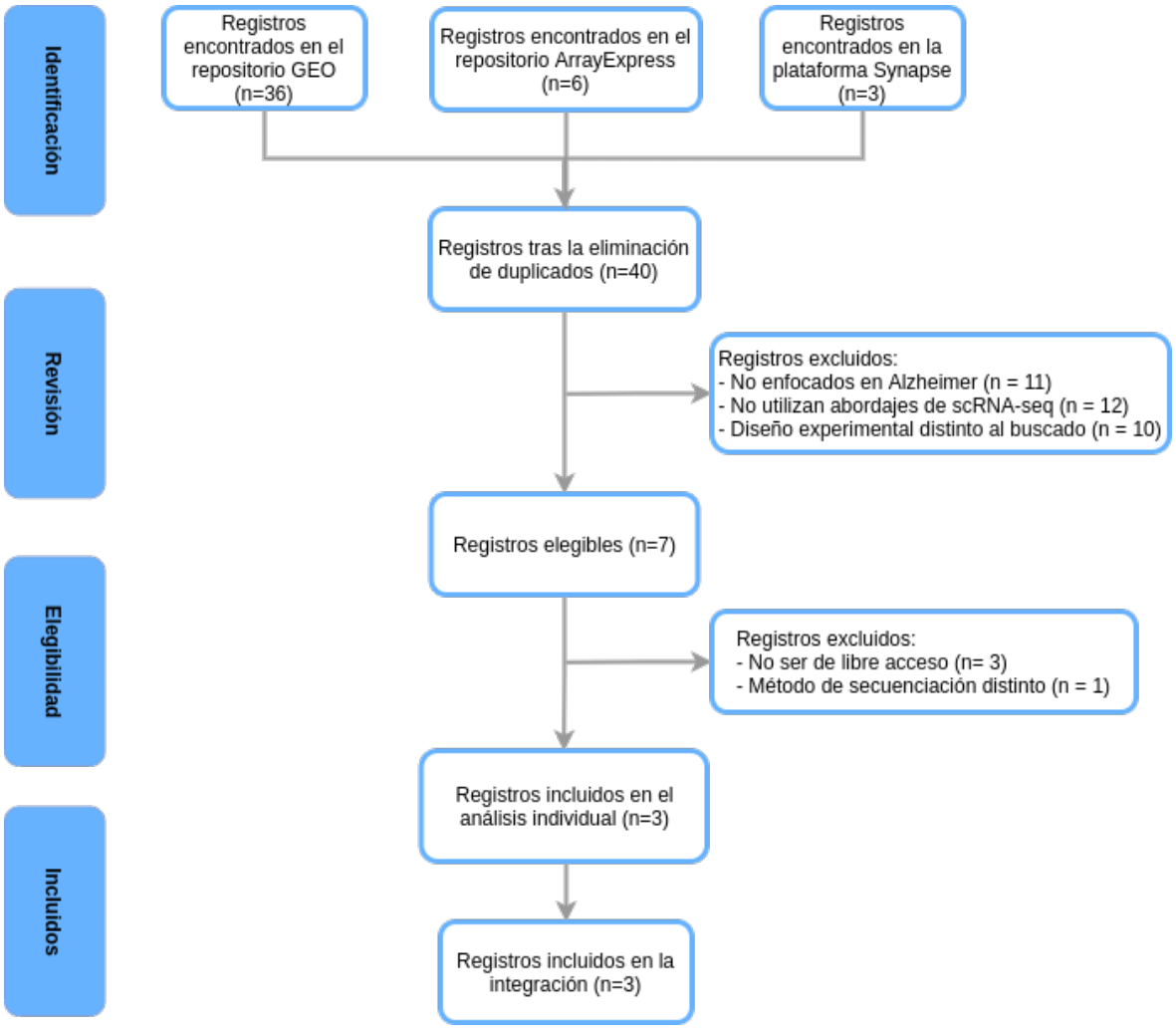

Figura 10: Diagrama de flujo PRISMA de la revisión sistemática llevada a cabo en este trabajo. Imagen propia creada a partir del diagrama Prisma de referencia[40,62].

Tabla 3: Estudios seleccionados tras la revisión sistemática. Se indica el método y la plataforma de secuenciación utilizada, la región cerebral a la que pertenecen las muestras, los tipos celulares analizados, el número de células y genes de partida, y el artículo de cada estudio.

Estudio (GEO series)	GSE138852	GSE157827	GSE160936
Método de secuenciación	*10xGenomics*	*10xGenomics*	*10xGenomics*
Plataforma de secuenciación	Illumina NextSeq 500	Illumina NovaSeq 6000	Illumina HiSeq 4000
Región cerebral	Corteza entorrinal	Corteza prefrontal	Corteza somatosensorial y corteza entorrinal
Tipos celulares	Todos	Todos	Microglía y Astrocitos
Número de células	13214	179392	101906
Número de genes	10850	33538	58929
Artículo	48	63	64

Todos los estudios seleccionados presentaron información de sexo de los sujetos, así como representación de ambos sexos tanto en el grupo control como en el grupo de pacientes con EA (Figura 11).

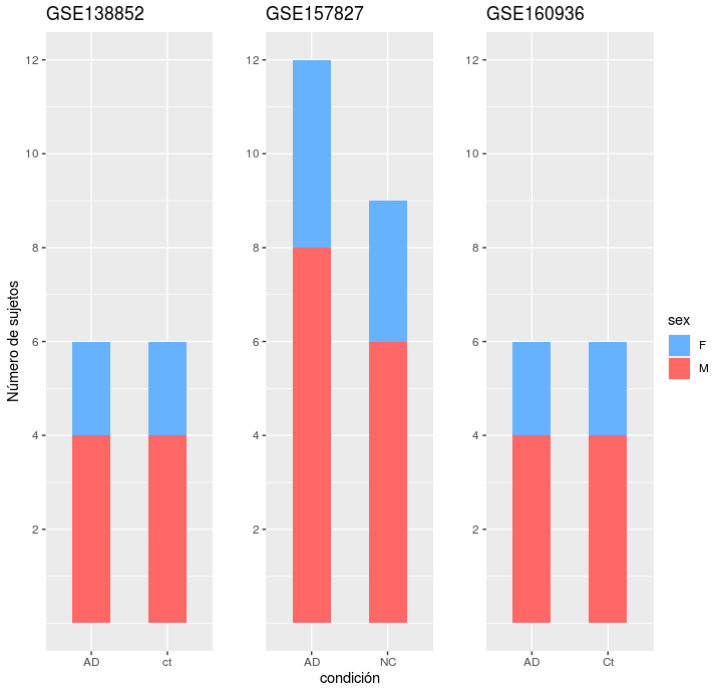

Figura 11: Distribución de las muestras por sexo, en los grupos control y EA en cada uno de los estudios. El sexo se denota como F y M para Mujeres y Hombres respectivamente, y la condición como AD para enfermos de Alzheimer y ct, NC y Ct para controles.

A pesar de los aspectos específicos de cada estudio, existen varias características comunes en los estudios seleccionados: a) utilizan un abordaje de transcriptómica de núcleo único, b) emplean el mismo método de secuenciación (*10xGenomics*), c) contienen muestras de la corteza cerebral tanto de pacientes de Alzheimer como de controles y d) detallan información del sexo de los mismos. Por todo ello, consideramos adecuada esta selección de estudios para llevar a cabo un análisis integrativo y tratar de responder la hipótesis planteada en este trabajo. En principio nos centramos en el estudio de la microglía y los astrocitos debido a su importante papel en la neuroinflamación, un proceso clave en la EA y otras enfermedades neurodegenerativas, con un importante potencial en la caracterización de estas patologías. Sin embargo, el análisis podría extenderse a otros tipos celulares como las neuronas o los oligodendrocitos.

4.2 Análisis individual de los estudios

4.2.1 Procesamiento y exploración de los datos

En esta sección se describen los resultados intermedios más relevantes obtenidos durante el procesamiento y la exploración de los datos. Como se llevaron a cabo los mismos pasos en los tres estudios, se describirań los resultados de los tres estudios, pero solo se mostrarán a modo de ejemplo las figuras del estudio GSE157827, salvo que se quiera resaltar alguna peculiaridad de otro estudio en concreto. El resto de figuras están disponibles en ZENODO, en el anexo III del siguiente enlace (https://zenodo.org/record/5457264)

En primer lugar, los datos de cada estudio necesarios para el análisis se descargaron del repositorio GEO. En el estudio GSE138852 partimos de la matriz de conteos, obtenida por los autores tras llevar a cabo el control de calidad. Como los autores construyeron las librerías por cada dos sujetos de la misma condición experimental, en algunas células no fueron capaces de determinar el paciente de procedencia. Estas células fueron eliminadas para nuestro análisis por la imposibilidad de asociar un sexo a las mismas. El número final de células utilizadas fue 12770. En los otros dos estudios se partió de los ficheros de salida del programa Cell Ranger.

Una vez importados los datos al entorno de R, se exploró y representó gráficamente el número de células, así como las tres covariables asociadas al control de calidad (número de conteos, número de genes detectados y proporción de genes mitocondriales) por sujeto para cada estudio.

Al analizar el número de células por paciente se puede observar una alta variabilidad entre sujetos, sin estar asociado el número de células al grupo experimental (Figura 12).

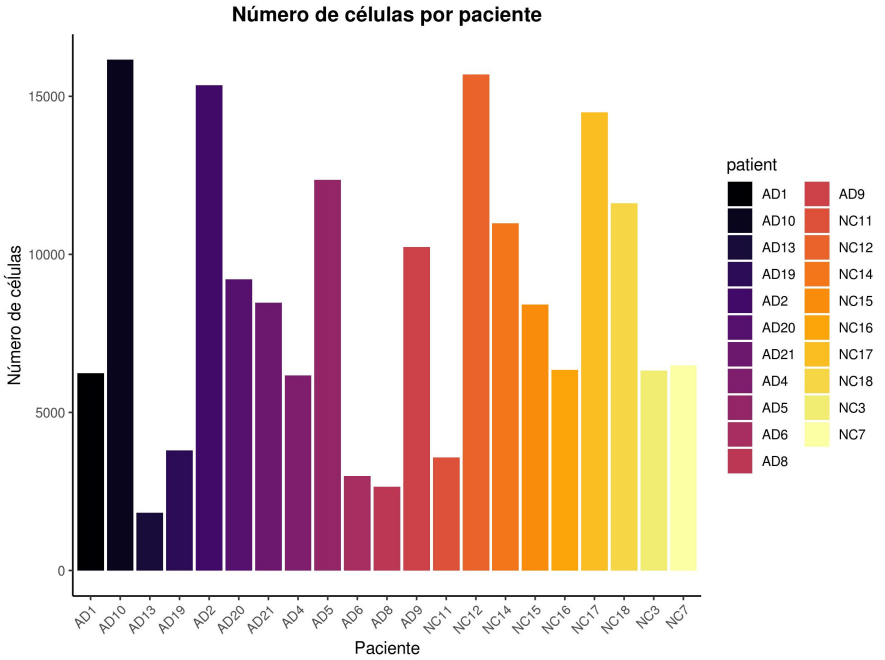

Figura 12: Diagrama de barras que representa el número de células por paciente del estudio GSE157827. AD y NC denotan pacientes de Alzheimer y controles respectivamente.

El resto de covariables se representaron en diagramas de violín, en los que cada punto representa una célula única. En el eje X se representan los sujetos y en el eje Y la covariable de interés. En estos diagramas se pueden explorar los datos para encontrar células con valores atípicos en alguna covariable o alguna muestra con baja calidad derivada del tratamiento experimental de la misma (Figuras 13, 14 y 15).

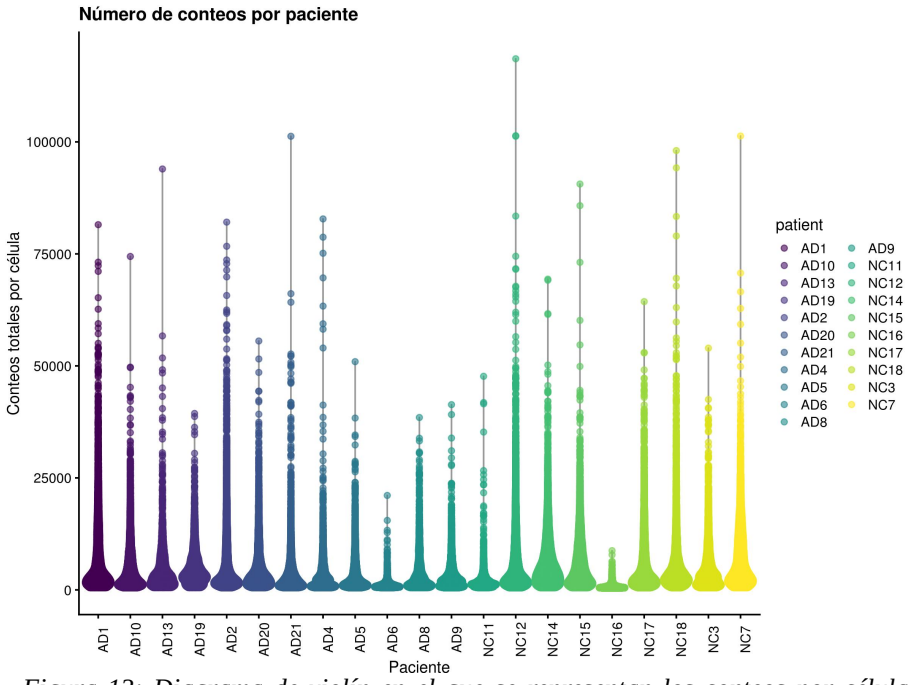

Figura 13: Diagrama de violín en el que se representan los conteos por célula y paciente del estudio GSE157827. Cada punto en el diagrama representa una célula. AD y NC denotan pacientes de Alzheimer y controles respectivamente.

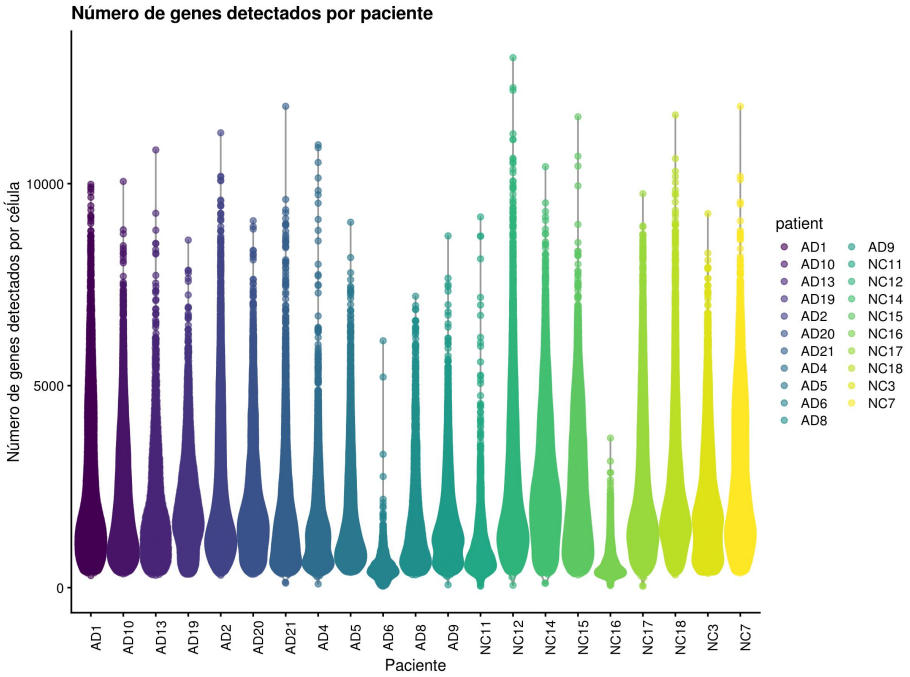

Figura 14: Diagrama de violín en el que se representa el número de genes detectados por célula y paciente del estudio GSE157827. Cada punto en el diagrama representa una célula. AD y NC denotan pacientes de Alzheimer y controles respectivamente.

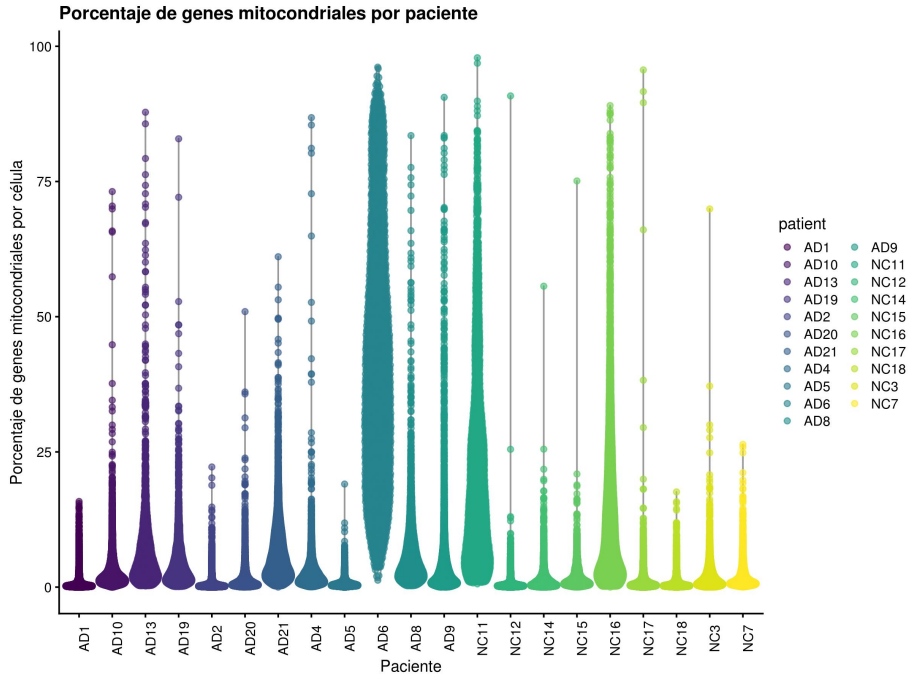

Figura 15: Diagrama de violín en el que se representa el porcentaje de genes mitocondriales por célula y paciente del estudio GSE157827. Cada punto en el diagrama representa una célula. AD y NC denotan pacientes de Alzheimer y controles respectivamente.

Tras analizar los gráficos se eliminó el paciente de Alzheimer número 6 del estudio GSE157827 (Figura 15) por su alto porcentaje de genes mitocondriales en la mayoría de sus células, lo cual puede ser indicativo de algún problema en el aislamiento de los núcleos, ya que en este tipo de abordajes no se deberían de encontrar genes mitocondriales.

Una vez explorados los datos se aplicó el control de calidad a todos los estudios, incluido GSE138852, para poder establecer los filtros en los tres estudios siguiendo el mismo criterio.

A continuación, se describen los resultados obtenidos al aplicar los filtros de calidad establecidos por la función isOutlier() (Tabla 2), ya que fue el abordaje finalmente elegido. Se eligió este abordaje porque al ver la distribución de tipos celulares en las muestras no se había perdido ningún tipo celular, que era la principal preocupación, confundir un tipo celular con células de baja calidad. Además, este abordaje al utilizar umbrales adaptativos se ajusta más a las peculiaridades de cada estudio. Por último, los porcentajes de genes mitocondriales permitidos establecidos por este abordaje fueron mucho menores. El umbral elegido en el otro abordaje del filtros comunes (10%) puede que sea demasiado alto para tratarse de estudios de núcleo único, ya que en estos casos el porcentaje de genes mitocondriales debería ser muy cercano a cero.

Finalizado el control de calidad, se volvieron a representar estas variables agrupando los pacientes por condición y sexo y coloreando en naranja aquellas células que fueron finalmente eliminadas por alguno de los criterios (incluso las eliminadas por ser identificadas como dobletes) (Figura 16). En general la mayoría de células fueron eliminadas por un alto porcentaje mitocondrial, aunque también se eliminó una alta cantidad por ser detectadas como dobletes (Tabla 4).

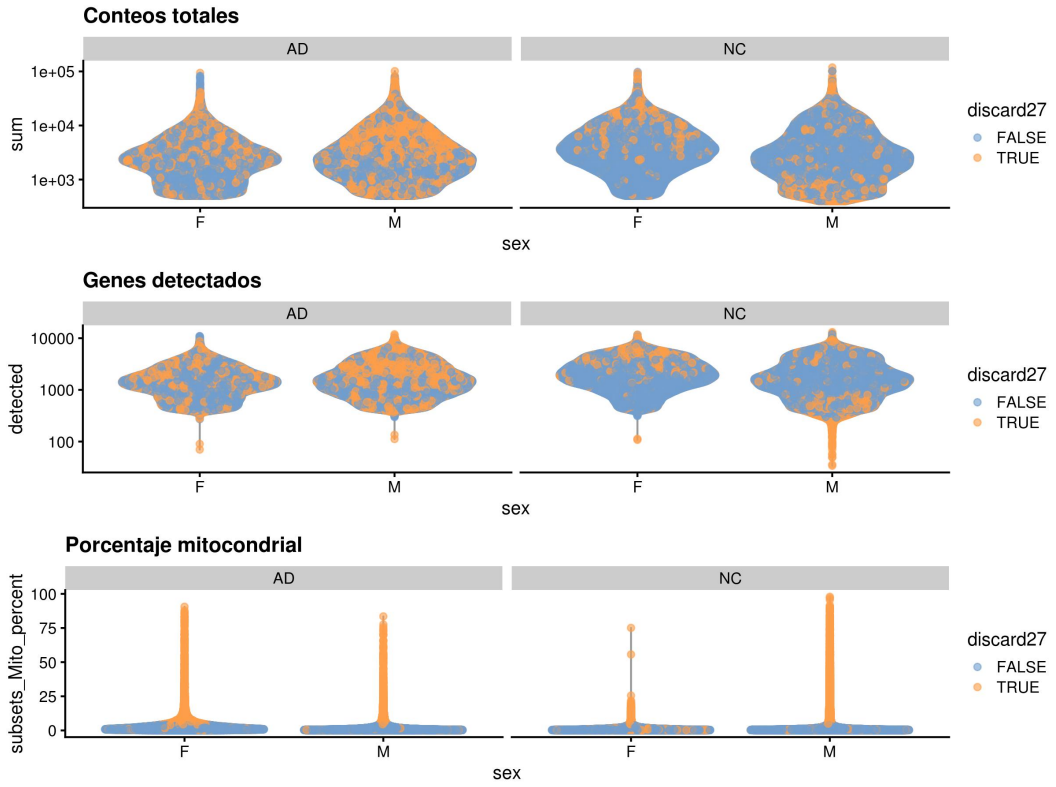

Figura 16: Diagramas de violín en los que se representan las tres covariables típicas del control de calidad: el número de conteos, el número de genes detectados y el porcentaje mitocondrial, por célula del estudio GSE157827. Cada punto representa una célula y el color de la misma indica si finalmente esa célula fue eliminada por algunos de los filtros, incluidas las eliminadas por ser identificadas como dobletes. Las células se agruparon por sexo y condición experimental. El color naranja indica que la célula fue eliminada y el azul que la célula fue conservada. Las siglas F y M denotan mujeres y hombres respectivamente, así como las siglas AD y NC denotan los pacientes y controles respectivamente.

Tabla 4: Número de células eliminadas por cada uno de los filtros del control de calidad.

Filtro del control de calidad	GSE138852	GSE157827	GSE160936
Tamaño librería	0	0	0
Número de genes	0	32	4064
% Genes mitocondriales	1454	26535	15480
Dobletes	274	19607	8323

Por último en la Tabla 5, se resume la evolución del número inicial de células y genes de partida, y los valores finales, tras aplicar los filtros de calidad. Este último número de células y genes serán los utilizados para los análisis posteriores de cada estudio.

Tabla 5: Número de células y genes que componen cada uno de los estudios antes y después del control de calidad.

Estudio (GEO series)	Número de células iniciales	Número de genes iniciales	Número de células finales	Número de genes finales
GSE138852	12770	10850	11064	8734
GSE157827	179392	33538	132429	18743
GSE160936	101906	58929	77517	21496

Los siguientes resultados a resaltar son los obtenidos en el paso de reducción de la dimensionalidad. El número de componentes principales necesarios para capturar la heterogeneidad de los datos, determinado por el método del codo, fue 7 en el estudio GSE138852, 6 en el estudio GSE157827 (Figura 17) y 3 en el estudio GSE160936.

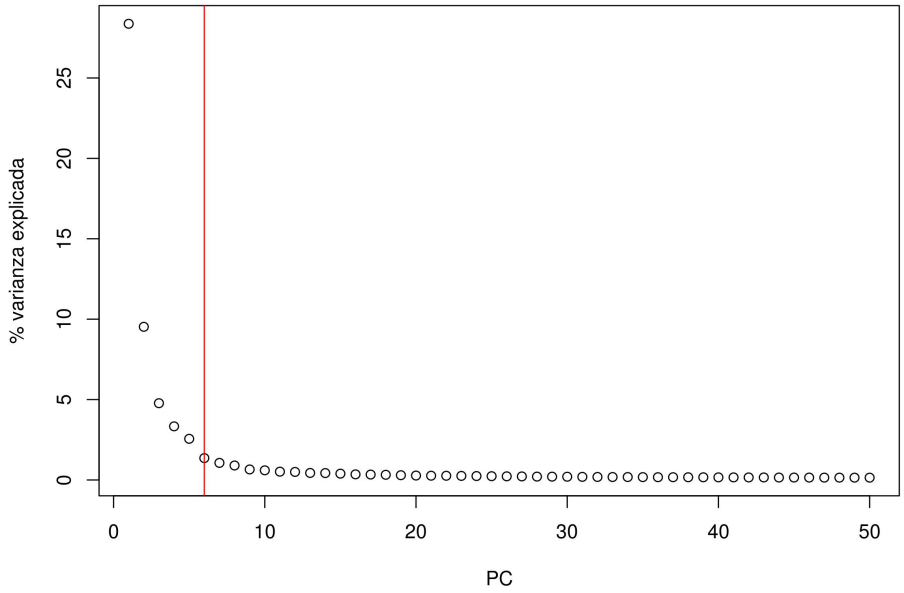

Figura 17: Gráfico del método del codo por el cual se identifican el número de componentes principales necesarios para capturar la heterogeneidad de los datos. En este caso del estudio GSE157827 serían necesarias 6 componentes principales.

Una vez llevado a cabo el paso de reducción de la dimensionalidad, se representaron gráficamente las dos primeras componentes principales coloreando las células por la variable que combina el sexo y la condición experimental (Figura 18). En ninguno de los estudios se observa una separación de las muestras según la covariable representada. Esto se debe a que las mayores fuentes de variabilidad son el sujeto de procedencia y el tipo celular.

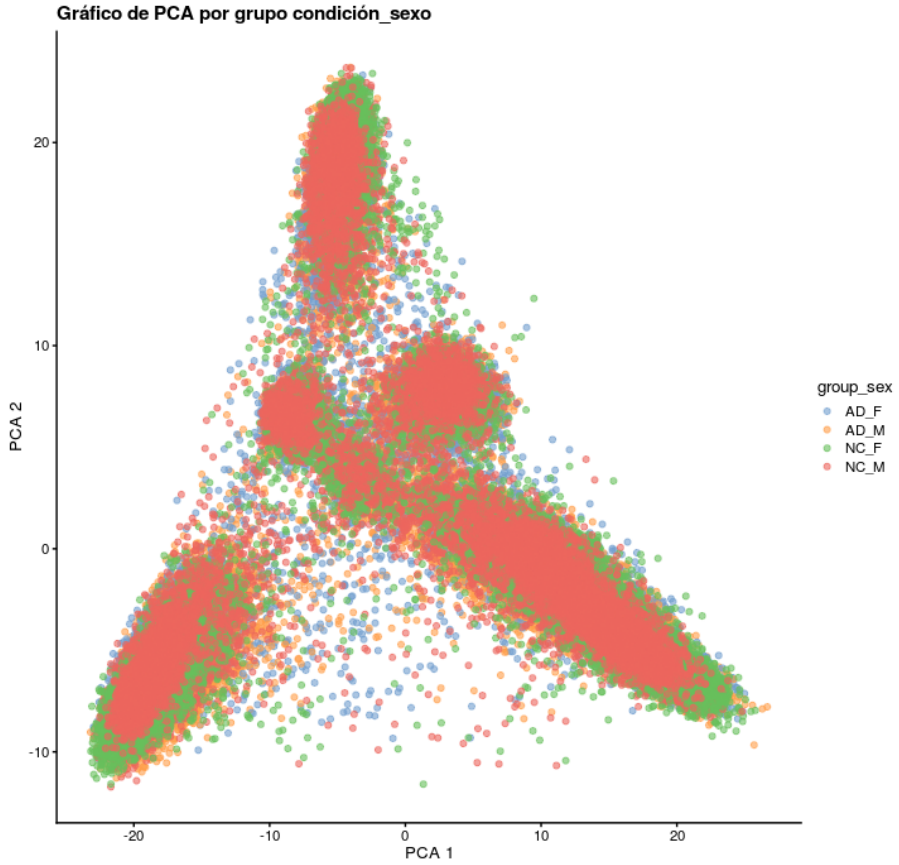

Figura 18: Representación de las 2 primeras componentes principales del estudio GSE157827. El color de las células viene determinado por el grupo que separa a las muestras según el sexo y la condición experimental, al que pertenecen.

A partir de los datos reducidos dimensionalmente se llevó a cabo el paso de *clustering* o agrupamiento de las células por similitud de expresión génica. Para ello se usó el mismo número de vecinos (K=20) para todos los estudios, aunque al ser tan diferentes el número de células y el número de genes de cada estudio, el número de *clusters* obtenidos es distinto (Tabla 6).

Tabla 6: Número de células, de genes y de clusters obtenidos en cada estudio usando una K igual a 20.

	GSE138852	GSE157827	GSE160936
Número de células	11064	132429	77517
Número de genes	8734	18743	21496
Número de *clusters*	16	24	28

La determinación del número adecuado de *clusters* dependerá de la resolución que se quiera conseguir, si el objetivo es obtener un *cluster* por tipo celular o si se quieren explorar los *subclusters* dentro de los mismos. Para ver la distribución de estos *clusters* se volvió a hacer una representación de las dos primeras componentes principales, pero esta vez coloreando las células según el *cluster* al que pertenezcan (Figura 19).

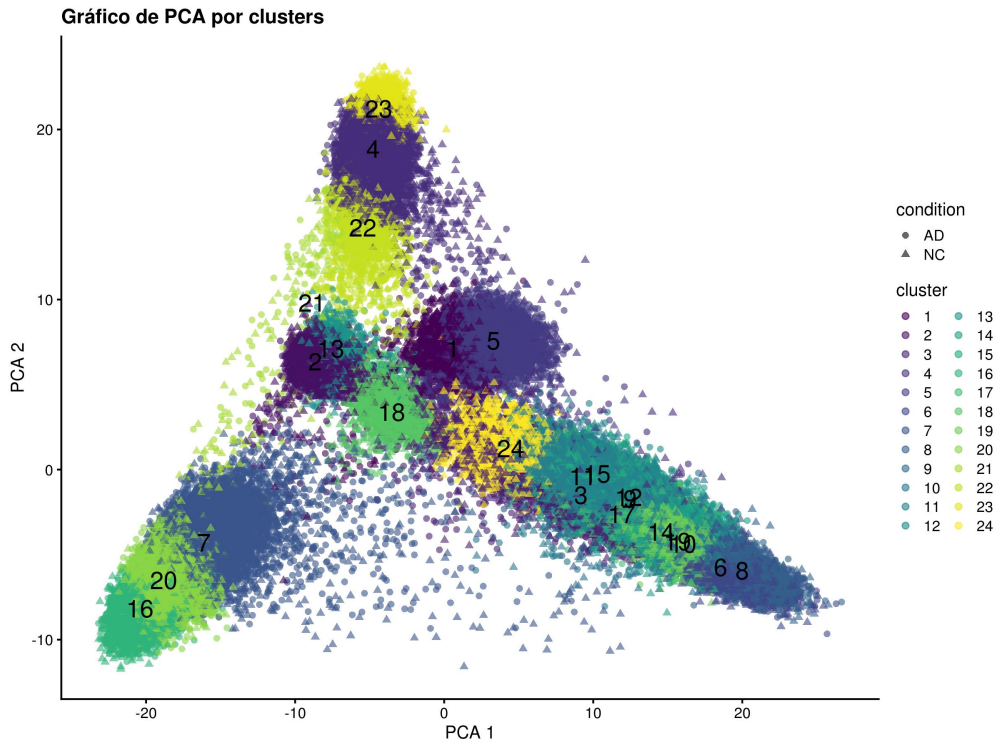

Figura 19: Representación de las 2 primeras componentes principales del estudio GSE157827. El color de las células viene determinado por el cluster al que pertenecen. Las células de pacientes de EA se representan con círculos y las de controles con triángulos.

A continuación, en la Tabla 7 se muestra el resultado de la identificación y la anotación del tipo celular, un paso fundamental en este tipo de abordajes, ya que nos permitirá realizar los análisis de expresión diferencial de cada tipo celular por separado, y por tanto revelar las peculiaridades de cada tipo celular.

Tabla 7: Número de células de cada tipo celular identificadas en cada estudio por el paquete BRETIGEA, así como el porcentaje del total de células que representa cada tipo celular.

	GSE138852		GSE157827		GSE160936	
Tipo celular	Número	%	Número	%	Número	%
Microglía	678	6,13	6854	5,18	24294	31,34
Astrocitos	1878	16,97	15213	11,49	20621	26,60
Neuronas	979	8,85	54360	41,05	2786	3,59
Oligodendrocitos	4670	42,21	37378	28,22	11615	14,98
OPCs	1289	11,65	10795	8,15	14816	19,11
Endoteliales	1570	14,19	7829	5,91	3385	4,37
Totales	11064	100	132429	100	77517	100

El porcentaje de células de cada tipo celular detectado varía de un estudio a otro. En el estudio GSE138852 destaca el número de oligodendrocitos, representando un 42% del número total de células. Sin embargo, en el estudio GSE157827 el tipo celular predominante es la neurona (41%). Estos porcentajes no coinciden con los descritos en la bibliografía. En el cortex cerebral de *Homo sapiens* sano se esperaría encontrar un 25% de neuronas, un 50% de células gliales y un 25% de células endoteliales[65]. Sin embargo, al contener estos estudios distintas condiciones experimentales y distinto número de sujetos por condición, se podrían ver modificados los porcentajes de estos tipos celulares. Tras el análisis de abundancia diferencial se podrá ver si los porcentajes de cada tipo celular obtenidos por los distintos estudios se ven modificados por la EA y/o el sexo, o si estas diferencias se deben al procedimiento experimental de las muestras llevado a cabo por los autores, los pasos de control de calidad y anotación del tipo celular realizados en este trabajo o a una combinación de varios de estos factores. Por último, en el estudio GSE160936 predominan la microglía y los astrocitos, lo cual concuerda con el diseño experimental del estudio, ya que los autores utilizaron anticuerpos específicos contra neuronas y oligodendrocitos para enriquecer sus muestras en los tipos celulares de interés.

Una vez obtenidos los tipos celulares, se volvieron a representar las dos primeras componentes principales, coloreando las células por tipo celular. En esta representación sí se ven los distintos tipos celulares separados. Además, al analizar los tipos celulares presentes en cada *cluster* se observó que en la mayoría de los *clusters* predominaba un único tipo celular, lo que resalta las diferencias genéticas entre tipos celulares, justificando el uso de este tipo de abordaje.

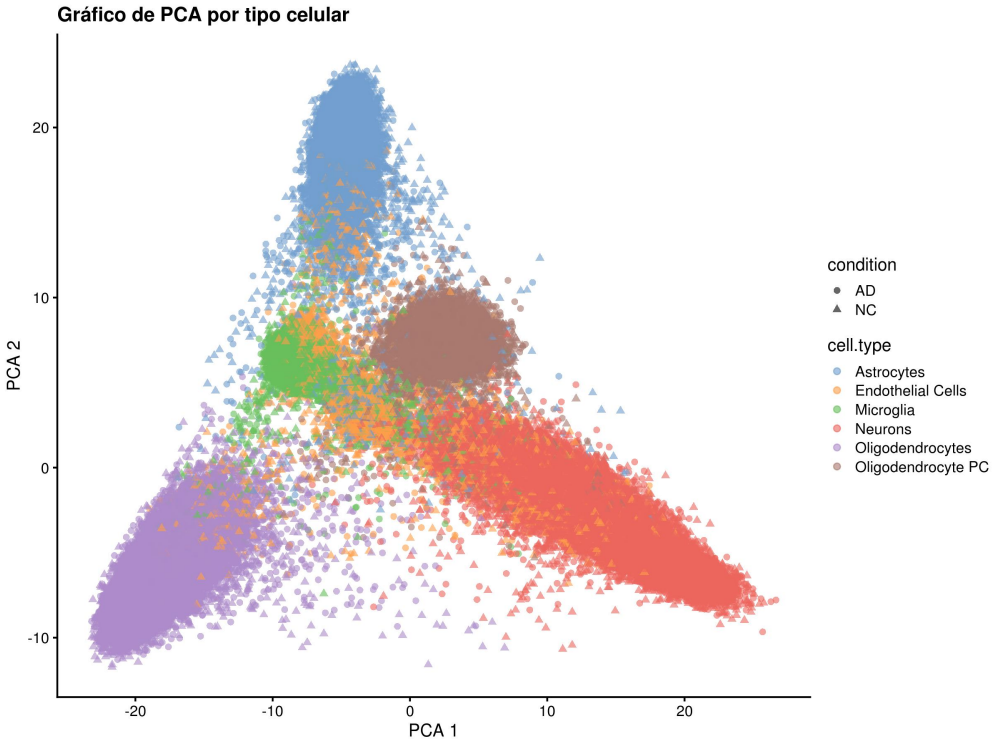

Figura 20: Representación de las 2 primeras componentes principales del estudio GSE157827. El color de las células viene determinado por el tipo celular al que pertenecen. Las células de pacientes de EA se representan con círculos y las de controles con triángulos.

4.2.2 Análisis de expresión diferencial

En el análisis de expresión diferencial realizado con el paquete de edgeR solo se obtuvieron genes significativos en el primer estudio (Tablas 8 y 9).

Tabla 8: Número de genes diferencialmente expresados detectados en las células de la microglía en los distintos estudios al realizar el análisis de expresión diferencial con el paquete edgeR. Descripción de contrastes: Mujer EA – Mujer Control (Mujer); Hombre EA – Hombre Control (Hombre); (Mujer EA – Mujer Control) - (Hombre EA – Hombre Control) (Diferencias de sexo).

Microglía				
Estudio (GEO series)	Contraste mujer	Contraste hombre	Contraste diferencias de sexo	Genes de partida
GSE138852	9	294	42	1520
GSE157827	0	0	0	10072
GSE160936	0	0	0	18272

Tabla 9: Número de genes diferencialmente expresados detectados en astrocitos en los distintos estudios al realizar el análisis de expresión diferencial con el paquete edgeR. Descripción de contrastes: Mujer EA – Mujer Control (Mujer); Hombre EA – Hombre Control (Hombre); (Mujer EA – Mujer Control) - (Hombre EA – Hombre Control) (Diferencias de sexo).

Astrocitos				
Estudio (GEO series)	Contraste mujer	Contraste hombre	Contraste diferencias de sexo	Genes de partida
GSE138852	132	1518	377	5774
GSE157827	0	0	0	13184
GSE160936	0	0	0	19983

La falta de señal significativa en dos de los tres estudios llevó a la revisión de la estrategia seguida. Cabe recordar que para este abordaje se sumaron los conteos de todas las células del mismo tipo en cada paciente, obteniendo una pseudo-célula de cada tipo celular por cada paciente. Sin embargo, el número de conteos de cada paciente podría estar influenciado por el número de células presentes en el mismo y no por el nivel de expresión de los genes. Para comprobarlo, se representó el número de células por pacientes y se hizo un gráfico de escalado multidimensional (MDS) de los perfiles de expresión de las pseudo-células para detectar si las células se separan por condición y/o sexo o si se agrupan por el número de células sumadas (Figuras 21 y 22).

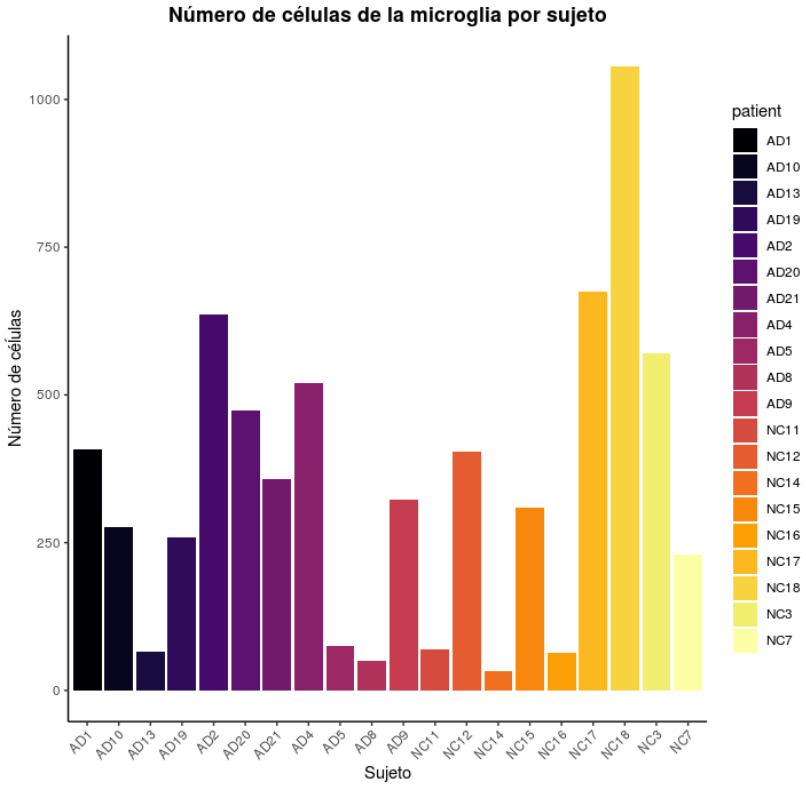

Figura 21: Número de células de microglía por sujeto en el estudio GSE157827. AD y NC denotan pacientes de Alzheimer y controles respectivamente.

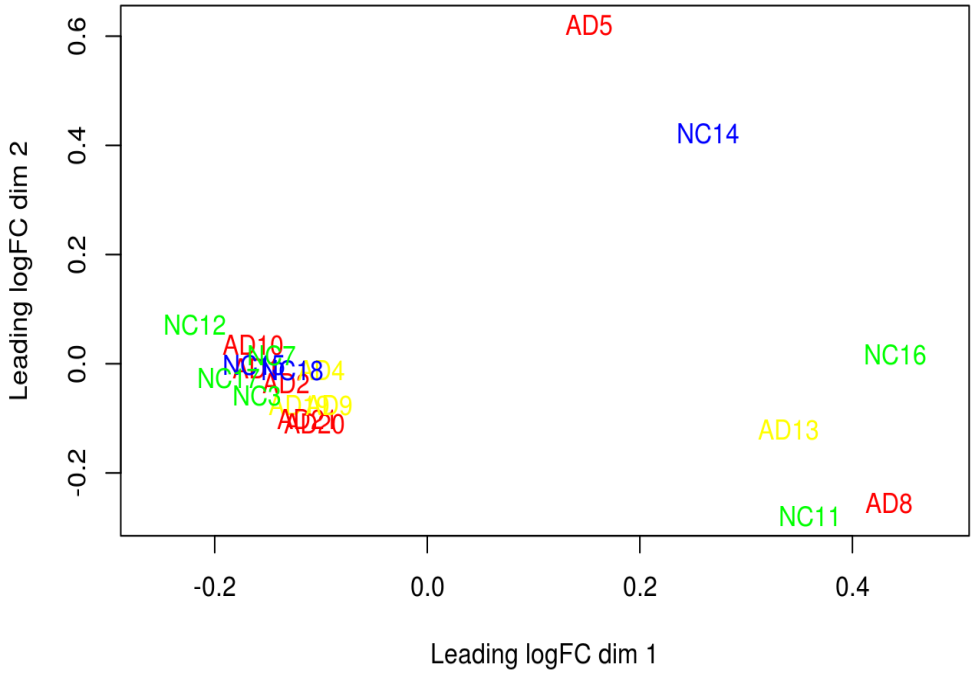

Figura 22: Gráfico de escalado multidimensional (MDS) de los perfiles de expresión de las pseudo-células de microglía por sujeto en el estudio GSE157827. Cada pseudo-célula se representa con el identificador del sujeto al que pertenece, coloreado en función de su sexo y condición experimental. En rojo se representan los hombres con EA, en amarillo las mujeres con EA, en verde los hombres control y en azul las mujeres control.

En los gráficos de barras se puede observar que el número de células por paciente es extremadamente variable y no parace estar relacionado con las covariables condición experimental y/o sexo (Figura 21). Además, las pseudo-células no se agrupan en función de los grupos experimentales establecidos por la variable que combina condición y sexo. Por ejemplo, en el estudio GSE157827 se observa una clara separación entre las pseudo-células obtenidas según el número de células sumadas (Figura 22).

Como el número de células de cada paciente es muy distinto y no parece estar asociado a la condición experimental y/o al sexo, y el número de células afecta a los resultados, se optó por hacer una modificación del abordaje, eligiendo en cada estudio el mismo número de células (seleccionadas de forma aletaoria) de cada paciente, fijado por el sujeto que tuviese menor número de células. De este modo en todos los pacientes se sumaron los conteos del mismo número de células (Tabla 10).

Tabla 10: Número de genes diferencialmente expresados detectados en microglía en los distintos estudios al realizar el análisis de expresión diferencial con el paquete edgeR, sumando el mismo número de células por paciente. Descripción de contrastes: Mujer EA – Mujer Control (Mujer); Hombre EA – Hombre Control (Hombre); (Mujer EA – Mujer Control) - (Hombre EA – Hombre Control) (Diferencias de sexo).

Estudio (GEO series)	Contraste mujer	Contraste hombre	Contraste diferencias de sexo	Genes de partida	Número de células
GSE138852	0	17	7	128	11
GSE157827	0	0	0	1924	33
GSE160936	0	0	0	15524	855

El número de genes significativos detectados se redujo aún más. Esto podría deberse a que el número de células seleccionadas era demasiado bajo como para que la pseudo-célula resultante se asemeje a datos de bulk RNA-seq y se puedan aplicar métodos propios de este tipo de abordaje. Si el número de muestras de los estudios fuese mayor se podrían eliminar aquellos sujetos que tuvieran un número de células demasiado bajo, pero el número de sujetos en este tipo de experimentos es bajo y además, al ser necesario un número de muestras representativas de cada sexo en cada condición, es muy complicado eliminar un sujeto.

Por último, también se llevó a cabo un abordaje específico para datos de scRNA-seq con el paquete MAST. En este caso se identificaron numerosos genes significativos en todos los estudios (Tablas 11 y 12).

Tabla 11: Número de genes diferencialmente expresados detectados en microglía en los distintos estudios al realizar el análisis de expresión diferencial con el paquete MAST. Descripción de contrastes: Mujer EA – Mujer Control (Mujer); Hombre EA – Hombre Control (Hombre); (Mujer EA – Mujer Control) - (Hombre EA – Hombre Control) (Diferencias de sexo).

Microglía				
Estudio (GEO series)	Contraste mujer	Contraste hombre	Contraste diferencias de sexo	Genes de partida
GSE138852	98	788	228	8734
GSE157827	1464	1303	1380	18743
GSE160936	6207	9674	8905	21496

Tabla 12: Número de genes diferencialmente expresados detectados en astrocitos en los distintos estudios al realizar el análisis de expresión diferencial con el paquete MAST. Descripción de contrastes: Mujer EA – Mujer Control (Mujer); Hombre EA – Hombre Control (Hombre); (Mujer EA – Mujer Control) - (Hombre EA – Hombre Control) (Diferencias de sexo).

Astrocitos				
Estudio (GEO series)	Contraste mujer	Contraste hombre	Contraste diferencias de sexo	Genes de partida
GSE138852	488	2280	842	8734
GSE157827	9655	11901	11502	18743
GSE160936	3956	3415	2423	21496

Aunque un mayor número de genes detectados como diferencialmente expresados no tiene por qué ser indicativo de un mejor abordaje, se decidió continuar el análisis con los resultados del abordaje realizado con el paquete MAST por varios motivos: a) este abordaje es específico de este tipo de datos, b) no presenta el problema del número de células por paciente porque las células se analizan de forma individual y c) la señal biológica detectada es más parecida entre los tres estudios. Para evaluar la similitud global entre los estudios, se determinó el nivel de asociación entre todos ellos, mediante la determinación de la correlación entre los estadísticos del contraste de diferencias de sexo de los genes coincidentes en los tres estudios para los tres abordajes. En las Tablas 13, 14 y 15 se detallan los resultados.

Tabla 13: Correlación entre los distintos estudios, a partir de los niveles de expresión diferencial obtenidos con el paquete edgeR. En los cuadros rosas se muestran los coeficientes de correlación y en los azules los p-valores.

Estudios	GSE138852	GSE157827	GSE160936
GSE138852	1	0.0230441	-0.04207465
GSE157827	0.3806	1	0.007412405
GSE160936	0.1093	0.7779	1

Tabla 14: Correlación entre los distintos estudios, a partir de los niveles de expresión diferencial obtenidos con el paquete edgeR, sumando el mismo número de células en cada paciente. En los cuadros rosas se muestran los coeficientes de correlación y en los azules los p-valores.

Estudios	GSE138852	GSE157827	GSE160936
GSE138852	1	0.1250205	0.1693789
GSE157827	0.1793	1	0.03228878
GSE160936	0.0679	0.7296	1

Tabla 15: Correlación entre los distintos estudios, a partir de los niveles de expresión diferencial obtenidos con el paquete MAST. En los cuadros rosas se muestran los coeficientes de correlación y en los azules los p-valores.

Estudios	GSE138852	GSE157827	GSE160936
GSE138852	1	0.323786	0.264085
GSE157827	2.2e-16	1	0.2424143
GSE160936	2.2e-16	2.2e-16	1

Como se puede ver en la Tabla 13 las correlaciones en el abordaje con edgeR son bajas e incluso negativas y no significativas, lo que significa que no existe una señal común en los tres estudios. En el abordaje de edgeR utilizando el mismo número de células (Tabla 12), mejoran ligeramente las correlaciones, pero siguen siendo bajas y no significativas. Por último, las correlaciones en el abordaje con MAST (Tabla 15) son positivas, más altas y significativas, lo que indica una señal común, aportando robustez al abordaje. Por tanto, como se ha descrito anteriormente, el abordaje seleccionado para este estudio fue el del paquete MAST.

4.2.3 Análisis de abundancia diferencial

La única diferencia de abundancia se ha detectado en el estudio GSE138852 en el contraste de mujer. El análisis muestra una diferencia significativa de cantidad de células precursoras de oligodendrocitos con un FDR igual a 0,0011 y un logFC de -2,24, que indica que el número de OPC es mayor en las mujeres control respecto a las mujeres con EA. En el resto de contrastes y estudios no se han detectado diferencias de abundancia

significativas. El número de células de cada estudio en función del sexo y la condición experimental se describen en las Tablas 16, 17 y 18.

Tabla 16: Número y porcentaje de tipo celular por sexo y condición experimental del estudio GSE138852. Entre paréntesis se indica el número de sujetos de cada grupo.

GSE138852								
Grupo	Mujer EA (2)		Hombre EA (4)		Mujer control (2)		Hombre control (4)	
Células	n	%	n	%	n	%	n	%
Microglía	147	6,03	183	6,76	95	10,34	253	5,06
Astrocitos	207	8,50	333	12,3	123	13,38	1215	24,29
Neuronas	187	7,68	259	9,57	165	17,95	368	7,36
Oligodendrocitos	1305	53,57	1331	49,19	295	32,10	1739	34,76
OPCs	69	2,83	294	10,86	110	11,97	816	16,31
Endoteliales	521	21,39	306	11,31	131	14,25	612	12,23
Totales	2436	100	2706	100	919	100	5003	100

Tabla 17: Número y porcentaje de tipo celular por sexo y condición experimental del estudio GSE157827. *Entre paréntesis se indica el número de sujetos de cada grupo.*

GSE157827								
Grupo	Mujer EA (4)		Hombre EA (7)		Mujer control (3)		Hombre control (6)	
Células	n	%	n	%	n	%	n	%
Microglía	1167	7,97	2276	4,27	1398	5,38	2013	5,23
Astrocitos	2000	13,66	5882	11,04	2946	11,33	4385	11,39
Neuronas	4434	30,28	21354	40,07	12420	47,76	16152	41,97
Oligodendrocitos	4179	28,54	16600	31,15	5681	21,85	10918	28,37
OPCs	1552	10,60	4059	7,62	2228	8,57	2956	7,68
Endoteliales	1311	8,95	3124	5,86	1332	5,12	2062	5,36
Totales	14643	100	53295	100	26005	100	38486	100

Tabla 18: Número y porcentaje de tipo celular por sexo y condición experimental del estudio GSE160936. *Entre paréntesis se indica el número de sujetos de cada grupo.*

GSE160936								
Grupo	Mujer EA (2)		Hombre EA (4)		Mujer control (2)		Hombre control (4)	
Células	n	%	n	%	n	%	n	%
Microglía	3152	27,53	7888	30,93	4330	30,24	8924	34,00
Astrocitos	2982	26,05	7311	28,66	4658	32,53	5670	21,60
Neuronas	703	6,14	1045	4,10	532	3,72	506	1,93
Oligodendrocitos	1867	16,31	3477	13,63	1216	8,49	5055	19,26
OPCs	1869	16,32	4943	19,38	3125	21,82	4879	18,59
Endoteliales	876	7,65	840	3,29	459	3,21	1210	4,61
Totales	11449	100	25504	100	14320	100	26244	100

4.3 Comparación e integración de los resultados individuales

4.3.1 Intersecciones de los resultados individuales

Para identificar qué genes presentaban un patrón común en los tres estudios, se llevó a cabo un análisis de intersección de los genes significativos de los estudios individuales (Tablas 11 y 12), para los distintos escenarios de interés, según el tipo celular y contraste. El resumen de los resultados se muestra en la Tabla 19. Las listas completas de genes obtenidas están disponibles en ZENODO, en el anexo IV del siguiente enlace (https://zenodo.org/record/5457264)

Tabla 19: Número de genes significativos comunes en los tres estudios para cada uno de los contrastes. El signo positivo indica genes sobreexpresados y el negativo genes infraexpresados.

Tipo celular		Microglía	Astrocitos
Contraste mujer	+	5	56
	-	9	55
	Total	14	111
Contraste hombre	+	7	172
	-	22	126
	Total	29	298
Contraste diferencias de sexo	+	2	36
	-	4	23
	Total	6	59
Genes comunes		7941	

Los genes que nos permitirirán estudiar las diferencias de sexo en la enfermedad son los obtenidos en el tercer contraste o contraste de diferencias de sexo (Tabla 20). Sin embargo, los genes obtenidos en los otros dos contrastes también son interesantes, ya que nos informan de la enfermedad en mujeres y hombres de manera independiente. Asimismo los resultados de los contrastes individuales en hombres y mujeres, han permitido identificar genes comunes en ambos sexos, que constituyen potenciales biomarcadores de la EA.

Tabla 20: Genes expresados diferencialmente comunes en todos los estudios en el contraste de diferencias de sexo. El color rojo en los genes indica un logFC positivo (gen sobreexpresado), mientras que el color azul un logFC negativo (gen infraexpresado).

Contraste	Tipo celular	Genes significativos
Diferencias de sexo	Microglía	NEAT1, RGS16, CKB, GPC1, LINGO1, MT3
	Astrocitos	ADAMTS9-AS2, ADCY2, ADGRA3, AHNAK, ANOS1, BOC, BTBD9, DOCK4, DOCK7, DTNA, FAM189A2, FRMPD2, HSPB8, MACF1, MALAT1, MARCH3, MED13L, MIR4300HG, MPP6, NAV2, NEAT1, NFAT5, NPAS3, PARD3, PLCE1, PLEKHA5, PLPP1, PPFIA2, PPP1R12B, PRKG1, RANBP3L, REV3L, SNED1, SNX29, VCAN, WDR49, ATP1B1, C1orf61, CALM1, COX5B, CSTB, FABP7, GAPDH, GRIN1, GRM3, HNRNPA2B1, HS3ST5, HSPA9, LAMTOR4, OLIG1, PHLDB1, PRNP, RGCC, SCD, SDF4, TAGLN3, THY1, WIF1, ZMYM3
	Ambos	NEAT1

Las intersecciones de las distintas listas de genes obtenidas en los distintos tipos celulares, los distintos contrastes y con distinto signo de logFC pueden revelar información interesante. En primer lugar, se llevaron a cabo las intersecciones de las listas de genes sobreexpresados en los contrastes de mujer (A) y hombre (B) (logFC positivo) con las listas de genes detectados en el contraste de diferencias de sexo (C; logFC positivo, y D; logFC negativo) en microglía (Figura 23).

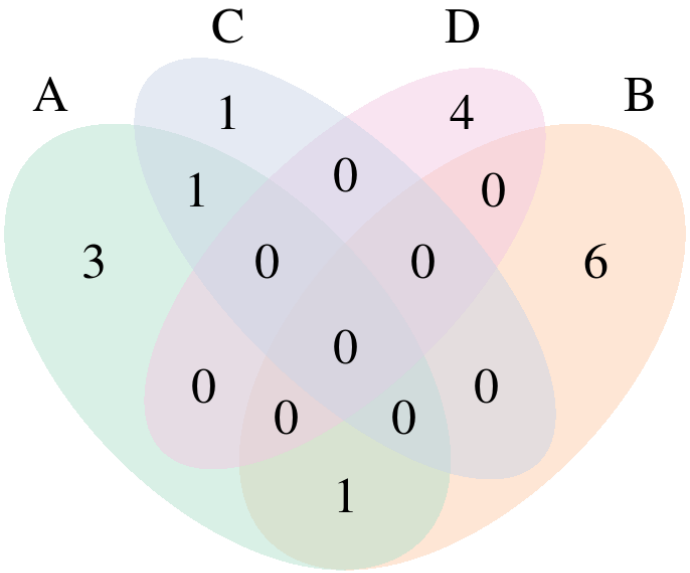

Figura 23: Diagrama de Venn que representa las intersecciones entre las listas de genes sobreexpresados en los contrastes de mujer (A) y hombre (B) y las listas de genes obtenidas en el contraste de diferencias de sexo (C; logFC +, y D; logFC -) en las células de la microglía.

Se detectó un gen común al intersectar las listas de genes sobreexpresados en los contrastes de mujer y hombre. Se trata del gen **DPYD**, un gen detectado en este análisis como sobreexpresado en pacientes con EA de ambos sexos. También, se detectó un gen común al intersectar los genes sobreexpresados en el contraste de mujer y los genes sobreexpresados en el contraste de diferencias de sexo, es decir que están sobreexpresados en mujeres con EA respecto a los hombres con EA, sin tener en cuentas las diferencias de sexo no debidas únicamente a la enfermedad. Se trata del gen **NEAT1**.

En segundo lugar, se llevaron a cabo las intersecciones de las listas de genes infraexpresados en los contrastes de mujer (A) y hombre (B) (logFC negativo) con las listas de genes detectados en el contraste de diferencias de sexo (C, logFC positivo, y D, logFC negativo) en microglía (Figura 24).

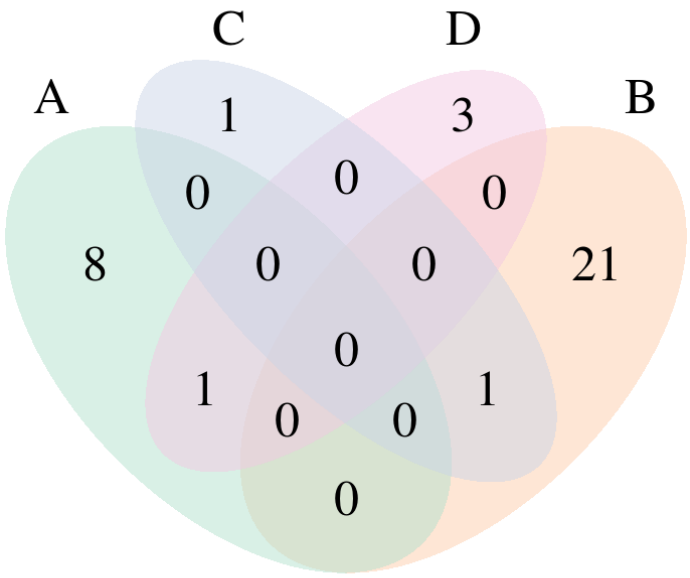

Figura 24: Diagrama de Venn que representa las intersecciones entre las listas de genes infraexpresados en los contrastes de mujer (A) y hombre (B) y las listas de genes obtenidas en el contraste de diferencias de sexo (C; logFC +, y D; logFC -) en las células de la microglía.

Se detectó un gen común al intersectar la lista de genes infraexpresados en el contraste de hombre, es decir sobreexpresados en los hombres controles y la lista de genes sobreexpresados en el contraste de diferencias de sexo, es decir que están sobreexpresados en mujeres con EA respecto a los hombres con EA, sin tener en cuenta las diferencias de sexo no debidas únicamente a la enfermedad. Se trata del gen **RGS16**, la detección de este gen en estos contrastes parece indicar que este gen está infraexpresado en hombres con EA con respecto a los hombres control, pero no en mujeres con EA, por eso al comparar los genes en el contraste de diferencias de sexo aparece como sobreexpresado en mujeres EA al compararlo con hombres. Se detectó un gen común al intersectar la lista de genes infraexpresados en el contraste de mujer, es decir sobreexpresados en mujeres controles y la lista de genes infraexpresados en el contraste de diferencias de sexo, es decir que están sobreexpresados en hombres con EA respecto a las mujeres con EA, sin tener en cuentas las diferencias de sexo no debidas únicamente a la enfermedad. Se trata del gen **CKB,** la detección de este gen en estos contrastes parece indicar que este gen está infraexpresado en mujeres con EA con respecto a las mujeres control, pero no en hombres con EA, por eso al comparar los genes en el contraste de diferencias de sexo aparece como sobreexpresado en hombres.

A continuación, se llevó a cabo el mismo procedimiento con los genes detectados en astrocitos, los resultados se encuentran en las Figuras 25 y 26 , y en la Tabla 21.

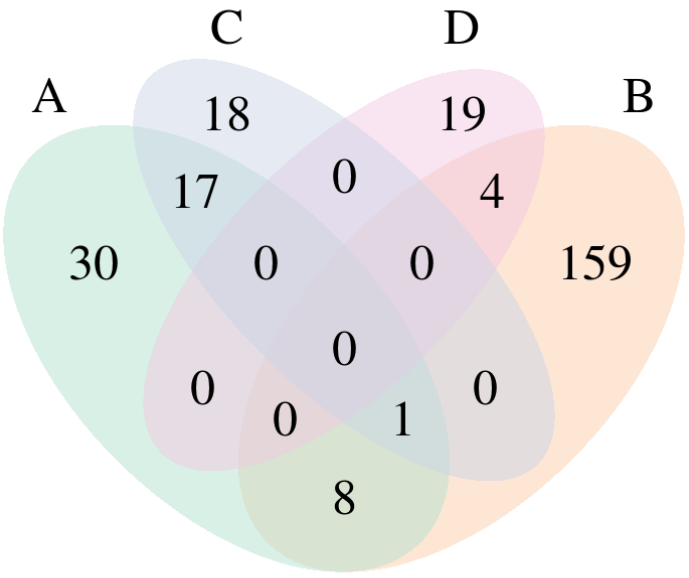

Figura 25: Diagrama de Venn que representa las intersecciones entre las listas de genes sobreexpresados en los contrastes de mujer (A) y hombre (B) y las listas de genes obtenidas en el contraste de diferencias de sexo (C; logFC +, y D; logFC -) en astrocitos.

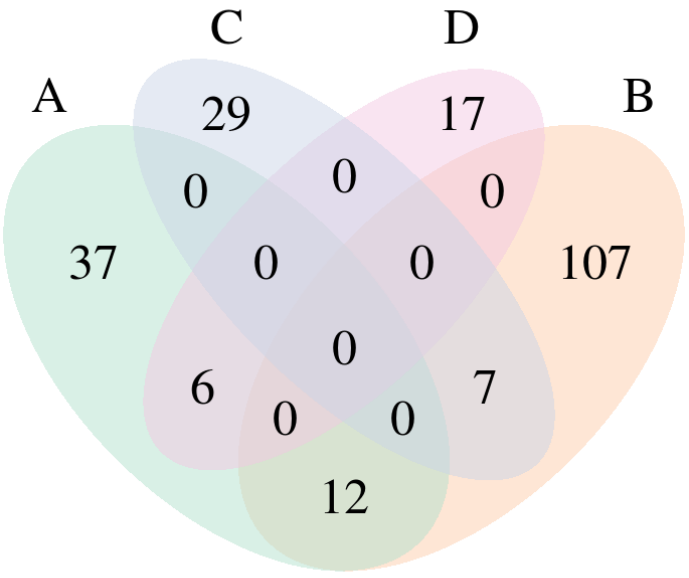

Figura 26: Diagrama de Venn que representa las intersecciones entre las listas de genes infraexpresados en los contrastes de mujer (A) y hombre (B) y las listas de genes obtenidas en el contraste de diferencias de sexo (C; logFC +, y D; logFC -) en astrocitos.

Tabla 21: Genes resultado de las intersecciones llevadas a cabo con las listas de genes obtenidas de astrocitos.

Figura	Intersección	Genes comunes
25	A∩B	ARFGAP1, CRB2, FAM107A, HSPB1, ITGB8, MRAS, PLXNB1, SLC7A2
	A∩C	ADAMTS9-AS2, ADCY2, AHNAK, DTNA, HSPB8, MALAT1, MARCH3, MIR4300HG, NAV2, PARD3, PLCE1, PLEKHA5, PLPP1, PPFIA2, PRKG1, SNED1, VCAN
	B∩D	HNRNPA2B1, HSPA9, RGCC, ZMYM3
	A∩B∩C	NEAT1
26	A∩B	ADGRV1, ARHGAP24, CSGALNACT1, FAM155A, GABRB1, GPC5, LSAMP, NHSL1, NTM, PTN, SLC1A2, SLC4A4
	B∩C	ANOS1, BTBD9, FAM189A2, FRMPD2, NPAS3, PRKG1, REV3L
	A∩D	ATP1B1, C1orf61, GRIN1, GRM3, THY1, WIF1

En la Figura 25 y primera parte de la Tabla 21, los genes de la intersección de las listas A y B son genes detectados como sobreexpresados en pacientes con EA de ambos sexos. Los de la intersección de las listas A y C son genes detectados como sobreexpresados en mujeres con EA respecto a mujeres control y respecto a hombres con EA. Los de la intersección de las listas B y D son genes detectados como sobreexpresados en hombres con EA respecto a hombres control y respecto a mujeres con EA. Por último, el gen NEAT1 encontrado en la intersección de las listas A, B y C estaría sobrerreprensentado tanto en hombres como en mujeres respecto a los controles de su mismo sexo, sin embargo, en mujeres con EA su nivel de expresión es más alto que en hombres con EA.

En la Figura 26 y segunda parte de la Tabla 21, los genes de la intersección de las listas A y B son genes detectados como infraexpresados en pacientes con EA de ambos sexos. Los de la intersección de las listas B y C son genes detectados como infraexpresados en hombres con EA con respecto a los hombres control, pero no en mujeres con EA. Por último, los de la intersección de las listas A y D son genes detectados como infraexpresados en mujeres con EA con respecto a las mujeres control, pero no en hombres con EA.

Aunque no es el objetivo principal del trabajo, la intersección de los contrastes de mujer y de hombre permiten identificar posibles marcadores de la EA no asociados al sexo (Tabla 22). Aproximadamente la mitad de estos genes están descritos para EA, proporcionando robustez al análisis llevado a cabo en este trabajo.

Tabla 22: Genes diferencialmente expresados en la EA no específicos de sexo detectados en el estudio. En azul los genes asociados previamente a la EA en las bases de datos de NCBI y Open Targets.

Tipo celular	Genes sobreexpresados en EA
Microglía	DPYD
Astrocitos	ARFGAP1, CRB2, FAM107A, HSPB1, ITGB8, MRAS, PLXNB1, SLC7A2
Tipo celular	**Genes infraexpresados en EA**
Astrocitos	ADGRV1, ARHGAP24, CSGALNACT1, FAM155A, GABRB1, GPC5, LSAMP, NHSL1, NTM, PTN, SLC1A2, SLC4A4

A priori podría pensarse que la comparación de los genes detectados en los contrastes de mujer y hombre, bastaría para revelar las diferencias de sexo en la EA. Sin embargo, como reflejan las intersecciones realizadas esto no se cumple en todos los casos. El contraste de diferencias de sexo permite detectar de forma más precisa las diferencias de sexo en la EA, ya que al tener en cuenta el contraste completo en el análisis de expresión diferencial, también se evalúa si la diferencia de expresión genica entre hombres y mujeres asociadas a la EA es significativa.

Por último, se llevó a cabo la intersección entre los resultados obtenidos en el contraste de diferencias de sexo para microglía y astrocitos, para saber si hay genes diferencialmente expresados comunes entre ambos tipos celulares. En el diagrama de Venn de la Figura 27 se muestra el resultado, los grupos A y B se corresponden con los genes sobre e infra expresados respectivamente, en microglía; mientras que los grupos C y D en astrocitos. Como se puede observar solo se ha detectado un gen común en ambos tipos celulares. Se trata de gen **NEAT1**, detectado en este análisis en ambos tipos celulares como sobreexpresado en mujeres con EA o infraexpresado en hombres con EA.

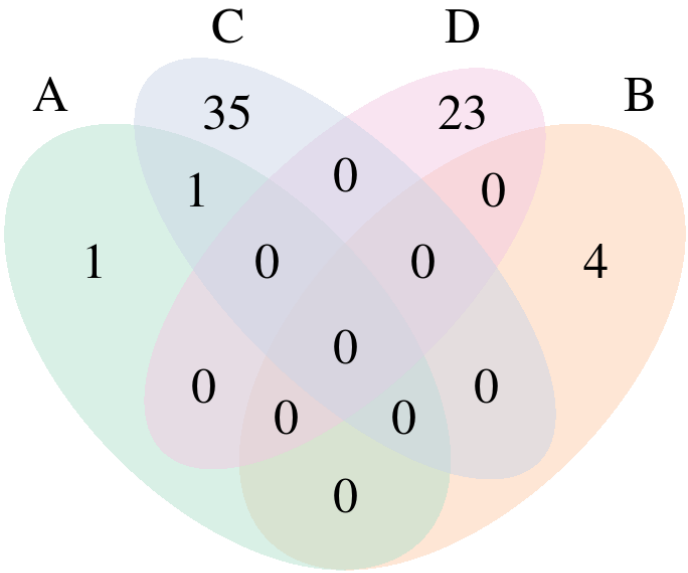

Figura 27: Diagrama de Venn que representa las intersecciones entre la listas de genes obtenidas en el contraste de diferencias de sexo en microglía y en astrocitos. A) Genes sobreexpresados en microglía de mujeres. B) Genes sobreexpresados en microglía de hombres. C) Genes sobreexpresados en astrocitos de mujeres. D) Genes sobreexpresados en astrocitos de hombres.

4.3.2 Caracterización funcional

En primer lugar, se revisó el conjunto de genes identificados con el método de expresión diferencial implementado en el paquete MAST, y que fueran comunes en los tres estudios, en las plataformas Open Targets y NCBI, con el objetivo de comprobar si previamente habían sido asociados a EA. El 41% de los genes estaban asociados a nuestra patología de interés, aspecto que confirma la robustez de la metodología empleada para la detección de genes con un perfil consenso. Por otra parte, la presencia de un alto porcentaje de genes sin describir, también proporciona la identificación de nuevos biomarcadores candidatos en EA.

A continuación, se llevó a cabo una descripción funcional de los genes detectados en el contraste de diferencias de sexo comunes en los tres estudios, mediante la información que proporciona la Gene Ontology a través de sus procesos biológicos y que está disponible en diversos recursos como GeneCards. Los genes detectados en microglía fueron 6 (2 sobreexpresados en mujeres y 4 en hombres) y su caracterización funcional se resume en las Tablas 23 y 24.

Tabla 23: Procesos biológicos de la GO asociados a los genes detectados en el contraste de diferencias de sexo del análisis de expresión diferencial como sobreexpresados en mujeres.

Sobreexpresados en mujeres	
NEAT1	
GO ID	Qualified GO Term
GO:0035195	involved_in gene silencing by miRNA
GO:0050729	involved_in positive regulation of inflammatory response
GO:0060965	involved_in negative regulation of gene silencing by miRNA
GO:1901647	involved_in positive regulation of synoviocyte proliferation
RGS16	
GO ID	Qualified GO Term
GO:0007186	involved_in G protein-coupled receptor signaling pathway
GO:0007601	involved_in visual perception
GO:0008277	involved_in regulation of G protein-coupled receptor signaling pathway
GO:0009968	involved_in negative regulation of signal transduction
GO:0043547	involved_in positive regulation of GTPase activity

Tabla 24: Procesos biológicos de la GO asociados a los genes detectados en el contraste de diferencias de sexo del análisis de expresión diferencial como sobreexpresados en hombres.

Sobreexpresados en hombres	
CKB	
GO ID	Qualified GO Term
GO:0006600	involved_in creatine metabolic process
GO:0007420	brain development
GO:0016310	involved_in phosphorylation
GO:0021549	involved_in cerebellum development
GO:0021762	involved_in substantia nigra development
GPC1	
GO ID	Qualified GO Term
GO:0001523	involved_in retinoid metabolic process
GO:0006024	involved_in glycosaminoglycan biosynthetic process
GO:0006027	involved_in glycosaminoglycan catabolic process
GO:0007411	involved_in axon guidance
GO:0009966	regulation of signal transduction

LINGO1	
GO:0050771	involved_in negative regulation of axonogenesis
MT3	
GO:0001666	involved_in response to hypoxia
GO:0001934	involved_in positive regulation of protein phosphorylation
GO:0006112	involved_in energy reserve metabolic process
GO:0006707	involved_in cholesterol catabolic process
GO:0006829	involved_in zinc ion transport

En el caso de los astrocitos el número de genes detectados es mucho mayor, 59 genes, 36 sobreexpresados en mujeres y 23 en hombres. Por esta razón, se llevó a cabo un análisis de sobrerrepresentación con la herramienta Panther. Sin embargo, no se detectaron procesos biológicos de la GO significativos enriquecidos para ninguno de estos grupos de genes. Posteriormente se utilizó la herramienta web STRING para identificar las relaciones descritas entre las proteínas codificadas por los distintos genes en caso de ser codificantes. La red obtenida a partir de los genes sobreexpresados en mujeres está poco interconectada, sin embargo la red obtenida a partir de los genes sobreexpresados en hombres sí está más interconectada (Figura 28). Además, 4 de estas proteínas pertenecen a la ruta KEGG hsa05010 de la EA (Figura 29).

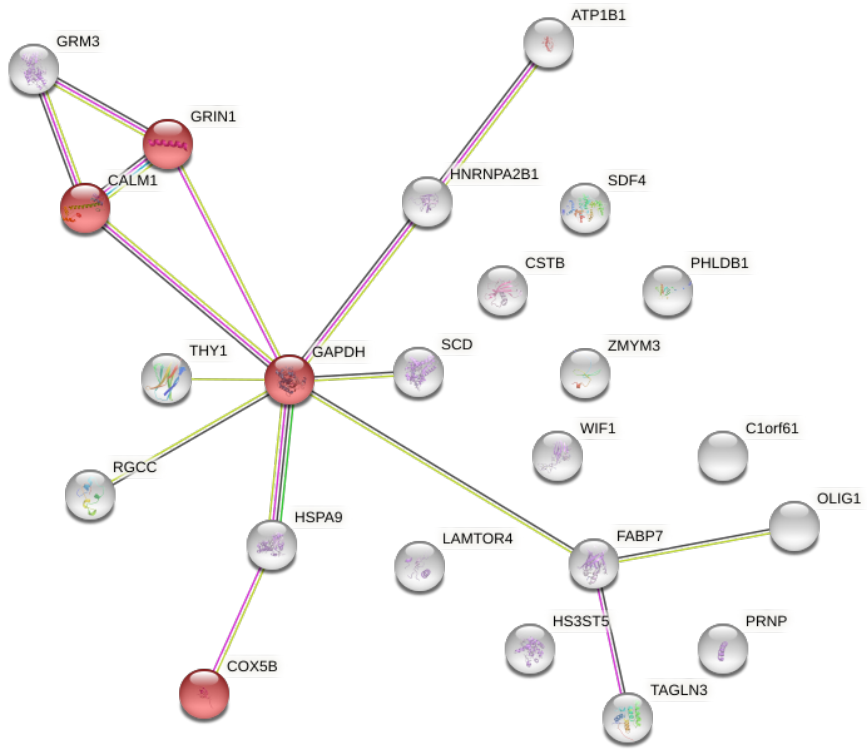

Figura 28: Red formada por las proteínas codificadas por los genes detectados en el contraste de diferencias de sexo en el análisis de expresión diferencial como sobreexpresados en hombres. En rojo las proteínas pertenecientes a la ruta KEGG hsa05010 de la EA.

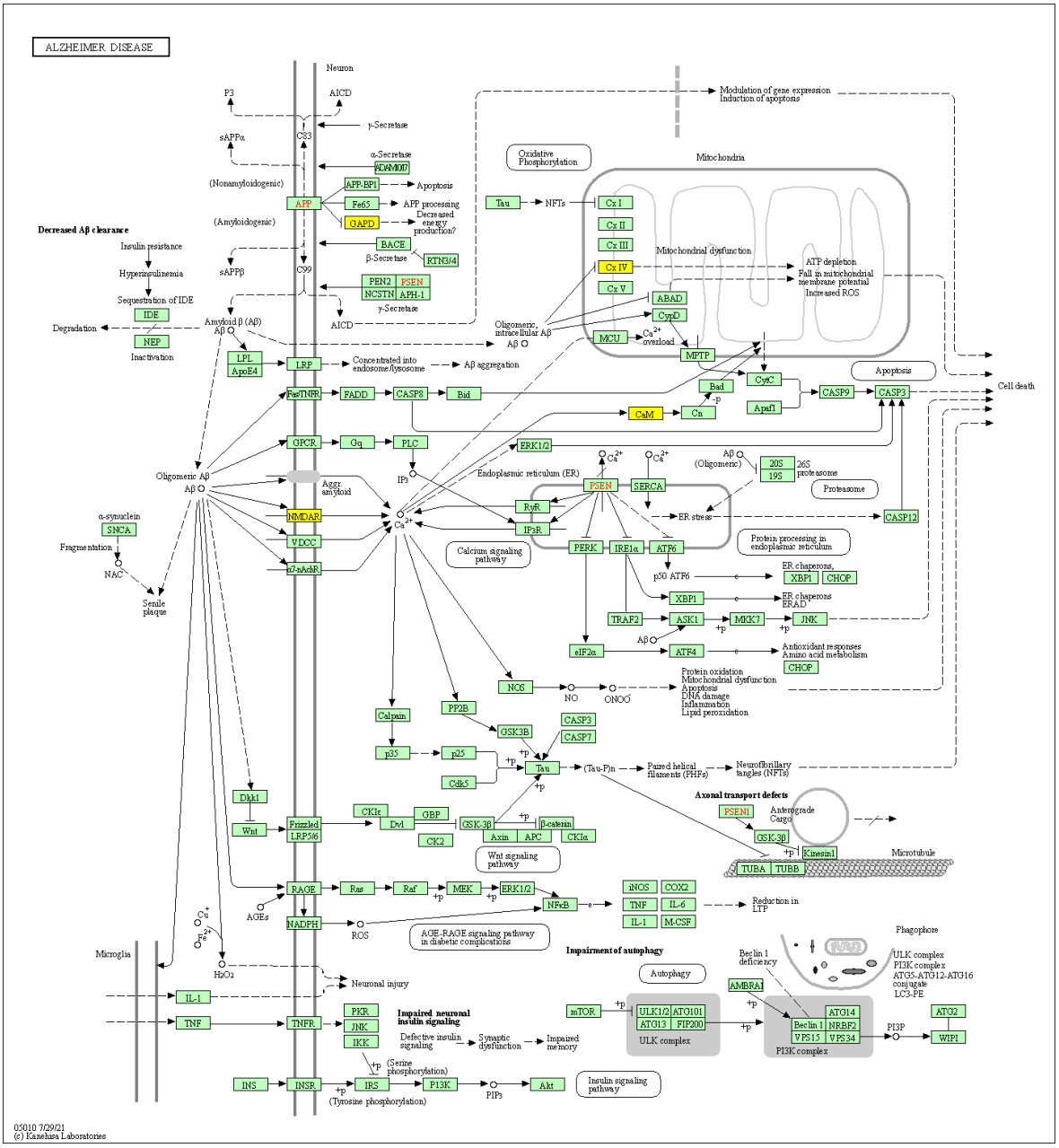

Figura 29: Ruta KEGG de la EA, en amarillo 4 de los genes detectados en el contraste de diferencias de sexo en el análisis de expresión diferencial como sobreexpresados en hombres.

Finalmente, se utilizó la herramienta STRING para la caracterización funcional de los resultados de los contrastes individuales en hombres y mujeres comunes en ambos sexos. La red obtenida (Figura 30) no está muy conectada, pero está enriquecida en proteínas asociadas a la membrana y a las uniones celulares.

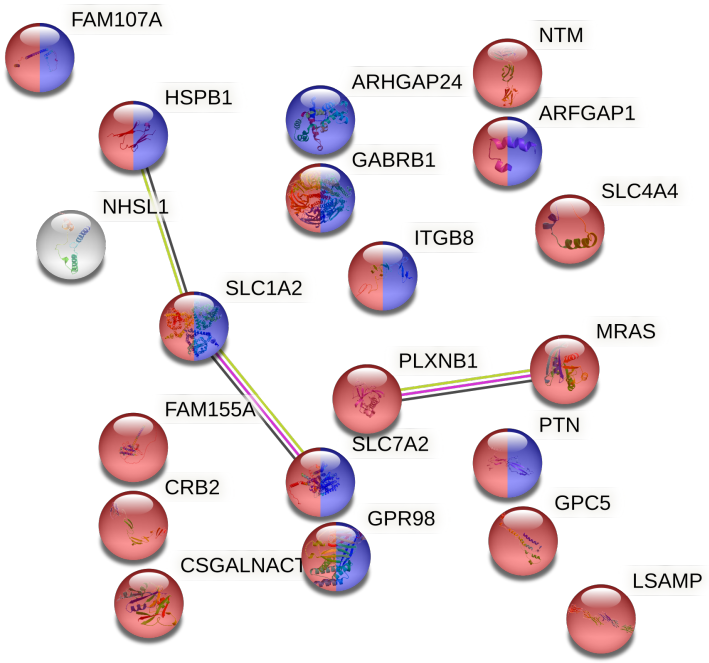

Figura 30: Red formada por las proteínas codificadas por los genes detectados en los contrastes individuales de hombres y mujeres comunes en ambos sexos. En rojo las proteínas asociadas al componente celular "membrana" (GO:0016020) y en azul las asociadas al componente celular "unión celular" (GO:0030054).

4.4 Requerimientos computacionales

El coste computacional de este abordaje es alto debido a la enorme cantidad de células y genes que se evalúan de cada sujeto. Como ejemplo se muestran en la Tabla 25 los tiempos computacionales empleados en el paso de análisis de expresión diferencial llevado a cabo con el paquete MAST, uno de los pasos más costosos. Además, se precisó un alto requerimiento de memoria para almacenar los datos. Tan solo los datos de partida de uno de los tres estudios descargados ocupaba 2,1 GB (GSE160936).

Tabla 25: Tiempos de ejecución empleados por el script del análisis de expresión diferencial llevado a cabo con el paquete MAST en cada uno de los estudios, así como otro tipo de datos relevantes para reflejar el coste computacional asociado a este trabajo.

Estudio	Tipo celular	Tiempo de ejecución (días-horas:minutos:segundos)	NCPUS /cores	Número de células	Número de genes
GSE138852	Microglía	00:13:33	1	678	8734
	Astrocitos	00:46:49	10	1878	8734
GSE157827	Microglía	10:23:26	10	6854	18743
	Astrocitos	18:52:55	15	15213	18743
GSE160936	Microglía	1-16:18:57	10	24294	21496
	Astrocitos	1-07:08:46	15	20621	21496

5. Discusión

5.1 Limitaciones y fortalezas del trabajo

Tras la revisión sistemática se obtuvieron un total de 7 estudios enfocados en la Enfermedad de Alzheimer, que empleasen un **abordaje de transcriptómica de célula o núcleo único** en *Homo sapiens* y que además presentaran tanto muestras de pacientes como de controles. Este número reducido de estudios refleja la novedad de este tipo de abordajes. En general, el número de trabajos que utilizan scRNA-seq o snRNA-seq es bajo si lo comparamos con el número de estudios, donde se han empleado otras tecnologías transcriptómicas como los microarrays o RNA-seq. De hecho, en los repositorios como GEO y ArrayExpress ni siquiera existe un filtro para seleccionar los estudios que hayan utilizado este tipo de abordaje, dificultando su búsqueda. Sin embargo, un aspecto positivo y poco usual es que **los 7 estudios contenían información del sexo** de los sujetos y representación de ambos sexos tanto en el grupo control como en el grupo EA, lo cual **arroja luz a la inclusión de la perspectiva de sexo en la investigación biomédica**. La falta de la información de sexo o la falta de representación de algunos de los sexos son criterios de exclusión bastante frecuentes en las revisiones sistemáticas realizadas en trabajos[66,67] basados en el estudio de las diferencias de sexo en distintas enfermedades.

Por otra parte, cabe destacar la **potencialidad de los abordajes *in silico* y de la reutilización de los datos**. La investigación abierta es fundamental para el avance científico, así como para una eficiente gestión de los recursos en la investigación. Con ese objetivo se establecieron los **principios FAIR**[68] (FINDABLE (Encontrables), ACCESSIBLE (Accesibles), INTEROPERABLE (Interoperables), REUSABLE (Reutilizables)) para el manejo y administración de datos científicos. Específicamente el

desarrollo de este trabajo, se ha podido llevar a cabo gracias a la implementación de estos principios también en el área de investigación biomédica.

El análisis de datos de experimentos de scRNA-seq es complejo por su **alto coste computacional y por la falta de estandarización** asociada a la novedad del campo. Por una parte, los tiempos de ejecución de los programas y la memoria requerida son altos, siendo necesaria una infraestructura computacional como el cluster del CIPF. Por otra parte, existen muchas opciones para realizar cada paso del análisis, por lo que para definir la *pipeline* se tuvieron que realizar muchas bifurcaciones y volver continuamente a modificar pasos anteriores. Un ejemplo de ello es el paso de **análisis de expresión diferencial**, en el que se llevaron a cabo varios abordajes. Finalmente, se consideró que el abordaje llevado a cabo con el **paquete MAST era el más adecuado** para el análisis de este tipo de datos. Además, aunque se está avanzando mucho y cada vez son más las herramientas disponibles para el análisis de este tipo de datos en Python y sobre todo en R, todavía sigue siendo común en algunos pasos la **reutilización y readaptación de paquetes de bulk RNA-seq**. Estos paquetes al no estar creados para analizar este tipo de datos **no son los más adecuados** porque no suelen considerar las peculiaridades de los mismos. Un ejemplo de ello lo encontramos en el método de análisis de expresión diferencial implementado por el paquete edgeR, inicialmente desarrollado para datos de RNA-seq. Su adaptación a datos de scRNA-seq no ha dado buenos resultados en este trabajo.

El **análisis integrativo de estudios de transcriptómica de núcleo único** llevado a cabo en este trabajo es **novedoso** y **prometedor** para el estudio de enfermedades como la EA. Además, la **perpectiva de sexo** lo hace aún más interesante, constituyendo un paso fundamental hacia la **medicina personalizada**.

5.2 Diferencias detectadas en la composición celular en los estudios de EA

En primer lugar, el porcentaje de células de cada tipo celular detectado varía de un estudio a otro y no coincide con los descritos en la bibliografía. En el estudio GSE160936 el cambio de los porcentajes de cada tipo celular podría explicarse por el procedimiento experimental de enriquecimiento de los tipos celulares de interés. En el resto de estudios podría pensarse que estos cambios se deben a los pacientes de EA incluidos, ya que la EA se caracteriza por la pérdida neuronal y la proliferación de la microglía[69] y los porcentajes de referencia utilizados se corresponden con los de un cerebro sano. Sin embargo, al analizar los porcentajes de cada tipo celular encontrados en los distintos grupos definidos por el sexo y la condición experimental, no parecen estar asociados a estas covariables. Además, en el análisis de abundancia diferencial, por lo general no se han detectado diferencias en la composición celular entre pacientes y controles ni entre sexos. Solo se ha detectado una diferencia significativa en la cantidad de células precursoras de oligodendrocitos en el contraste de mujer en el estudio GSE138852. Este resultado indica que el número de OPC es mayor en mujeres control respecto a las mujeres con EA. Esto concuerda con otros estudios que asocian la disminución de OPC con la EA, la pérdida de la mielina y el deterioro cognitivo[70].

Estos resultados sugieren que las diferencias detectadas en la composición celular al comparar los resultados obtenidos en este trabajo y los porcentajes descritos en la bibliografía no se deben principal ni exclusivamente a diferencias de abundancia celular

asociadas a la EA y/o al sexo. Estas diferencias podrían estar asociadas al procedimiento experimental o a algún paso del análisis bioinformático de los datos, como el control de calidad o la anotación del tipo celular.

5.3 Diferencias transcriptómicas específicas y no específicas de sexo detectadas en la EA.

Como resultado del análisis integrativo llevado a cabo en este trabajo, se han detectado genes expresados diferencialmente entre mujeres y hombres con EA, descartando las diferencias entre sexos no asociadas a la enfermedad. Estos genes podrían ser potenciales **marcadores de la enfermedad específicos de sexo.** Además, se han detectado genes expresados diferencialmente en pacientes con EA respecto a controles, comunes en ambos sexos, los cuales podrían ser **potenciales marcadores de la enfermedad no específicos de sexo.**

El **41%** de los genes detectados estaban **asociados previamente a la EA**, aportando **robustez** al análisis utilizado. En el resto de genes no hay asociación a la EA descrita, sin embargo, un estudio más profundo de estos genes podría servir para **encontrar nuevos genes implicados**, que podrían ayudar a descifrar los enclaves de esta compleja enfermedad.

Entre los potenciales marcadores de la enfermedad no específicos de sexo, encontramos **DPYD**, detectado en microglía como sobreexpresado en pacientes con EA respecto a controles. Aunque no es un marcador asociado a la EA establecido, sí se ha detectado en el estudio syn18485175[71] como un gen que sufre cambios de expresión asociados a la progresión de la EA. Por otra parte, en astrocitos se han detectado 20 genes, 8 sobreexpresados y 12 infraexpresados, en EA respecto a controles. La mayoría de estos genes están asociados a membrana y aproximadamente un 50% presenta asociación descrita a la EA.

Aunque los genes expresados diferencialmente entre pacientes de EA y controles independientemente del sexo son resultados interesantes derivados de este trabajo, el objetivo principal del mismo es la caracterización de las **diferencias de sexo** en la EA. Por tanto, la interpretación biológica se centrará en genes detectados en el contraste de diferencias de sexo.

Para la interpretación biológica de estos resultados se debe tener en cuenta que para obtener estos datos se han utilizado núcleos en lugar de células completas. Aunque se suele hacer la suposición de que los niveles de expresión nucleares son buenos indicadores del perfil de expresión general de la célula, esto no es siempre cierto. Por ejemplo, en este tipo de abordajes podría perderse la señal de algún gen fuertemente expresado que tienda a localizarse en el citoplasma para mejorar la eficiencia de la traducción. Asimismo, el nivel de expresión detectado de los genes depende también de la velocidad de exportación nuclear. Por último, el secuestro patológico de los transcritos de un gen codificante en el núcleo, podría llevar a su detección como gen sobreexpresado, aunque en realidad se tradujese en una reducción de la actividad llevada a cabo por la proteína resultante[29].

En microglía se han detectado 6 genes diferencialmente expresados entre mujeres y hombres con la EA, 2 sobreexpresados en mujeres y 4 sobreexpresados en hombres. En astrocitos el número de genes diferencialmente expresados asciende a 59, de los cuales 36

están sobreexpresados en mujeres y 23 sobreexpresados en hombres. De estos genes solo coincide un gen en ambos tipos celulares. Esto refleja la **necesidad de los abordajes de célula única o núcleo único**, ya que permite **revelar la heterogeneidad enmascarada** en los abordajes de bulk RNA-seq.

NEAT1 es el gen detectado en este trabajo en ambos tipos celulares como sobrexpresado en mujeres con EA respecto a hombres con EA. La asociación de este gen con la EA ha sido descrita por la comunidad científica. Se trata de un gen que da lugar a un transcrito largo no codificante (Long-non-coding RNAs, en inglés) que queda retenido en el núcleo y que probablemente actúe como **regulador transcripcional** de numerosos genes, como apuntan los procesos biológicos de la GO asociados al mismo (Tabla 23)[72]. Se ha descrito como oncogen en varios tipos de cáncer y se ha asociado a enfermedades neurodegenerativas, pero se sabe muy poco acerca de su desregulación en la EA y su papel en la neurodegeneración. Un estudio reciente ha descrito que NEAT1 regula negativamente los niveles de CDK5R1, gen que codifica para la proteína p35 y que se encuentra sobreexpresado en la EA. Esto sugiere un **papel neuroprotector** de NEAT1 en la EA para compensar los altos niveles de CDK5R1. También coloca a NEAT1 como **posible marcador molecular** de la EA y como **posible diana farmacológica**[73,74].

Además de NEAT1, en microglía también se ha detectado el gen **RGS16** como sobreexpresado en mujeres con EA. No se ha encontrado asociación previa de este gen a la EA. Se trata de un gen que codifica para una proteína que pertenece a la familia de proteínas reguladoras de la señalización de las proteínas G. Otras proteínas de esta misma familia como RGS4 sí han sido asociadas a la EA[75].

Por otra parte, los genes detectados como sobreexpresados en hombres con EA respecto a mujeres con EA son CKB, GPC1, LINGO1 y MT3. En primer lugar, CKB, un gen con asociación descrita a la EA que codifica para la creatina quinasa B, una enzima citoplasmática involucrada en la homeostasis de la energía[76]. CKB es diana del alcaloide TGN, un posible candidato a fármaco para la EA, que se ha probado en ratones y conlleva a la **activación de CKB** que provoca **formación axonal**[77]. En segundo lugar, **GPC1**, otro gen con asociación descrita a la EA. Codifica para un proteoglicano con cadenas del polisacárido heparán sulfato (en inglés heparan sulfate proteoglycans (HSPGs)). Se ha demostrado que GPC1 **se puede unir a las fibrillas de Aβ** mediante las cadenas de heparán sulfato. Se piensa que esta proteína interacciona con los oligómeros o polímeros de Aβ, dando lugar a la **deposición de amiloide en las placas seniles en los cerebros de pacientes con EA** y acelerando la muerte neuronal en respuesta al estrés y a Aβ[78]. En cuanto a **LINGO1**, codifica para una proteína transmembrana muy abundante en el cerebro implicada en muchas enfermedades neurodegenerativas. Existen indicios de que esta proteína también esté asociada a la EA, **favoreciendo** la proteolisis de APP por la ruta amiloidogénica y por tanto **la generación de Aβ**, pero también mediante la activación de rutas de señalización claves en la inhibición del crecimiento y la supervivencia neuronal[79]. Por último, **MT3** codifica para una proteína de unión a metales predominante en el cerebro. Se encarga de **mantener la homeostasis del cobre y del zinc en las células**, protegiendolas del estrés oxidativo y regulando el crecimiento y la diferenciación celular. La distribución anormal de metales está estrechamente relacionada con muchas enfermedades como la diabetes, el cáncer o la EA[80].

Entre los genes detectados en el contraste de diferencias de sexo en astrocitos podemos resaltar **PPP1R12B**, gen que codifica para la subunidad reguladora 12B de la

proteína fosfatasa 1, y **PRKG1**, que codifica una quinasa dependiente de GMP, por ser la única conexión encontrada por la herramienta STRING entre los genes detectados como sobreexpresados en mujeres en el contraste de diferencias de sexo. Ninguno de estos genes está asociado con EA en la bibliografía revisada. La relación fue establecida por STRING porque homólogos de estas proteínas interaccionan y se coexpresan.

Por último, sería interesante resaltar 4 genes detectados en astrocitos en el contraste de diferencias de sexo en EA como sobreexpresados en hombres: CALM1, COX5B, GAPDH y GRIN1. Estos genes codifican proteínas pertenecientes a la ruta de KEGG de la EA (*hsa05010*). En primer lugar, **CALM1** es un gen importante en la **señalización del calcio**. La desregulación del calcio interrumpe la homeostasis y conlleva a la neurodegeneración[81]. La unión de la proteína codificada por este gen a los oligómeros Aβ conlleva a la desregulación del calcio y provoca la activación de la proteína MAPK que es responsable de la disfunción sináptica y del deterioro de la memoria[82,83]. En segundo lugar **COX5B**, cuya asociación a la EA está ya descrita. Este gen codifica la subunidad nuclear del complejo enzimático citocromo oxidasa C (COX). COX es el último **complejo enzimático la cadena respiratoria mitocondrial**. Está formado por unidades codificadas por genes mitocondriales (subunidades catalíticas) y por unidades codificadas por genes nucleares. Se piensa que las funciones de las unidades codificadas por genes nucleares podrían participar en la regulación y ensamblaje del complejo[84]. Varios estudios han descrito defectos en el complejo COX en la EA[85,86]. En cuanto a **GAPDH**, codifica para una enzima que participa en el metabolismo aeróbico de la glucosa, pero que también lleva a cabo **funciones no glicolíticas**, interesantes para la investigación de enfermedades neurodegenerativas. Existe una fuerte evidencia de la implicación directa e indirecta de GAPDH en la EA[87]. Por último, **GRIN1**, la proteína codificada por este gen es una subunidad fundamental de los **receptores de N-metil-D-aspartato** implicados en la plasticidad de la sinapsis neuronales[88], aunque también se han descrito en las membranas celulares de astrocitos. Los astrocitos son indispensables para la neurotransmisión neuronal, proporcionando apoyo físico y metabólico a las neuronas. Para ello, utilizan receptores de neurotransmisores localizados en sus membranas para comunicarse con las neuronas[89]. Esto explicaría la detección de este gen en astrocitos en este trabajo. Se piensa que la **desregulación del tráfico de estos receptores** podría estar **implicada en los síntomas de comportamiento** asociados a trastornos neuropsiquiátricos como la esquizofrenia o la EA[90].

En resumen, este trabajo ha permitido la caracterización de las diferencias transcriptómicas en microglía y astrocitos, específicas y no específicas de sexo, en la EA. Si bien el análisis integrativo llevado a cabo confirma las conclusiones de otros estudios, también proporciona nuevos genes cuyo estudio más profundo podría ayudar comprender esta compleja enfermedad, y mejorar el tratamiento y el diagnóstico a través de la identificación de biomarcadores.

6. Conclusiones

1. En base a la bibliografía revisada, este es el primer estudio que lleva a cabo un análisis integrativo con datos de snRNA-seq para caracterizar las diferencias de sexo en la EA.

2. Los abordajes de transcriptómica de célula o núcleo único presentan una gran potencialidad para el estudio de enfermedades humanas, ya que son capaces de revelar la heterogeneidad celular enmascarada en los resultados de bulk RNA-seq.
3. El análisis de datos de experimentos de scRNA-seq es complejo por su alto coste computacional, su alto requerimiento de memoria y por la falta de estandarización asociada a la novedad del campo.
4. No se han detectado diferencias sexo en la composición celular de pacientes con EA. Sin embargo, en uno de los estudios se detectó una mayor abundancia de células precursoras de oligodendrocitos en mujeres control respecto a mujeres con EA.
5. En base a los resultados obtenidos, el método de análisis expresión diferencial implementado por el paquete MAST es más adecuado para datos de snRNA-seq que el implementado por el paquete edgeR.
6. Este trabajo ha permitido la identificación de genes diferencialmente expresados en la EA específicos y no específicos de sexo, en microglía y astrocitos.
7. El 41% de los genes detectados como diferencialmente expresados estaban asociados previamente a la EA, aportando robustez al análisis utilizado. El 59% restante podrían ser potenciales marcadores de la enfermedad todavía no descritos.
8. Las células de la microglía y los astrocitos presentan patrones de expresión diferencial en función del sexo en pacientes con EA. Esto apoya la necesidad de la inclusión de la perspectiva de sexo en la investigación biomédica.

7. Perspectivas futuras

Como continuación del trabajo se han propuesto las siguientes líneas de actuación:

1. Ampliación del análisis integrativo con los 4 estudios descartados en la fase final de la revisión sistemática, así como la inclusión de posibles nuevos estudios derivados de la actualización de la revisión sistemática.
2. Refinamiento de la *pipeline* creada para el análisis de los datos de snRNA-seq.
3. Optimización de la estrategia computacional para el análisis de datos que permitirá una reducción del tiempo de ejecución de los trabajos.
4. Incluir la aplicación de metaanálisis de genes y funciones en la estrategia de integración.
5. Extensión del análisis descrito a otros tipos celulares como neuronas y oligodendrocitos.
6. Por último, ampliar el estudio de los genes identificados en los múltiples escenarios, con el objetivo de profundizar en su interpretación biológica, dentro del marco del trabajo presentado.

8. Bibliografía

1. Mauvais-Jarvis, F. *et al.* Sex and gender: modifiers of health, disease, and medicine. *The Lancet* vol. 396 565–582 (2020).
2. Erkkinen, M. G., Kim, M. O. & Geschwind, M. D. Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harb. Perspect. Biol.* **10**, (2018).
3. Kovacs, G. G. Concepts and classification of neurodegenerative diseases. in *Handbook of Clinical Neurology* vol. 145 301–307 (2018).
4. Yanguas-Casás, N. Sex Differences in Neurodegenerative Diseases. *SM J. Neurol. Disord. Stroke* (2017).
5. Dugger, B. N. & Dickson, D. W. Pathology of neurodegenerative diseases. *Cold Spring Harbor Perspectives in Biology* vol. 9 (2017).
6. Dubal, D. B. Sex difference in Alzheimer's disease: An updated, balanced and emerging perspective on differing vulnerabilities. *Handb. Clin. Neurol.* **175**, 261–273 (2020).
7. Soria Lopez, J. A., González, H. M. & Léger, G. C. Alzheimer's disease. in *Handbook of Clinical Neurology* vol. 167 231–255 (2019).
8. Lane, C. A., Hardy, J. & Schott, J. M. Alzheimer's disease. *European Journal of Neurology* vol. 25 59–70 (2018).
9. Mantzavinos, V. & Alexiou, A. Biomarkers for Alzheimer's Disease Diagnosis. *Curr. Alzheimer Res.* **14**, (2017).
10. Nebel, R. A. *et al.* Understanding the impact of sex and gender in Alzheimer's disease: A call to action. *Alzheimer's Dement.* **14**, 1171–1183 (2018).
11. Serrano-Pozo, A., Frosch, M. P., Masliah, E. & Hyman, B. T. Neuropathological alterations in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **1**, (2011).
12. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. *Alzheimer's and Dementia: Translational Research and Clinical Interventions* vol. 4 575–590 (2018).
13. Hane, F. T., Lee, B. Y. & Leonenko, Z. Recent Progress in Alzheimer's Disease Research, Part 1: Pathology. *Journal of Alzheimer's Disease* vol. 57 1–28 (2017).
14. Zhou, Y., Sun, Y., Ma, Q. H. & Liu, Y. Alzheimer's disease: Amyloid-based pathogenesis and potential therapies. *Cell Stress* **2**, 150–161 (2018).
15. Congdon, E. E. & Sigurdsson, E. M. Tau-targeting therapies for Alzheimer disease. *Nature Reviews Neurology* vol. 14 399–415 (2018).
16. Fakhoury, M. Microglia and Astrocytes in Alzheimer's Disease: Implications for Therapy. *Curr. Neuropharmacol.* **16**, (2018).
17. Robinson, M., Lee, B. Y. & Hane, F. T. Recent Progress in Alzheimer's Disease Research, Part 2: Genetics and Epidemiology. *Journal of Alzheimer's disease : JAD* vol. 57 317–330 (2017).
18. Shi, Y. & Holtzman, D. M. Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight. *Nat. Rev. Immunol.* **18**, 759–772 (2018).

19. Hane, F. T. *et al.* Recent Progress in Alzheimer's Disease Research, Part 3: Diagnosis and Treatment. *Journal of Alzheimer's Disease* vol. 57 645–665 (2017).

20. Chandra, A., Dervenoulas, G. & Politis, M. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. *Journal of Neurology* vol. 266 1293–1302 (2019).

21. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* vol. 50 (2018).

22. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome profiling in human diseases: New advances and perspectives. *International Journal of Molecular Sciences* vol. 18 (2017).

23. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, (2017).

24. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* vol. 9 (2017).

25. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1–9 (2020).

26. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, (2017).

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

28. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

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

30. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).

31. Zhang, X. *et al.* Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell* **73**, 130-142.e5 (2019).

32. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).

33. Price, A. J. *et al.* Characterizing the nuclear and cytoplasmic transcriptomes in developing and mature human cortex uncovers new insight into psychiatric disease gene regulation. *Genome Res.* **30**, 1–11 (2020).

34. Barthelson, R. A., Lambert, G. M., Vanier, C., Lynch, R. M. & Galbraith, D. W. Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. *BMC Genomics* **8**, (2007).

35. Grindberg, R. V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

36. Armand, E. J., Li, J., Xie, F., Luo, C. & Mukamel, E. A. Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron* vol. 109 11–26 (2021).

37. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2020).

38. Zenodo. https://about.zenodo.org/.

39. Morgan M, Obenchain V, Lang M, Thompson R, Turaga N. BiocParallel: Bioconductor facilities for parallel evaluation. R package version 1.26.1. (2021).

40. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**, (2009).

41. 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).

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

43. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

44. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, (2004).

45. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 1–6 (2018).

46. 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. *Bioinformatics* **33**, 1179–1186 (2017).

47. Germain, P. scDblFinder: scDblFinder. R package version 1.4.0. (2020).

48. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).

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

50. 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**, (2016).

51. Blighe, K. & Lun, A. PCAtools: PCAtools: Everything Principal Components Analysis. R package version 2.4.0. (2021).

52. Csardi, G. & Nepusz, T. The igraph software package for complex network research. InterJournal, Complex Systems, 1695. (2006).

53. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Reports 2018 81* **8**, 1–19 (2018).

54. Robinson, M., McCarthy, D. & Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140. (2010).

55. Ritchie, M. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47. (2015).

56. 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).

57. 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).

58. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

59. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1-1.30.33 (2016).

60. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).

61. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

62. https://www.diagrams.net/doc/.

63. Lau, S. F., Cao, H., Fu, A. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25800–25809 (2020).

64. Smith, A. M. *et al.* Diverse human astrocyte and microglial transcriptional responses to Alzheimer's pathology. *bioRxiv* 2021.07.19.452932 (2021).

65. Bahney, J. & Bartheld, C. S. von. The Cellular Composition and Glia–Neuron Ratio in the Spinal Cord of a Human and a Nonhuman Primate: Comparison With Other Species and Brain Regions. *Anat. Rec.* **301**, 697–710 (2018).

66. Català-Senent, J. F. *et al.* Hepatic steatosis and steatohepatitis: a functional meta-analysis of sex-based differences in transcriptomic studies. *Biol. Sex Differ.* **12**, (2021).

67. Pérez-Díez, I. *et al.* Functional Signatures in Non-Small-Cell Lung Cancer: A Systematic Review and Meta-Analysis of Sex-Based Differences in Transcriptomic Studies. *Cancers 2021, Vol. 13, Page 143* **13**, 143 (2021).

68. https://datos.gob.es/es/noticia/principios-fair-buenas-practicas-para-la-gestion-y-administracion-de-datos-cientificos.

69. Johnson, T. S. *et al.* Spatial cell type composition in normal and Alzheimers human brains is revealed using integrated mouse and human single cell RNA sequencing. *Sci. Reports 2020 101* **10**, 1–14 (2020).

70. Vanzulli, I. *et al.* Disruption of oligodendrocyte progenitor cells is an early sign of pathology in the triple transgenic mouse model of Alzheimer's disease. *Neurobiol. Aging* **94**, 130–139 (2020).

71. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

72. NEAT1 nuclear paraspeckle assembly transcript 1 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/283131.

73. Zhao, M.-Y. *et al.* The long-non-coding RNA NEAT1 is a novel target for Alzheimer's disease progression via miR-124/BACE1 axis. *Neurol. Res.* **41**, 489–497 (2019).

74. Spreafico, M., Grillo, B., Rusconi, F., Battaglioli, E. & Venturin, M. Multiple Layers of CDK5R1 Regulation in Alzheimer's Disease Implicate Long Non-Coding RNAs. *Int. J. Mol. Sci. 2018, Vol. 19, Page 2022* **19**, 2022 (2018).

75. Squires, K. E., Montañez-Miranda, C., Pandya, R. R., Torres, M. P. & Hepler, J. R. Genetic Analysis of Rare Human Variants of Regulators of G Protein Signaling Proteins and Their Role in Human Physiology and Disease. *Pharmacol. Rev.* **70**, 446 (2018).

76. CKB creatine kinase B [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/1152.

77. Farid, M. M., Yang, X., Kuboyama, T. & Tohda, C. Trigonelline recovers memory function in Alzheimer's disease model mice: evidence of brain penetration and target molecule. *Sci. Reports 2020 101* **10**, 1–10 (2020).

78. Watanabe, N. *et al.* Glypican-1 as an Abeta binding HSPG in the human brain: its localization in DIG domains and possible roles in the pathogenesis of Alzheimer's disease. *FASEB J.* **18**, 1013–1015 (2004).

79. Fernandez-Enright, F. & Andrews, J. L. Lingo-1: a novel target in therapy for Alzheimer's disease? *Neural Regen. Res.* **11**, 88 (2016).

80. Koh, J.-Y. & Lee, S.-J. Metallothionein-3 as a multifunctional player in the control of cellular processes and diseases. *Mol. Brain* **13**, (2020).

81. Noman Bin, A., Muhammad Imran, N. & Myeong Ok, K. Comparative Gene-Expression Analysis of Alzheimer's Disease Progression with Aging in Transgenic Mouse Model. *Int. J. Mol. Sci. 2019, Vol. 20, Page 1219* **20**, 1219 (2019).

82. Corbacho, I., Berrocal, M., Török, K., Mata, A. M. & Gutierrez-Merino, C. High affinity binding of amyloid β-peptide to calmodulin: Structural and functional implications. *Biochem. Biophys. Res. Commun.* **486**, 992–997 (2017).

83. Shetty, M. S. & Sajikumar, S. Differential involvement of Ca2+/calmodulin-dependent protein kinases and mitogen-activated protein kinases in the dopamine D1/D5 receptor-mediated potentiation in hippocampal CA1 pyramidal neurons. *Neurobiol. Learn. Mem.* **138**, 111–120 (2017).

84. COX5B cytochrome c oxidase subunit 5B [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/1329.

85. Mastroeni, D. *et al.* Nuclear but not mitochondrial-encoded oxidative phosphorylation genes are altered in aging, mild cognitive impairment, and Alzheimer's disease. *Alzheimers. Dement.* **13**, 510 (2017).

86. Feldhaus, P. *et al.* Evaluation of respiratory chain activity in lymphocytes of patients with Alzheimer disease. *Metab. Brain Dis.* **26**, 229–236 (2011).

87. Butterfield, D. A., Hardas, S. S. & Lange, M. L. B. Oxidatively Modified Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) and Alzheimer Disease: Many Pathways to Neurodegeneration. *J. Alzheimers. Dis.* **20**, 369 (2010).

88. GRIN1 glutamate ionotropic receptor NMDA type subunit 1 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/2902.

89. Skowrońska, K., Obara-Michlewska, M., Zielińska, M. & Albrecht, J. NMDA Receptors in Astrocytes: In Search for Roles in Neurotransmission and Astrocytic Homeostasis. *Int. J. Mol. Sci.* **20**, (2019).

90. Lau, C. G. & Zukin, R. S. NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. *Nat. Rev. Neurosci. 2007 86* **8**, 413–426 (2007).