Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

ANÁLISIS DEL TRANSCRIPTOMA CON RESOLUCIÓN ESPACIAL PARA EL ESTUDIO DE LAS DIFERENCIAS DE SEXO EN ESCLEROSIS MÚLTIPLE

AUTORA:
NATALIA DEL REY DÍEZ

TUTORES:
FRANCISCO GARCÍA GARCÍA
CRISTINA GALIANA ROSELLÓ
VICENTE ARNAU LLOMBART
IRENE SOLER SÁEZ
CARLA PERPIÑÁ CLÉRIGUES

JULIO, 2025

RESUMEN

La esclerosis múltiple (EM) es la principal causa de discapacidad no ocasionada por accidentes entre los adultos jóvenes. Se trata de una enfermedad inflamatoria crónica del Sistema Nervioso Central (SNC), caracterizada por ataques autoinmunes a las vainas de mielina, que generan lesiones diseminadas por el SNC. La composición celular de las distintas regiones lesionales, así como sus cambios entre la fase crónica activa e inactiva, aún no se comprenden completamente. Además, al igual que en otras enfermedades autoinmunes y neurodegenerativas, también se han descrito diferencias entre sexos en cuanto a la susceptibilidad y progresión de la EM. En este trabajo, a través de un abordaje *in silico* con datos de transcriptómica espacial y unicelular, se caracterizaron los principales tipos celulares de las regiones espaciales asociadas a áreas lesionadas y no lesionadas de 19 muestras de tejido cerebral humano, y se evaluaron las diferencias de expresión génica entre sexos con resolución espacial. Para el análisis de transcriptómica espacial, se implementó el flujo de trabajo de *Giotto*, una *suite* de análisis computacional diseñada específicamente para datos espaciales, explorando las diferentes alternativas metodológicas de cada etapa del análisis bioinformático y seleccionando en cada caso el enfoque más adecuado a las características específicas de los datos. El perfil de expresión diferencial de los genes permitió caracterizar regiones específicas como el núcleo de la lesión de EM (alta abundancia de astrocitos y disminución de oligodendrocitos), el borde de la lesión (enriquecido en microglía activada), la sustancia blanca sana (elevada densidad de astrocitos) y la sustancia gris (alta proporción de neuronas). A partir de esta caracterización, se delimitó el análisis diferencial entre sexos a la región del núcleo de la lesión. Los hombres presentaron una sobreexpresión de genes relacionados con el estrés y la muerte celular, asociada a una mayor atrofia cerebral y pérdida neuronal respecto a las mujeres. En contraste, en las mujeres se observaron sobreexpresados genes vinculados a procesos inflamatorios y a la remodelación del citoesqueleto. Estos resultados ofrecen información sobre los cambios en la composición celular en las lesiones de la EM y destacan las diferencias de expresión génica entre sexos, las cuales podrían estar asociadas a la progresión diferencial de esta enfermedad. En conjunto, este trabajo demuestra la utilidad de la transcriptómica espacial para el estudio de la heterogeneidad celular y su organización dentro de un tejido en el contexto de enfermedades neurodegenerativas.

Palabras clave: esclerosis múltiple, transcriptómica espacial, snRNA-Seq, diferencias de sexo, neurodegeneración.

ABSTRACT

Multiple sclerosis (MS) is the leading cause of non-accident-related disability among young adults. It is a chronic inflammatory disease of the Central Nervous System (CNS), characterized by autoimmune attacks on myelin sheaths, leading to disseminated lesions throughout the CNS. The cellular composition of the different lesion regions, as well as their changes between chronic active and inactive phases, are not completely understood. Moreover, as in other autoimmune and neurodegenerative diseases, sex differences in MS susceptibility and progression have been described. In this study, through an *in silico* approach using spatial and single-cell transcriptomics data, we characterized the main cell types in spatial regions associated with lesioned and non-lesioned areas of 19 human brain tissue samples, and assessed spatially resolved sex differences in gene expression. For the spatial transcriptomics analysis, we implemented the *Giotto* workflow, a toolbox designed specifically for spatial data, systematically exploring different methodological alternatives at each stage of the bioinformatics analysis and selecting in each case the approach best suited to the specific characteristics of the data. Differential gene expression profiling enabled the characterization of specific regions such as the MS lesion core (high abundance of astrocytes and decreased oligodendrocytes), lesion rim (enriched in activated microglia), healthy white matter (high density of astrocytes) and gray matter (high proportion of neurons). Based on this characterization, the analysis of sex differences was focused on the lesion core. Males showed overexpression of genes related to stress and cell death, associated with greater brain atrophy and neuronal loss compared to females. In contrast, females exhibited overexpression of genes linked to inflammatory processes and cytoskeleton remodeling. These results provide insights into changes in cellular composition in MS lesions and highlight sex differences in gene expression, which may be associated with the differential progression of this disease.

Keywords: multiple sclerosis, spatial transcriptomics, snRNA-Seq, sex differences, neurodegeneration.

RESUM

L'esclerosi múltiple (EM) és la principal causa de discapacitat no ocasionada per accidents entre els adults joves. Es tracta d'una malaltia inflamatòria crònica del Sistema Nerviós Central (SNC), caracteritzada per atacs autoimmunes a les beines de mielina, que generen lesions disseminades per tot el SNC. La composició cel·lular de les diferents regions lesionals, així com els seus canvis entre les fases cròniques actives i inactives, no es comprenen completament. A més, igual que en altres malalties autoimmunes i neurodegeneratives, s'han descrit diferències entre sexes quant a la susceptibilitat i progressió de l'EM. En aqueste treball, a través d'un abordatge *in silico* amb dades de transcriptòmica espacial i unicel·lular, es van caracteritzar els principals tipus cel·lulars de les regions espacials associades a àrees lesionades i no lesionades de 19 mostres de teixit cerebral humà, i es van avaluar les diferències d'expressió gènica entre sexes amb resolució espacial. Per a l'anàlisi de transcriptòmica espacial, es va implementar el flux de treball de *Giotto*, una *suite* d'anàlisi computacional dissenyada específicament per a dades espacials, explorant les diferents alternatives metodològiques de cada etapa de l'anàlisi bioinformàtica i seleccionant en cada cas l'estratègia més adequat a les característiques específiques de les dades. El perfil d'expressió diferencial dels gens va permetre caracteritzar regions específiques com el nucli de la lesió d'EM (alta abundància d'astròcits i disminució d'oligodendròcits), la vora de la lesió (enriquit en microglia activada), la substància blanca sana (elevada densitat d'astròcits) i la substància grisa (alta proporció de neurones). A partir d'esta caracterització, es va delimitar l'anàlisi diferencial entre sexes a la regió del nucli de la lesió. Els hòmens van presentar una sobreexpressió de gens relacionats amb l'estrés i la mort cel·lular, associada a una major atròfia cerebral i pèrdua neuronal respecte a les dones. En contraposició, en aquestes últimes es van observar sobreexpresats gens vinculats a processos inflamatoris i a la remodelació del citoesquelet. Estos resultats oferixen informació sobre els canvis en la composició cel·lular en les lesions de l'EM i destaquen les diferències d'expressió gènica entre sexes, les quals podrien estar associades a la progressió diferencial d'esta malaltia. En conjunt, este treball demostra la utilitat de la transcriptòmica espacial per a l'estudi de l'heterogeneïtat cel·lular i la seua organització dins d'un teixit en el context de malalties neurodegeneratives.

Paraules clau: esclerosi múltiple, transcriptòmica espacial, snRNA-Seq, diferències de sexe, neurodegeneració.

ÍNDICE

ABREVIATURAS Y ACRÓNIMOS

A

ADN: Ácido desoxirribonucleico.
ADNc: Ácido desoxirribonucleico complementario.
ARN: Ácido ribonucleico.
ARNm: Ácido ribonucleico mensajero.

B

BHE: Barrera hematoencefálica.

C

CIPF: Centro de Investigación Príncipe Felipe.
COV: *Coefficients of variation* (coeficiente de variación.
CPU: *Central Processing Unit* (Unidad Central de Procesamiento).
CTRL: Control.

D

DWLS: *dampened weighted least squares* (mínimos cuadrados ponderados amortiguados).

E

EBT: *Extraction-based techniques* (técnicas basadas en la extracción).
EGA: *European Genome Archive*.
EM-A: Lesiones agudas de esclerosis múltiple.
EM-CA: Lesiones crónicas activas de esclerosis múltiple.
EM-CI: Lesiones crónicas inactivas de esclerosis múltiple.
EM: Esclerosis múltiple.
EMPP: Esclerosis múltiple primaria progresiva.
EMRR: Esclerosis múltiple remitente recurrente.
EMSP: Esclerosis múltiple secundaria progresiva.

F

FAIR: *Findable, Accessible, Interoperable and Reusable* (Localizable, accesible, interoperable y reutilizable).
FC: *Fold Change* (tasa de cambio).
FDR: *False disconvery Rate* (Tasa de Falso Descubrimiento).
FFPE: *Formalin-Fixed, Paraffin Embedded* (fijado con formalina, incluidas en parafina).
FISH: *Fluorescence in situ hybridization* (hibridación *in situ* de fluorescencia).

G

GEO: *Gene Expression Omnibus*.
GSEA: *Gene Set Enrichment Analysis* (análisis de enriquecimiento de conjuntos de genes).
GWAS: *Genome-wide association study* (estudios de asociación de genoma completo).

H

H&E: Hematoxilina y eosina.
HLA: *Human Leukocyte Antigens* (antígenos leucocitarios humanos).
HMRF: *Hidden Markov Random Field* (Campo Aleatorio Oculto de Markov).
HSP: *Heat Shock Proteínas* (Proteínas de Choque Térmico).
HVG: *Highly Variable Genes* (Genes Altamente Variables).

I

ID: Identificador.
IDF: *Inverse Document Frequency* (frecuencia inversa del documento).
IgG: Inmunoglobulinas G.
IRM: Imágenes de resonancia magnética.

K

kNN: *k-Nearest Neighbors* (k-vecinos más cercanos).

L

LCR: Líquido cefalorraquídeo.
LOESS: *Locally Weighted Scatterplot Smoothing* (Suavizado de Diagrama de Dispersión Ponderado Localmente).

M

MMP: *Matrix metalloproteinases* (metaloproteinasas de matriz).

N

NEBT: *Non-extraction-based techniques* (técnicas no basadas en la extracción).
NGS: *Next-generation sequencing* (secuenciación de próxima generación).

O

ORA: *Over-Representation Analysis* (análisis de sobrerrepresentación).

P

PAGE: *Parametric Analysis of Gene Set Enrichment* (Análisis paramétrico del enriquecimiento de conjuntos genéticos).
PCA: *Principal Component Analysis* (Análisis de Componentes Principales).
PRISMA: *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (elementos de información preferidos para revisiones sistemáticas y metaanálisis).

R

RAM: *Random Access Memory* (Memoria de Acceso Aleatorio).
RCA: *Rolling circle amplification* (amplificación de círculo rodante).
RNA-Seq: *Ribonucleic acid sequencing* (secuenciación de ácido ribonucleico).

S

SCA: Síndrome Clínico Aislado.
scRNA-Seq: *Single-cell ribonucleic acid sequencing* (secuenciación de ácido ribonucleico de célula única).
SLURM: *Simple Linux Utility for Resource Management* (Utilidad Simple de Linux para la Gestión de Recursos).
SNC: Sistema Nervioso Central.
sNN: *Shared Nearest Neighbor* (vecino más cercano compartido).
snRNA-Seq: *Single-nucleus ribonucleic acid sequencing* (secuenciación de ácido ribonucleico de núcleo único).
SOAR: *Spatial transcriptOmics Analysis Resource*.
ST: *Spatial transcriptomics* (transcriptómica espacial).
STOmics DB: *Spatial TranscriptOmics DataBase*.
SVG: *Spatially Variable Genes* (Genes Espacialmente Variables).

T

TF-IDF: *Term Frequency - Inverse Document Frequency* (Frecuencia de término - Frecuencia inversa del documento).
TF: *Term Frequency* (frecuencia del término).
TME: Terapias modificadoras de la enfermedad.

U

UMAP: *Uniform Manifold Approximation and Projection*.
UMI: *Unique molecular identifier* (identificador molecular único).

V

VEB: Virus Epstein-Barr.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1. Esclerosis múltiple

El sistema inmunológico es el encargado de proteger al organismo de enfermedades e infecciones, para lo que debe ser capaz de diferenciar los antígenos propios de los ajenos, desarrollando tolerancia hacia los primeros y respuestas inmunes contra los segundos. Sin embargo, bajo ciertas circunstancias, se producen defectos en los mecanismos que controlan la autorreactividad, lo que altera la autotolerancia inmunológica y provoca que el sistema inmunitario ataque de forma errónea a las células y tejidos sanos del huésped, causando lo que se conoce como enfermedades autoinmunes. Como consecuencia, se desencadenan respuestas inmunitarias inflamatorias, en su mayoría de manera crónica, que deterioran los tejidos del individuo (1,2).

En función de la extensión de los tejidos afectados, este tipo de patologías se clasifican en dos categorías principales: sistémicas y órgano-específicas. Una característica típica de las enfermedades autoinmunes sistémicas —como la artritis reumatoide, el lupus eritematoso sistémico o el síndrome de Sjögren— es la presencia de anticuerpos en la sangre y su depósito en distintos órganos, lo que provoca un estado inflamatorio generalizado y desencadena una respuesta autorreactiva que no se limita a una zona específica del organismo (3,4). Por el contrario, las enfermedades autoinmunes órgano-específicas están dirigidas a tejidos concretos, de manera que es el tipo de autoantígenos reconocidos lo que determina qué órganos y/o sistemas van a verse afectados. Se caracterizan por una gran heterogeneidad sintomatológica, ya que su presentación clínica depende del lugar en el que se producen los ataques autoinmunes. Además, muestran una naturaleza poligénica y una etiología multifactorial, resultado de la interacción entre la predisposición genética y los factores ambientales (5,6).

Entre las más de 80 enfermedades autoinmunes que se conocen en la actualidad, en su mayoría órgano-específicas, hay patologías que afectan al sistema nervioso central (SNC), el páncreas o la glándula tiroides, entre otros. En este contexto, la esclerosis múltiple (EM) es una enfermedad autoinmune órgano-específica, crónica y neuroinflamatoria que afecta al SNC, específicamente al cerebro y a la médula espinal. Las vainas de mielina, encargadas de aislar y proteger las fibras nerviosas, son el objetivo de los ataques erróneos del sistema inmunológico en la EM. El daño a esta cubierta protectora recibe el nombre de desmielinización, y es el proceso responsable de que las neuronas pierdan su capacidad de transmitir rápida y eficientemente las señales eléctricas. En respuesta a las lesiones inducidas por la inflamación mediada por células T, se produce una gliosis reactiva, caracterizada por la activación y proliferación de las células gliales, lo que se acompaña de pérdida neuronal y una neurodegeneración progresiva (7,8).

La EM presenta una amplia gama de síntomas neurológicos, que varían en función de la región neuronal afectada, la gravedad y la progresión de la enfermedad. Por tanto, las manifestaciones clínicas difieren entre pacientes e incluso cambian durante la progresión temporal de la enfermedad en un mismo paciente. La sintomatología más frecuente incluye fatiga extrema, deterioro cognitivo, entumecimiento y sensación de hormigueo, problemas de visión, incontinencia urinaria, disfunción sexual, debilidad muscular, cambios de humor o mareos (9).

1.1.1. Epidemiología

Según la información más reciente recogida en el Atlas de Esclerosis Múltiple (datos de 2023) (10), la EM afecta a aproximadamente 2,9 millones de personas a nivel mundial, de las cuales casi 60.000 son españolas. Estas cifras sitúan a la EM como la enfermedad inflamatoria más común del SNC. En comparación con la primera edición del Atlas de 2013, se ha registrado un aumento de más de un millón de casos, atribuido principalmente a las mejoras en el diagnóstico y los tratamientos, que han contribuido a una mayor esperanza de vida de los pacientes con EM (10–12).

La prevalencia de la EM no muestra una distribución geográfica homogénea, sino que presenta una considerable variación entre localizaciones. Particularmente, se observa una mayor prevalencia en la población europea y americana, con tasas que oscilan entre los 100 y 300 casos por cada 100.000 personas. En contraste, las zonas del Sudeste Asiático, África y el Pacífico Occidental son las que reportan menor prevalencia, con valores que fluctúan entre los 5 y 10 casos por cada 100.000 personas (Figura 1). Estas discrepancias pueden atribuirse principalmente a las limitaciones de diagnóstico que se presentan en algunas regiones, así como a diferencias en los perfiles étnicos y demográficos entre países (10,11).

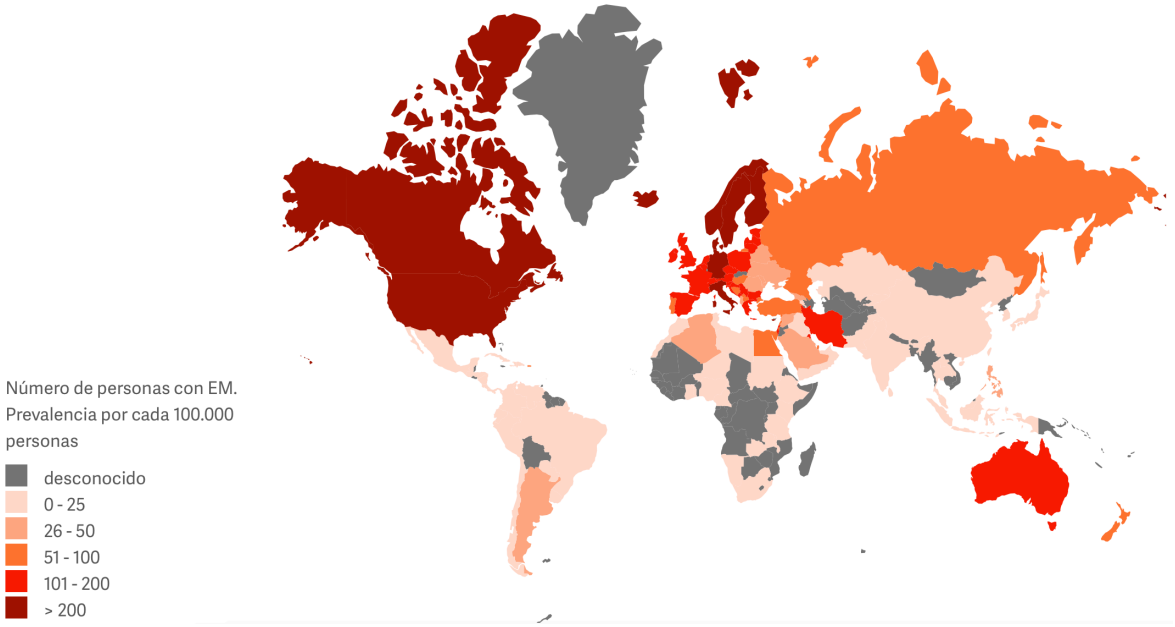

Figura 1. Prevalencia a nivel global de la esclerosis múltiple. El código de colores hace referencia al número de personas de cada país con diagnóstico de esclerosis múltiple por cada 100.000 habitantes según los datos recopilados en la última edición del Atlas de Esclerosis Múltiple. Figurada adaptada de Multiple Sclerosis International Federation (MSIF) (10). EM: Esclerosis múltiple.

Independientemente de la localización geográfica, la EM se manifiesta predominantemente durante los primeros años de vida adulta (entre los 18 y los 40 años), y la edad media de diagnóstico a nivel global se sitúa en 32 años. De hecho, es considerada una de las causas más frecuentes de discapacidad no ocasionada por lesiones traumáticas entre los adultos jóvenes. Sin embargo, es preciso destacar que esta enfermedad no se presenta exclusivamente en personas adultas. En los últimos años, se ha evidenciado un crecimiento en la cantidad de casos de niños y adolescentes con un diagnóstico de EM, con una cifra actual que supera los 30.000 pacientes menores de 18 años (10,12).

1.1.2. Factores de riesgo

La etiología de la EM es aún incierta. En concordancia con lo observado en otras enfermedades de naturaleza autoinmune, se postula que la EM podría ser el resultado de la interacción sinérgica de factores genéticos y ambientales. La principal hipótesis planteada es que, ante una cierta predisposición genética, algunos factores ambientales específicos u otros relacionados con el estilo de vida, podrían actuar como el desencadenante de la respuesta inmunológica aberrante. La EM no se puede considerar una enfermedad hereditaria, pues no se transmite de padres a hijos, aunque existe una vulnerabilidad genética, que sí que puede heredarse. El riesgo de EM se acentúa cuando existe un mayor grado de parentesco con un familiar afectado (13).

Las variantes genéticas (alelos) de más de 200 genes han sido identificadas como potencialmente implicadas en el desarrollo de la EM. Concretamente, algunos estudios de asociación del genoma completo (GWAS, por sus siglas en inglés) han descrito un total de 32 variantes genéticas en los genes de los antígenos leucocitarios humanos (HLA, por sus siglas en inglés), además de 201 variantes genéticas fuera de esta región específica, que están directamente relacionadas con la susceptibilidad de desarrollar EM (14). No obstante, la asociación más habitual con la EM a nivel genético se atribuye a los genes HLA. Este conjunto de genes participa en la presentación de antígenos a los linfocitos T, lo que los hace imprescindibles para discriminar las proteínas del propio organismo de las proteínas extrañas de virus o bacterias. Destaca especialmente el haplotipo HLA-DRB1*15:01 por triplicar el riesgo de EM en individuos que portan al menos una copia de este alelo, siendo además la variante más frecuente en personas con EM. Los polimorfismos en algunos genes relacionados con la inmunidad (como IL7RA o IL2RA), el metabolismo de la vitamina D, el DNA mitocondrial o la reparación del SNC, aunque con menor frecuencia, también se han asociado con un mayor riesgo de EM. Ser portador de estos alelos no se traduce necesariamente en padecer la enfermedad, ya que muchos individuos poseen esta predisposición genética sin nunca manifestar EM (14–18).

Los principales responsables de desarrollar EM son los factores ambientales y el estilo de vida. Uno de los factores de riesgo que más se ha relacionado con la EM es la deficiencia de vitamina D, una molécula que regula la producción de citoquinas antiinflamatorias. A este respecto, algunos estudios han demostrado que la prevalencia de la EM está directamente relacionada con la latitud. Las personas que viven en países situados en latitudes más altas tienen mayor riesgo de desarrollar EM al estar expuestas a menores niveles de luz solar y, por tanto, presentar niveles más bajos de vitamina D. Sin embargo, este riesgo disminuye en países próximos al ecuador, donde los niveles de luz solar son superiores (19). Además, se ha demostrado que la infección previa del virus Epstein-Barr (VEB) aumenta hasta 32 veces el riesgo de EM, lo cual es probablemente debido al mimetismo molecular entre el factor de transcripción EBNA1 del VEB con GlialCAM, una proteína de adhesión de las células gliales cerebrales (20,21). Finalmente, el tabaquismo y la obesidad adolescente, en combinación con factores de riesgo genéticos, son otros de los factores de riesgo de la EM relacionados con el estilo de vida (22).

1.1.3. Fisiopatología y evolución de las lesiones

Los axones de las neuronas están recubiertos por una capa formada por proteínas y una mezcla de lípidos. Esta envoltura, conocida como vaina de mielina, en el SNC es producida por los oligodendrocitos. En un estado fisiológico, la vaina de mielina actúa como un aislante eléctrico, permitiendo la correcta transmisión de los impulsos nerviosos a través de las conexiones neuronales. La EM es un proceso autoinmune provocado por la autorreactividad de las células T contra fragmentos específicos de la mielina, destruyéndola y bloqueando la comunicación entre el SNC y el resto del cuerpo. Sin embargo, todavía no se ha logrado identificar el autoantígeno que desencadena el proceso (23,24).

La barrera hematoencefálica (BHE) es la estructura encargada de regular el paso de sustancias entre la circulación sanguínea y el cerebro. En condiciones fisiológicas, el paso de linfocitos a través de la BHE depende de un mecanismo de transporte asistido por selectinas. Sin embargo, tras la activación de las células T autorreactivas, se generan interrupciones en la BHE que facilitan que los linfocitos T puedan alcanzar el SNC al unirse a determinadas moléculas de adhesión. La entrada de linfocitos T autorreactivos al SNC supone la producción de metaloproteinasas de matriz (MMP) y citoquinas proinflamatorias, que degradan y debilitan aún más la BHE, favoreciendo la infiltración de más tipos celulares inmunitarios. Como consecuencia, los linfocitos T se reactivan e inician una cascada inflamatoria, que resulta en el ataque directo a los oligodendrocitos y las neuronas por parte de los autoanticuerpos de mielina producidos por las células B. Las células mieloides también participan en la patogénesis de la EM al promover la fagocitosis y la desmielinización, provocando inflamación y daño neuronal que inhibe la remielinización (24,25).

Como resultado de la inflamación focal y el daño citotóxico inmunomediado a los oligodendrocitos y axones neuronales, se generan lesiones o placas diseminadas por todo el SNC. Las lesiones de la EM manifiestan un patrón temporal y espacial bien definido, lo que brinda información sobre la evolución de la EM. Las lesiones agudas (EM-A) son las primeras que aparecen y suelen centrarse alrededor de los vasos sanguíneos, lo que supone la acumulación de células B y macrófagos en el espacio perivascular. Se caracterizan por la pérdida activa de mielina debida a la destrucción de oligodendrocitos, con presencia de células fagocíticas de mielina, tales como macrófagos y células de la microglía activadas, que se infiltran a través de la lesión (26–28).

Tras esta fase inicial, las lesiones pueden entrar en un estado crónico activo (EM-CA), en el que el núcleo de la lesión está completamente desmielinizado (pérdida total de oligodendrocitos). En este tipo de lesiones, la desmielinización se asocia con la activación de los astrocitos. Además, se suelen expandir hacia la sustancia blanca de apariencia normal, dando lugar a un borde inflamado bien definido, donde la presencia de células de la microglía activadas es más pronunciada que en el estado agudo (26–28).

Con el tiempo, las lesiones EM-CA pueden pasar a una fase crónica inactiva (EM-CI), en la que el borde no presenta tanta inflamación, lo que se evidencia con una menor infiltración de células inmunes, pero el núcleo de la lesión presenta una densa cicatriz glial (o esclerosis), que le da nombre a esta enfermedad. La formación de estas cicatrices también se asocia a la activación de astrocitos, de igual modo que ocurre en el estado activo, pero en este caso la abundancia de microglía se reduce respecto a las lesiones activas (26–28).

1.1.4. Diagnóstico y tratamientos actuales

Para diagnosticar la EM no se dispone de una única prueba específica. En su lugar, el diagnóstico se basa en la integración de evidencias clínicas con pruebas de imagen y de laboratorio. Se trata de un diagnóstico diferencial enfocado en la exclusión de otras patologías neurológicas con signos y síntomas similares (29).

En cuanto a las evidencias clínicas, se considera una recaída de EM cuando se manifiestan síntomas neurológicos nuevos o recurrentes sin fiebre simultánea durante un período mínimo de 24 horas consecutivas, y están separados de la recaída anterior por un estado de estabilidad de al menos 30 días. En general, los pacientes suelen presentar entre 1 y 2 recaídas al año (30).

Las imágenes de resonancia magnética (IRM) del cerebro y la médula espinal son un componente clave para el diagnóstico diferencial de la EM. Las IRM permiten definir las lesiones producidas en el SNC y determinar su diseminación en el tiempo y el espacio, lo que es imprescindible para discriminar otros tipos de enfermedades neuroinflamatorias. La difusión en el tiempo hace referencia a la aparición de nuevas lesiones en el SNC a lo largo del tiempo, y la difusión en el espacio describe la formación de lesiones en varias regiones anatómicas dentro del SNC, indicando un daño multifocal. En el caso de que la evidencia clínica y las IRM no sean suficientes para confirmar el diagnóstico, se recurre a una examinación del líquido cefalorraquídeo (LCR) con el fin de detectar bandas oligoclonales, indicadoras de inflamación en el SNC (12,29).

Si los resultados de la integración de todas estas pruebas no apuntan a otra patología neurológica, se aplican los criterios McDonald de 2017 para ratificar el diagnóstico de EM. Estos criterios consideran el número de recaídas clínicas, el número y localización de las lesiones en el SNC detectadas en las pruebas de imagen y la presencia de bandas oligoclonales en el LCR (31).

Actualmente, no existe ningún tratamiento capaz de curar la EM. Uno de los elementos que más compromete la calidad de vida de los pacientes con EM es el deterioro cognitivo. Por ello, efectuar un diagnóstico precoz es esencial para poder iniciar de inmediato un tratamiento que disminuya el número de recaídas y minimice el daño neurológico a largo plazo, paliando además la sintomatología asociada. Lo más frecuente son las terapias modificadoras de la enfermedad (TME), que limitan la tasa de recaída y la formación de nuevas lesiones (32). Este tipo de tratamientos no mitigan todos los síntomas, por lo que se suelen combinar con otros medicamentos específicos para combatir los restantes. Asimismo, para no perder fuerza muscular y aliviar algunos síntomas, se suele incentivar la terapia física (33).

1.1.5. Subtipos de esclerosis múltiple

La primera clasificación de la EM en subtipos clínicos, fundamentada exclusivamente en los distintos fenotipos clínicos, fue propuesta en 1996 por el Comité Asesor Internacional sobre Ensayos Clínicos en Esclerosis Múltiple (34). Sin embargo, gracias al incremento en los conocimientos referentes a la patología de la EM, junto a los avances desarrollados en imagen y biomarcadores tisulares, en el año 2013 se realizó una revisión para perfeccionar la nomenclatura (35). Esta última es la clasificación que se emplea hoy en día.

Se denomina Síndrome Clínico Aislado (SCA) a los primeros episodios de desmielinización inflamatoria en el SNC, generalmente monofocales, aunque no se cumplan los criterios de diseminación a lo largo del tiempo como para diagnosticar EM. El riesgo de desarrollar EM se eleva considerablemente en los pacientes que han sufrido previamente un ataque de SCA. En general, el SCA es la primera presentación clínica y tiende a progresar hacia lo que se denomina EM remitente recurrente (EMRR) (30,35).

La EMRR es el diagnóstico inicial más habitual, y se caracteriza por ataques con síntomas neurológicos nuevos o empeoramiento de los ya existentes (recaídas) intercalados con períodos de recuperación parcial o completa sin progresión aparente de la enfermedad (remisión) (Figura 2A). En algunas ocasiones, las recaídas resultan en disfunción neurológica o discapacidad residual, que conlleva a la acumulación del deterioro cognitivo (30,35).

El agravamiento gradual de la EMRR supone la transición al estado de EM secundaria progresiva (EMSP) a los 10-15 años desde el inicio de la enfermedad, especialmente en pacientes que no han sido tratados. La EMSP se caracteriza por una progresión continua con empeoramiento de las funciones neurológicas a lo largo del tiempo. En este caso, también se pueden observar períodos de recaídas y períodos de estabilidad en los que la discapacidad no aumenta (Figura 2B) (30,35).

El último subtipo es la EM primaria progresiva (EMPP), aunque es el menos frecuente. En la EMPP, a diferencia de lo que ocurre en la EMSP, la enfermedad se desarrolla de forma progresiva desde el inicio, es decir, sin estar precedida por períodos de remisión seguidos de exacerbaciones agudas. La progresión de la enfermedad no sigue un patrón uniforme a lo largo del tiempo, sino que también es posible que se produzcan recaídas y períodos de estabilidad (Figura 2C) (30,35).

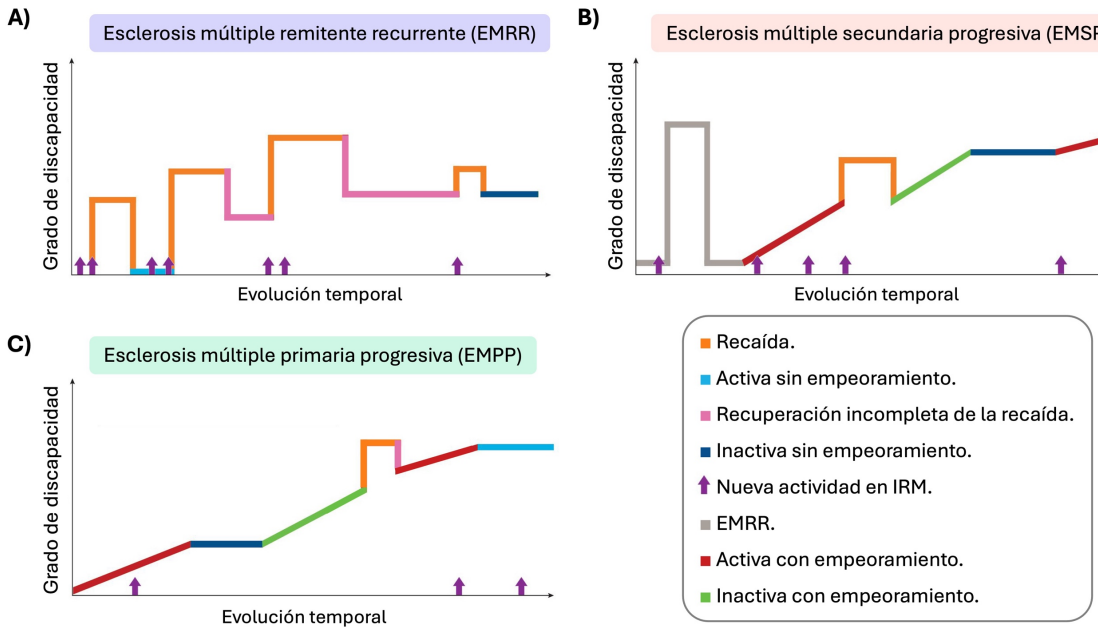

Figura 2. Evolución temporal del grado de discapacidad en los tres subtipos de esclerosis múltiple. A) EM remitente recurrente, B) EM secundaria progresiva y C) EM primaria progresiva. El color de cada una de las líneas representa una fase clínica de EM. El término activa o inactiva hace referencia a nueva actividad en las imágenes de resonancia magnética (IRM). Figura modificada de Klineova et al. (30) con los criterios de Lublin et al. (35). EM: esclerosis múltiple.

1.1.6. Diferencias de sexo en esclerosis múltiple

La variable sexo influye significativamente en el desarrollo, susceptibilidad y evolución de la mayoría de las enfermedades autoinmunes y neurodegenerativas, y la EM no es una excepción (36).

Aunque todavía se desconocen las razones, la prevalencia de la EM manifiesta un sesgo de sexo, siendo más frecuente en mujeres que en hombres. En el territorio español, el 68 % de los casos de EM son diagnosticados en mujeres, frente al 32 % restante que corresponde a hombres. En otras palabras, la proporción de mujeres con EM excede en más del doble a la de hombres (proporción de 2:1). Estos datos ponen de manifiesto las diferencias de sexo en la incidencia de la enfermedad, siendo aún más notorias en las regiones del Sudeste Asiático y el Pacífico Occidental, donde la proporción es de 3 a 4 mujeres con EM por cada hombre. La evaluación de estas proporciones a lo largo del tiempo ha revelado que esta tendencia es debida al aumento en la incidencia de mujeres, y no a una disminución en la cifra de hombres afectados (10,37).

Mientras las mujeres son más susceptibles a padecer EM; los hombres, en promedio, son diagnosticados a edades más avanzadas, tienen una progresión más acentuada de la enfermedad y presentan manifestaciones clínicas de mayor severidad. Además, el sexo masculino se ha asociado con la acumulación más rápida de discapacidad respecto al femenino, lo que se traduce en una conversión del subtipo EMRR al EMSP en tiempos más cortos (36,37)

Las razones de las diferencias de sexo todavía no se comprenden bien, aunque se cree que principalmente están relacionadas con los cromosomas y las hormonas sexuales (36). Por ejemplo, el gen TLR7, codificado en el cromosoma X, se ha identificado como uno de los potenciales responsables de la mayor neurodegeneración observada en los hombres, consecuencia del mayor nivel de expresión en las neuronas masculinas en comparación con las femeninas. Asimismo, debido a la compensación de la dosis del cromosoma X en mujeres, en la que cada célula inactiva aleatoriamente una de las dos copia del cromosoma X, la mitad de las células expresan el alelo X materno y la otra mitad el alelo X paterno; mientras que todas las células masculinas expresan el alelo X materno. Estas diferencias en la dosificación del cromosoma X, aunque todavía no hay evidencia consolidada, también podrían estar implicadas en las diferencias de sexo de la EM (38).

El efecto de los cromosomas sexuales en la susceptibilidad y progresión de la EM está modulado por las hormonas sexuales. En este contexto, algunas hormonas sexuales femeninas, como el estradiol y el estriol, se han asociado con efectos neuroprotectores al inducir la remielinización y mejorar la plasticidad cerebral (39). Además, se ha demostrado que los niveles altos de testosterona en hombres contribuyen a una menor discapacidad, lo que evidencia sus propiedades neuroprotectoras (40).

Sin embargo, aunque las diferencias de sexo en la EM son evidentes, muchos estudios de EM solo se realizan para un sexo u omiten esta variable biológica en sus análisis, creyendo que los hallazgos para un sexo son equivalentes para el sexo opuesto. Es por ello por lo que en el presente estudio se analizaron las diferencias de sexo en EM desde una perspectiva espacial.

1.2. Técnicas ómicas: transcriptómica aplicada al estudio de enfermedades neurodegenerativas

Los avances en Biotecnología de los últimos años han revolucionado el campo de la investigación biomédica, permitiendo responder de manera eficiente las preguntas biológicas desde un enfoque ómico. Las denominadas técnicas ómicas son el conjunto de tecnologías de alto rendimiento enfocadas al análisis masivo de las diferentes macromoléculas biológicas. La implementación de estas técnicas ha permitido identificar prácticamente todas las biomoléculas de la misma naturaleza presentes en una muestra, lo que resulta en la obtención de grandes volúmenes de datos en lapsos de tiempo cada vez menores. El análisis de los datos ómicos generados permite la caracterización del comportamiento de las células, tejidos y órganos a nivel molecular. Por tanto, no solo son útiles para entender los procesos fisiológicos normales, sino que también juegan un papel clave en el estudio de las enfermedades humanas (41,42).

Existe un extenso abanico de disciplinas ómicas, cada una orientada al análisis de una clase específica de moléculas. Las más relevantes son la genómica, la transcriptómica, la proteómica y la metabolómica. A rasgos generales, la genómica estudia las secuencias de ácido desoxirribonucleico (ADN) y los cambios genéticos, la transcriptómica evalúa los niveles de expresión génica, la proteómica cuantifica las proteínas e infiere sus funciones e interacciones, y la metabolómica analiza los metabolitos para componer una "imagen" del estado fisiológico de la célula. La integración de varias técnicas ómicas ofrece una comprensión más profunda de los sistemas biológicos, al reunir distintas capas moleculares de información y abordar su estudio desde una perspectiva holística. Sin embargo, esta integración, en ocasiones, resulta compleja a nivel técnico y computacional (42).

Todas las células de un organismo comparten el mismo material genético, pero no todos los genes se transcriben a ácido ribonucleico (ARN) en todo momento ni en todas las células. El conjunto de transcritos de ARN expresados en una célula u organismo, ya sean codificantes de proteínas como no codificantes, se denomina transcriptoma. La transcriptómica permite extraer información acerca de estos niveles de expresión de un genoma, en un momento dado y bajo unas condiciones fisiológicas o patológicas específicas. Esta estrategia se fundamenta en que la abundancia de ARN mensajero (ARNm) es una representación aproximada de la expresión activa del gen del que proviene, lo que brinda información sobre la biología de un organismo. Identificar qué genes tienen alterada su actividad ante diferentes condiciones es crítico para comprender los mecanismos de enfermedad y definir estrategias terapéuticas dirigidas a revertir los cambios moleculares (43,44).

El inicio de la transcriptómica surgió con el desarrollo de los microarrays, capaces de cuantificar simultáneamente la abundancia de miles de transcritos de una muestra biológica mediante su hibridación con unas sondas específicas de ADN. Su principal inconveniente es que restringen el estudio del transcriptoma a aquellos genes que han sido previamente depositados en el chip. El desarrollo de las técnicas de secuenciación de ARN (RNA-Seq, por sus siglas en inglés), basadas en la secuenciación masiva, permitió superar esta limitación al posibilitar el análisis de todos los transcritos presentes en un tejido o cultivo de células. La metodología de RNA-Seq consiste en generar y secuenciar una copia de ADN sintética y complementaria del ARNm (ADNc), obteniendo el perfil transcripcional equivalente a la expresión media de cada gen en el conjunto de células analizadas (45–47).

Sin embargo, los tejidos están formados por más de un tipo celular, cada uno con su propio transcriptoma. Los métodos tradicionales de secuenciación masiva no permiten explorar esta heterogeneidad celular y el origen de sus cambios. Los avances en microfluídica, que han permitido la reducción al máximo del volumen necesario para preparar las librerías de secuenciación, han posibilitado la secuenciación de ARN de células únicas (scRNA-Seq, por sus siglas en inglés), revelando la diversidad celular de los tejidos. No obstante, esta tecnología no contextualiza la composición celular y la dinámica molecular dentro del microambiente de las lesiones tisulares. La transcriptómica espacial (ST, por sus siglas en inglés) ha emergido como una solución innovadora al incorporar información sobre la organización espacial de los perfiles transcripcionales dentro de un tejido. Esta tecnología ofrece una nueva dimensión para identificar patrones espaciales de expresión génica y evaluar cómo influyen las interacciones celulares en la enfermedad (48–50).

1.2.1. Transcriptómica de núcleo único

El flujo de trabajo de scRNA-Seq inicia con un método de disgregación tisular para disociar las células. Sin embargo, para el estudio de enfermedades neurodegenerativas, como la EM, se suelen emplear muestras congeladas de tejido de cerebro *post-mortem*, donde la recuperación de células intactas se complica debido a la morfología irregular y a la interconexión de las neuronas y otras células cerebrales. En este contexto, surgió la secuenciación de ARN de núcleo único (snRNA-Seq, por sus siglas en inglés) como una variante alternativa a scRNA-Seq en la que se secuencian los núcleos en vez de las células completas. En general, la técnica de elección para estudiar los trastornos neurológicos suele ser snRNA-Seq (51).

La principal diferencia entre ambas técnicas radica en el tipo de transcritos que se analizan: en scRNA-Seq se secuencia el ARN presente en el núcleo, en el citoplasma y en las mitocondrias; mientras que en snRNA-Seq solo se analiza el ARN nuclear. Aunque *a priori* pueda parecer que con esta última aproximación se pierde parte de la información, se han realizado diversos estudios que demuestran que los resultados de ambos procedimientos son totalmente comparables (52–54).

El protocolo de snRNA-Seq comienza con la disociación proteolítica de las células, cuyo objetivo es obtener una suspensión de núcleos, evitando la necesidad de preservar la integridad de las células. La purificación de los núcleos se lleva a cabo por ultracentrifugación, proceso que permite su separación del resto del contenido citoplasmático. Los núcleos únicos se aíslan y capturan, generalmente mediante dispositivos de microfluídica o citometría de flujo. Posteriormente, se extraen las moléculas de ARNm de cada uno de los núcleos individuales y se utilizan como molde para generar las bibliotecas de ADNc. Estas moléculas de ADNc son las que se secuencian (55,56).

Como resultado de la secuenciación, se obtienen archivos que almacenan las lecturas de secuenciación, que deben ser procesadas a nivel bioinformático. Este procesamiento comienza con el control de calidad y la cuantificación del número de lecturas obtenidas para cada transcrito, ya que es una fiel medida del nivel de expresión de la respectiva región del genoma. Los siguientes pasos incluyen la normalización de los datos, la reducción de la dimensionalidad y la anotación de tipos celulares. Los análisis posteriores dependerán del propósito de los investigadores (57).

1.2.2. Transcriptómica espacial

El SNC está compuesto por tejido nervioso, que se caracteriza por ser de los más complejos debido a la conectividad funcional de la gran variedad de tipos celulares que lo conforman. A pesar de que la técnica de snRNA-Seq es una buena aproximación para entender los perfiles transcriptómicos de las células individuales, la ausencia de contexto espacial impide la comprensión integral del sistema. Para entender cómo las diferentes disposiciones de tipos celulares del SNC contribuyen a la funcionalidad biológica específica de cada región cerebral, es crucial correlacionar la identidad de las células obtenidas en snRNA-Seq con su ubicación espacial en el tejido intacto (58,59). Las ómicas espaciales analizan las biomoléculas en su contexto espacial nativo, lo cual es clave para interpretar la biología tisular de las enfermedades neurodegenerativas.

Concretamente, la ST es una técnica desarrollada para el análisis cuantitativo de la expresión del ARNm y su visualización con resolución espacial en el tejido. Esta aproximación, publicada por primera vez en 2016, posibilita la interpretación visual de los mecanismos moleculares subyacentes a la enfermedad y abre un nuevo campo de investigación biomédica (60). Se han desarrollado múltiples tecnologías de ST, que difieren principalmente en la forma de detectar la abundancia de los transcritos y en el nivel de resolución espacial. De este modo, se puede establecer una clasificación en dos categorías principales: ST basada en imágenes y ST basada en secuenciación (61).

En las técnicas de ST basadas en imágenes, los ARNm se detectan mediante imágenes de microscopía de fluorescencia. En función de cómo se cuantifican las moléculas de ARNm, este primer grupo se subdivide, a su vez, en técnicas basadas en hibridación *in situ* de fluorescencia (FISH, por sus siglas en inglés) y en técnicas basadas en secuenciación *in situ* (ISS, por sus siglas en inglés). En el método FISH, los ARNm se visualizan gracias a varias rondas de hibridación con sondas fluorescentes; y en el método ISS se detectan las señales de fluorescencia generadas tras varios ciclos de amplificación de círculo rodante (RCA, por sus siglas en inglés) seguidos de secuenciación en el propio tejido (Figura 3A). Ambos enfoques ofrecen una alta resolución espacial, pudiendo localizar los transcritos incluso a nivel subcelular. Sin embargo, el aumento de los ciclos de hibridación predispone a una mayor probabilidad de error de unión de las sondas, por lo que estas técnicas se limitan a analizar entre cientos y miles de genes (61–63).

En las tecnologías de ST basadas en secuenciación, la identificación y cuantificación de los transcritos se realiza combinando tecnologías clásicas de microarrays y técnicas de secuenciación de próxima generación (NGS, por sus siglas en inglés). En este caso, las muestras de tejido se depositan sobre una matriz de puntos con los reactivos necesarios para secuenciar los ARNm e integrarles unos códigos de barras espaciales con los que determinar posteriormente la localización espacial de cada transcrito y su nivel de expresión (Figura 3B). El nivel de detección de los ARNm y la resolución espacial es menor que en las técnicas basadas en imágenes, pero proporcionan un análisis del transcriptoma completo en tiempos más cortos (61–63).

No obstante, esta clasificación no siempre está clara, puesto que existen tecnologías que combinan ambos enfoques. Por ello, algunos autores han propuesto una clasificación alternativa en dos nuevos grupos: técnicas basadas en la extracción del ARNm (EBT, por sus siglas en inglés) y técnicas no basadas en la extracción del ARNm (NEBT, por sus siglas en inglés) (64). De acuerdo con esta nueva clasificación, la mayor parte de las tecnologías basadas en secuenciación son EBT, mientras que las NEBT engloban principalmente a las técnicas basadas en imágenes.

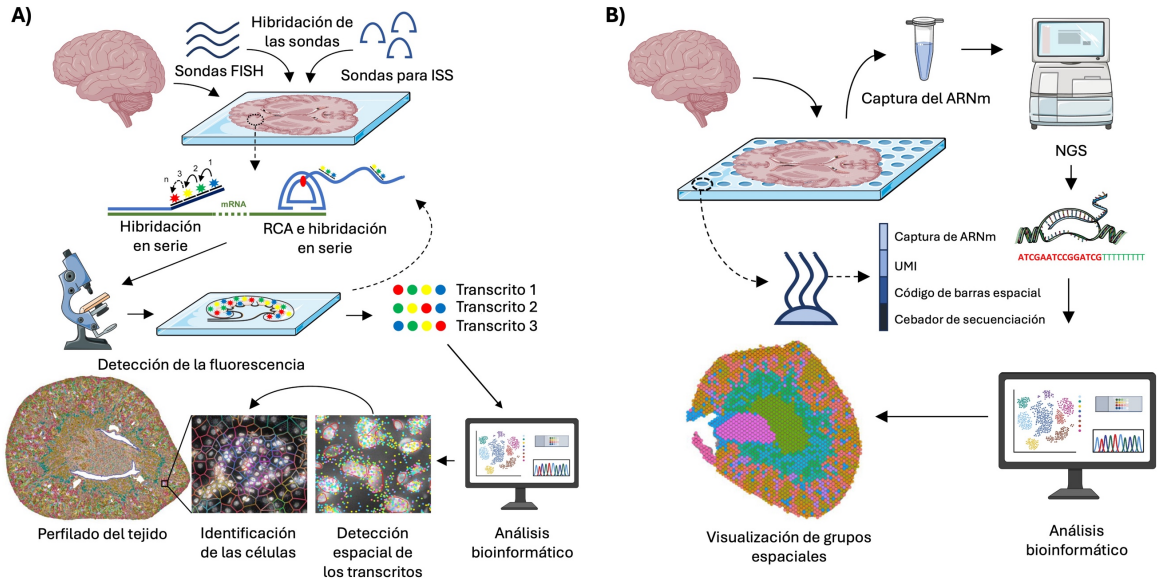

Figura 3. Representación esquemática del flujo de trabajo de las tecnologías de ST. A) Técnicas basadas en imagen. El tejido se coloca sobre un portaobjetos para su hibridación en serie con sondas FISH (hibridación *in situ* de fluorescencia) o con sondas para ISS (secuenciación *in situ*). La identidad de los transcritos se determina mediante imágenes secuenciales con microscopía de fluorescencia. Tras el análisis bioinformático de los datos, se analizan espacialmente los resultados. RCA: amplificación de círculo rodante. B) Técnicas basadas en secuenciación. La muestra se deposita sobre un microarray para la captura del ARNm y su posterior secuenciación de próxima generación (NGS, por sus siglas en inglés). Las sondas contenidas en cada punto del microarray incluyen cuatro secuencias: cola poli(dT) para capturar el ARNm, identificador molecular único (UMI, por sus siglas en inglés) para identificar cada transcrito, código de barras espacial para el mapeo de las localizaciones espaciales y cebador de secuenciación. Tras el análisis bioinformático de los datos, se visualizan espacialmente los grupos espaciales. Figura modificada con BioRender de Isnard et al. (64).

La plataforma comercial más empleada dentro de las tecnologías basadas en secuenciación, y a partir de la cual se han obtenido los datos utilizados en este trabajo, es *Visium*, comercializada por la empresa *10x Genomics*. El fundamento de *Visium* sigue el concepto de ST introducido por primera vez en 2016 por Ståhl et al. (60). Esta tecnología permite medir el ARNm total en secciones de tejido intacto y mapear la ubicación espacial donde se está produciendo la actividad genética. El portaobjetos incluye cuatro áreas de captura, cada una formada por un microarray cuadrado de 6,5 mm de lado, en cada uno de los cuales hay un total de 4.992 puntos de expresión o localizaciones de medición de 55 µM de diámetro, y separados por una distancia de 100 µM de centro a centro. Estos puntos, aunque físicamente son circulares, se organizan en una red hexagonal, asociándose cada uno de ellos a seis vecinos directos. Cada uno de estos puntos contiene unas sondas para capturar y secuenciar el ARNm, que además incluyen un código de barras espacial único en forma de secuencia y un identificador molecular único (UMI, por sus siglas en inglés) para identificar cada transcrito (Figura 4) (65,66).

El tejido, previamente fijado y teñido con hematoxilina y eosina (H&E), se coloca sobre el microarray y se permeabiliza para extraer todo el ARNm celular. El ARNm liberado hibrida con las sondas de captura y se retrotranscribe a ADNc para su secuenciación, etiquetando además cada molécula de ARNm con un UMI. Todas las moléculas de ADNc sintetizadas a partir del ARNm capturado en un punto específico comparten el mismo código espacial, por lo que tras su secuenciación se puede mapear el punto original del que proviene cada lectura y, por tanto, inferir su localización espacial en la sección de tejido. La combinación del código de barras espacial con el UMI de cada molécula asegura la especificidad de los datos (66).

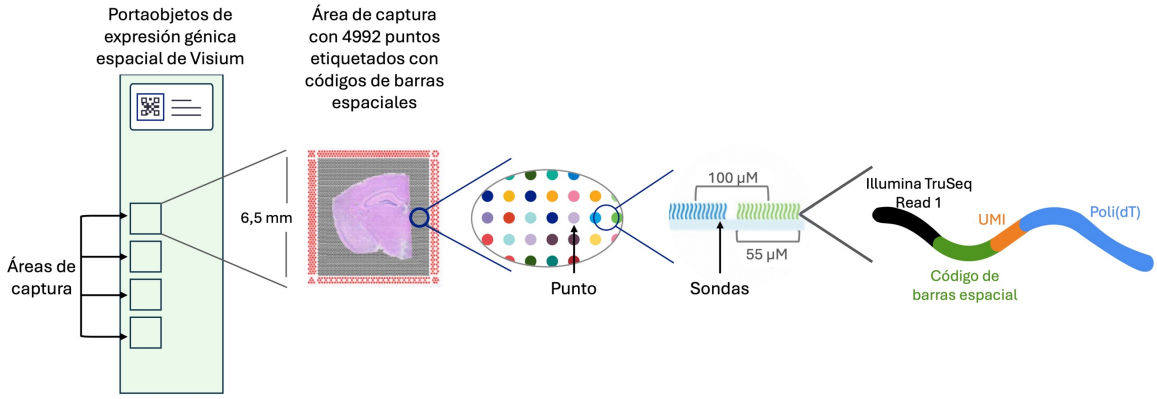

Figura 4. Esquema técnico del portaobjetos de expresión génica espacial de *Visium* (*Visium Spatial Gene Expression Slide*). Cada lámina tiene cuatro áreas de captura, cada una formada por un microarray de casi 5.000 puntos de medición. Cada punto contiene múltiples sondas, que incluyen una combinación de cuatro secuencias: cebador *Illumina TruSeq* para la secuenciación, código de barras espacial (único para cada punto), identificador molecular único (UMI, por sus siglas en inglés) (diferente en cada sonda) y cola poli(dT). Figura modificada de BMKGENE (65).

La principal limitación de *Visium* es la baja eficiencia de detección, evidenciada por el número de genes capturados, que es inferior al estimado en las muestras. Debido a las dimensiones de los puntos de captura, en cada uno de ellos puede haber entre 1 y 10 células. Por tanto, otra desventaja de *Visium* es que carece de resolución espacial a nivel de célula única. No obstante, gracias al corto tiempo de ejecución y a la capacidad de analizar el transcriptoma completo (y no solo unos miles de genes) tanto de muestras frescas congeladas como de muestras fijadas con formalina e incluidas en parafina (FFPE, por sus siglas en inglés), sigue siendo la técnica de elección en la mayoría de los estudios (62).

Independientemente de la tecnología de ST empleada o el tipo de tejido, la metodología experimental siempre requiere de un análisis bioinformático ulterior, que permita la interpretación y visualización de los datos. Los datos brutos generados deben ser organizados en una matriz de conteos (o matriz de expresión génica) de tantas filas como genes se hayan detectado, y de tantas columnas como ubicaciones espaciales se hayan medido. Una vez generada esta matriz de expresión, comienza el análisis bioinformático *per se* (67). Las etapas principales del análisis en ST suelen ser una práctica estandarizada, que se desarrolla a través del siguiente flujo de trabajo: control de calidad, normalización, selección de genes altamente variables, reducción de la dimensionalidad, agrupamiento, identificación de los tipos celulares mayoritarios (deconvolución), análisis de patrones espaciales e identificación de dominios espaciales. Actualmente, existen múltiples paquetes de *software* que permiten realizar el análisis computacional, destacando *Seurat* (68) y *Giotto* (69) en lenguaje R y *Scanpy* (70) en Python. Cada una de estas herramientas integra su propio flujo de trabajo y sus propias funciones para realizarlo. Uno de los desafíos más grandes en el ámbito de la ST es este análisis bioinformático, debido al gran número de flujos de trabajo disponibles y a la complejidad de los algoritmos subyacentes.

2. Objetivos

La ST es una herramienta que facilita la comprensión de los mecanismos subyacentes a la enfermedad, permitiendo el estudio simultáneo de la expresión génica y la organización espacial de los tejidos. En línea con el área de investigación del Laboratorio de Biomedicina Computacional del Centro de Investigación Príncipe Felipe (CIPF), este Trabajo de Fin de Máster se enmarca en una iniciativa para aplicar el potencial de la ST para caracterizar las diferencias de sexo en enfermedades neurodegenerativas mediante la implementación de un flujo de trabajo reproducible.

En este contexto, el objetivo principal de este trabajo es explorar el paquete *Giotto* (69) para el análisis completo de datos de ST y, mediante un abordaje *in silico*, caracterizar las diferencias de sexo en la EM con una resolución espacial. Para ello, se plantearon los siguientes objetivos específicos:

1. Revisión sistemática y selección de un estudio de ST centrado en la EM, cuyos datos sean de acceso público, y que incluya además información sobre el sexo de los donantes en los metadatos.
2. Comparación de las diferentes metodologías y algoritmos implementados en *Giotto* para cada paso del análisis bioinformático de los datos seleccionados.
3. Deconvolución de los tipos celulares y definición de regiones espaciales en el tejido en función de los perfiles transcriptómicos y su localización espacial.
4. Identificación de las diferencias de expresión génica significativas y patrones espaciales sexo-específicos en las lesiones de EM.

3. Materiales y métodos

La metodología empleada para la consecución de los objetivos se basa en un flujo de trabajo que consta de nueve etapas principales (Figura 5): 1) revisión sistemática, 2) control de calidad, 3) normalización, 4) selección de genes altamente variables, 5) reducción de la dimensionalidad, 6) agrupamiento según los perfiles de expresión, 7) anotación de tipos celulares utilizando un conjunto de datos de snRNA-Seq como referencia, 8) identificación de patrones espaciales y detección de dominios espaciales, y 9) análisis de expresión diferencial entre sexos y regiones tisulares específicas..

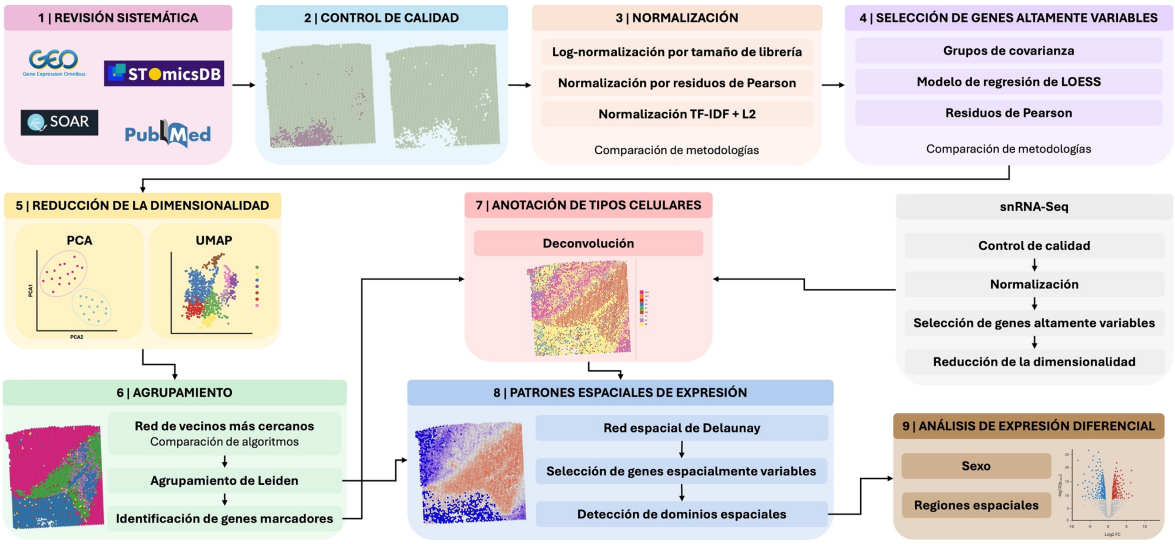

Figura 5. Flujo de trabajo. Representación esquemática del análisis bioinformático llevado a cabo en el presente trabajo: 1) Revisión sistemática para la selección de un estudio de transcriptómica espacial en esclerosis múltiple. 2) Control de calidad para filtrar los *spots* y genes que no cumplen los criterios de calidad. 3) Comparación de metodologías de normalización, 4) Comparación de varias metodologías de selección de genes altamente variables, 5) Reducción de la dimensionalidad, 6) Agrupamiento en función de los perfiles de expresion génica, 7) Anotación de la proporción de tipos celulares de cada *spot* (deconvolución), empleando para ello un conjunto de datos de snRNA-Seq como referencia, 8) Identificación de los patrones espaciales de expresión y detección de dominios espaciales considerando los perfiles de expresión y las localizaciones espaciales, y 9) Análisis de expresión diferencial para identificar diferencias de sexo entre regiones espaciales. PCA: *Principal Component Analysis*. UMAP: *Uniform Manifold Approximation and Projection*. snRNA-Seq: secuenciación de ARN de núcleo único.

A lo largo del trabajo, los puntos de medición de *Visium* se denominarán *spots*, en línea con la nomenclatura más común para referirse a las localizaciones espaciales dentro del tejido. Asimismo, se empleará el término "gen" para referirse a las regiones del genoma en las que se hayan cuantificado transcritos, independientemente de si codifican o no proteínas. Por último, la definición de hombre o mujer se estableció en función de la variable sexo, asumiendo que las variaciones cromosómicas y las hormonas gonadales son las principales responsables de las diferencias fisiológicas y biológicas entre hombres y mujeres; en contraposición a la combinación de factores sociales, culturales y ambientales que determinan el género (71).

3.1. Disponibilidad del código y recursos computacionales

Antes de comenzar con la descripción de la metodología, en los siguientes párrafos se expone información general acerca de la disponibilidad del código generado y de los recursos computacionales utilizados para el desarrollo de este trabajo.

Todo el análisis bioinformático se desarrolló en lenguaje de programación R (versión 4.4.2) (72). Se seleccionó este lenguaje por su robustez y extendido uso en el análisis de datos transcriptómicos. La versión de los paquetes citados a lo largo del trabajo, así como la disponibilidad del código implementado se encuentra en el repositorio público de GitHub: https://github.com/nataliadelrey/MS_SpatialTranscriptomics. A través de este enlace, cualquier usuario puede acceder al material, garantizando la reproducibilidad de los resultados obtenidos.

El gran volumen de datos derivado de las estrategias de ST requiere de una capacidad computacional de almacenamiento y procesamiento poco amigable con los ordenadores personales. Por tal razón, para el análisis bioinformático se empleó la infraestructura computacional del Centro de Investigación Príncipe Felipe (CIPF). Este clúster de computadores cuenta con 20 nodos, que permiten disponer de 744 unidades centrales de procesamiento (CPUs, por sus siglas en inglés). En conjunto, posee una memoria RAM (del inglés *Random Access Memory*) de 9 TeraBytes y una capacidad de almacenamiento de 1 PetaByte. Para gestionar los recursos de *hardware* y trabajar de forma ordenada con el resto de los usuarios, los diferentes *scripts* se ejecutaron a través del sistema de colas SLURM (del inglés *Simple Linux Utility for Resource Management*).

Giotto es un paquete de R desarrollado específicamente para el análisis integral y la visualización interactiva de datos transcriptómicos espaciales sin necesidad de instalar paquetes adicionales (69). Además, implementa algoritmos novedosos para abordar tareas propias de ST, como la detección de patrones espaciales o la identificación de interacciones célula-célula. Aunque otros paquetes como *Seurat* (68), originalmente implementados para scRNA-Seq, también han incorporado métodos para el análisis espacial, *Giotto* presenta mayor grado de especialización. Por todo ello, *Giotto* fue el paquete de elección para llevar a cabo el análisis bioinformático completo de los datos de ST. Puesto que prácticamente la totalidad de las funciones empleadas para este trabajo pertenecen a *Giotto*, se omitirá la especificación del paquete del que proceden. En caso de emplear funciones implementadas en otros paquetes, se indicará de forma explícita.

3.2. Revisión sistemática

Una revisión sistemática es un procedimiento que permite recopilar y analizar de manera rigurosa toda la evidencia científica generada hasta el momento sobre un área específica de investigación. Este primer análisis es fundamental para conocer el estado del arte de un tema de interés (ST en EM, en este caso) y analizar los datos de los estudios seleccionados. Para asegurar la objetividad, transparencia y reproducibilidad de los resultados, la revisión sistemática se realizó siguiendo las directrices de la declaración PRISMA (del inglés *Preferred Reporting Items for Systematic reviews and Meta-Analyses*), establecidas en 2009 (73) y actualizadas en 2020 (74).

El objetivo era seleccionar un estudio de ST que permitiera explorar la potencialidad de *Giotto* para realizar una caracterización de las diferencias de sexo en EM en un contexto unicelular y espacial (pregunta de investigación), de acuerdo con los criterios de inclusión y exclusión que enmarcaron las características requeridas para seleccionar los estudios. A través de una búsqueda en diferentes bases de datos y repositorios públicos, se identificaron los estudios que pudieran responder a la pregunta planteada, filtrando según los criterios de inclusión previamente establecidos (Tabla 1). De este modo, se recopilaron todos los estudios de ST centrados en EM, realizados en humanos o en modelos murinos, y que incluyeran tanto pacientes como individuos sanos.

Tabla 1. Criterios de inclusión utilizados para la revisión sistemática.

Criterio de inclusión	Descripción
Tipo de estudio	Transcriptómica espacial
Organismo	*Homo sapiens* o *Mus musculus*
Condición	Pacientes con esclerosis múltiple e individuos sanos.

La búsqueda se efectuó en las bases de datos públicas *Gene Expression Omnibus* (GEO) (75), *Spatial transcriptOmics Analysis Resource* (SOAR) (76) y *Spatial TranscriptOmics DataBase* (STOmicsDB)[1] (77); en el motor de búsqueda PubMed y en la herramienta de búsqueda de Google. Debido a la diversidad de formatos de las bases de datos y repositorios consultados, la estrategia de búsqueda se adaptó a cada plataforma. En GEO se realizó una búsqueda avanzada, empleando las palabras clave "*multiple sclerosis*" y "*spatial transcriptomics*", para identificar todos los conjuntos de datos disponibles. En SOAR y STOmics DB se revisó el listado de datos depositados relacionados con EM. En PubMed y Google se introdujeron los términos clave "*multiple sclerosis*" y "*spatial transcriptomics*", y se revisó la bibliografía resultante.

Todos los estudios identificados fueron evaluados de forma manual, descartando aquellos que cumplían algún criterio de exclusión (Tabla 2). Se excluyeron los estudios que no incluían información sobre el sexo de los participantes o que no ofrecían acceso público a sus datos. La tecnología de ST es relativamente novedosa y costosa, lo que supone que este tipo de trabajos cuenten con un número limitado de muestras. Sin embargo, para garantizar una robustez estadística a la hora de realizar el análisis comparativo, se exigió que hubiera al menos tres muestras por sexo en ambos grupos experimentales (individuos sanos y pacientes). Se priorizaron tanto los estudios que disponían de un *dataset* de snRNA-Seq emparejado con los datos de ST, necesario para caracterizar la diversidad celular dentro del tejido, como los estudios generados con la plataforma *Visium*, la tecnología de ST más extendida desde su comercialización.

Tabla 2. Criterios de exclusión utilizados para la revisión sistemática.

Criterio de exclusión	Descripción
Sexo	Ausencia de información sobre el sexo de los individuos.
Disponibilidad de datos	Ausencia de datos depositados en algún repositorio público*.
Tamaño muestral	Menos de tres muestras por condición y sexo**.
Diseño experimental	Ausencia de un conjunto de datos de snRNA-Seq generado a partir de los mismos individuos.
Tecnología de ST	Datos no generados con *Visium*.

* Se requieren, al menos, las imágenes de las tinciones de hematoxilina y eosina del tejido, los metadatos de las muestras y los archivos de expresión asociados a las ubicaciones espaciales medidas en el tejido.

** En caso de que ningún estudio cumpla este criterio, se considerarán aquellos en los que el número de muestras esté equilibrado entre sexos.

ST: *Spatial Transcriptomics* (transcriptómica espacial).

[1] SOAR y STOmicsDB son dos bases de datos específicas de datos de transcriptómica espacial, por lo que todos los *datasets* almacenados en ambas bases de datos han sido generados con esta tecnología.

Los datos del estudio seleccionado se descargaron de GEO (GSE279183) y almacenaron en el clúster computacional del CIPF para ejecutar el análisis bioinformático que permitiera dar respuesta a la pregunta inicialmente planteada. Se obtuvieron tanto los datos de snRNA-Seq como los de ST. Aunque el formato de los datos depositados en repositorios públicos suele ser variable, en el caso de los datos de ST generados con *Visium*, es frecuente encontrar los archivos generados por el *software* de análisis *Space Ranger*, desarrollado por la empresa *10x Genomics*. *Space Ranger* es una herramienta que permite procesar los datos espaciales de *Visium* y mapear la expresión de las lecturas transcriptómicas con su localización espacial en la imagen de microscopía de las tinciones H&E del tejido. Cada muestra se analiza de manera independiente, generando un directorio específico para cada una de ellas con los resultados: áreas de tejido detectadas y analizadas, matriz espacial de conteos (*genes x spots*), identificadores de los *spots*, identificadores de los genes y coordenadas de las posiciones de los *spots*, entre otros.

3.2.1. Importación y organización de los datos en R

El paquete *Giotto* ofrece varias funciones específicas para crear directamente objetos de clase *Giotto* a partir de los datos generados por diferentes plataformas de ST. Concretamente, la función *createGiottoVisiumObject()* importa los datos generados con *Visium*, para lo que extrae automáticamente toda la información necesaria desde las carpetas creadas por el *software* de análisis *Space Ranger*.

Los ficheros de ST descargados tras la revisión sistemática fueron organizados en el formato de salida de *Space Ranger*. A partir de ellos, se generó un objeto *Giotto* para cada muestra. Esta clase de objetos almacenan la matriz de expresión, las coordenadas espaciales, las imágenes de microscopía de los tejidos, los metadatos de los genes y los metadatos de los *spots*[2]. Además, también permiten guardar los resultados obtenidos a lo largo del análisis (Figura 6).

Los metadatos de los *spots* carecían de información relacionada con el individuo del que provenía cada sección de tejido. Por este motivo, estas variables se recuperaron del material suplementario de la publicación original y se añadieron *a posteriori* a cada objeto *Giotto*. Para ello, con la función *createcellMetaObj()*, se creó un nuevo objeto de clase *CellMetaObj* para cada muestra, en el que se incorporaron los metadatos correspondientes. Tras ello, cada *CellMetaObj* se integró en su respectivo objeto *Giotto* con la función *setCellMetadata()*. Disponer de esta información es fundamental para realizar todos los análisis posteriores e interpretar los resultados. Como identificador (ID) de los genes se empleó el GENE SYMBOL.

Las matrices de *Visium* están compuestas por 4.992 *spots*, pero no todos ellos son cubiertos por la sección de tejido, de manera que quedan *spots* "vacíos" que, al no haber sido secuenciados, carecen de información relevante y, sin embargo, aumentan el coste computacional derivado de los análisis bioinformáticos. Además, uno de los problemas más habituales en ST es la difusión de los transcritos durante el proceso de permeabilización y secuenciación del tejido, propagándose a áreas donde no se encontraban originalmente. Esto puede provocar que se detecten genes en posiciones que no corresponden con la superficie real del tejido (78).

[2] El término metadatos se refiere al conjunto de variables que caracterizan el elemento de interés.

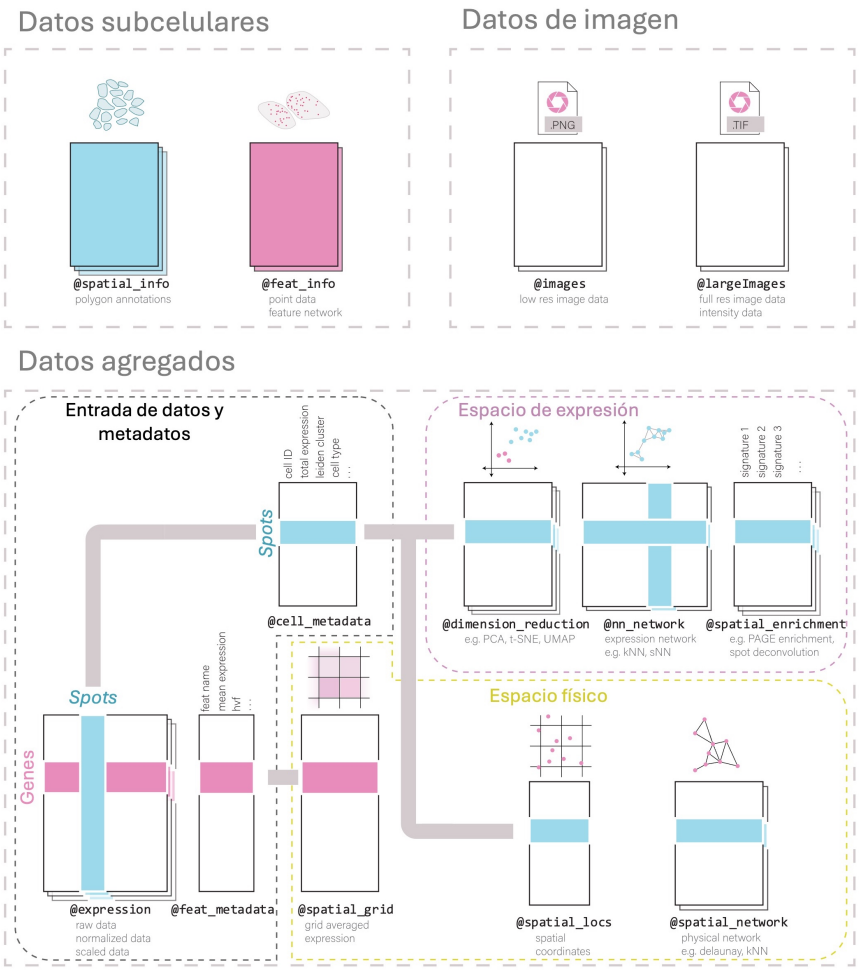

Figura 6. Esquema organizativo de los principales componentes (*slots*) de un objeto de clase *Giotto*. Hay un primer componente que almacena los datos subcelulares (anotaciones espaciales o metadatos subcelulares). Otro componente guarda las imágenes del tejido, que deben estar alineadas con las coordenadas espaciales. El tercer componente contiene a su vez tres subapartados: entrada de datos y metadatos (matriz de expresión génica y metadatos de los genes y *spots*), espacio de expresión (resultados de la reducción de dimensionalidad, del enriquecimiento espacial y de las redes de vecinos más cercanos) y espacio físico (coordenadas espaciales de los *spots* y redes espaciales). Figura adaptada de Del Rossi et al. (79).

Para abordar estos aspectos, el *software* de análisis *Space Ranger*, a partir de la imagen de la tinción H&E del tejido, realiza un proceso de alineación, detección y segmentación que permite identificar las regiones ocupadas por tejido. Esta información es almacenada en uno de los ficheros de salida de *Space Ranger*, y es automáticamente incorporada en los metadatos del objeto *Giotto* durante su creación. A fin de reducir el gran coste que suponía el almacenamiento y computación de los datos de ST, así como los posibles artefactos derivados de la difusión de los transcritos, todos los *spots* de cada muestra sobre los que no se distribuía la sección de tejido dispuesta en el microarray de *Visium* fueron eliminados del conjunto de datos, conservando únicamente los correspondientes al área real del tejido.

Para los análisis posteriores se requería de un único objeto *Giotto* que integrase la expresión génica y la información espacial de todas las muestras, garantizando que todos los análisis fueran idénticos y permitiendo la capacidad de realizar comparaciones directas entre ellas.

La agrupación se realizó con la función *joinGiottoObjects()*, la cual admite tres formas de realizar la unión, que difieren en cómo manejan las posiciones espaciales: *shift*, *z_stack* y *no_change*. El método *shift* desplaza mediante vectores numéricos la localización espacial de cada muestra a lo largo del eje de coordenadas *x* o *y*. Con *z_stack* se combinan las localizaciones espaciales incorporando una tercera dimensión (eje *z*), pero sin alterar los ejes *x* e *y*. Este tipo de unión resulta de interés para elaborar una reconstrucción tridimensional de muestras de tejido procedentes del mismo donante. Por último, con *no_change* no se aplican cambios, lo que puede provocar la superposición de las muestras en el espacio de coordenadas. Dada la naturaleza de las muestras empleadas, el método de combinación elegido fue *shift*, añadiendo una separación de 1.000 unidades a lo largo del eje *x* entre cada muestra.

En *Visium*, cada sección de tejido es secuenciada de manera independiente, por lo que en cada muestra se detectarán unos genes determinados, que no necesariamente coinciden entre todas ellas. Al combinar los objetos *Giotto* se genera una matriz de conteos que incluye todos los genes únicos detectados, independientemente de si inicialmente estaban presentes en todas las muestras o no. Por tanto, si un gen no se expresa en algunas muestras, pero sí en otras, se asigna un valor de 0 a todos los *spots* de aquellas muestras que no presentan dicho gen. Para evitar la pérdida de conteos de los genes poco expresados solamente en algunas muestras durante el filtrado posterior, se decidió construir el objeto combinado de forma previa a este.

3.3. Control de calidad

La calidad de los datos de ST está directamente influenciada por el flujo de trabajo experimental, la pericia del operador manual o el propio secuenciador, entre otros factores. Por tanto, evaluar la calidad de los datos es imprescindible para asegurar que las diferencias observadas entre muestras son atribuibles a factores biológicos, y no a artefactos técnicos o experimentales.

Los datos de ST siempre se preprocesan con algoritmos como *Space Ranger* para determinar la expresión génica y asociar cada transcrito a un determinado *spot*. Este tipo de *software* elimina aquellas lecturas que no conservan la secuencia que posibilita su localización espacial, que no pueden asociarse a un único *spot* o que no mapean con una única secuencia en el genoma de referencia. De este modo, únicamente las lecturas que superan todos los filtros son anotadas e incorporadas en la matriz de expresión. Cabe destacar que todo este proceso corresponde con la fase de preprocesamiento primaria, y los datos brutos utilizados en este trabajo ya habían pasado por ella, por lo que se procedió directamente al control de calidad a nivel de *spots* y genes.

En primer lugar, se utilizaron las medidas de expresión génica para establecer los criterios de calidad que debía cumplir cada *spot* individual, generalmente definidos por umbrales mínimos relacionados con la cantidad y tipo de genes anotados. Con la función *addCellStatistics()* se calculó el número de genes detectados en cada *spot* y la suma total de sus conteos (tamaño de librería). Valores pequeños en estas métricas pueden indicar baja carga de material genético. Puesto que se considera que un mayor número de células por *spot* se traduce en más conteos y genes diferentes, lo ideal sería normalizar estas métricas en base al número de células capturadas en cada *spot*, identificando localizaciones que, aunque tengan pocas lecturas o genes, también tienen pocas células. Sin embargo, esta normalización no pudo realizarse al no disponer de esta información.

Por otro lado, se calculó la proporción de conteos de genes mitocondriales con la función *addFeatsPerc()*. Un elevado número de este tipo de lecturas se suele asociar a daño celular. Para caracterizar la distribución de los *spots* de baja calidad en el tejido e identificar las regiones espaciales con propiedades anómalas, se visualizó el patrón espacial de todos los indicadores de calidad con la función *spatPlot2D()*.

Para el control de calidad a nivel de gen, se determinaron algunas estadísticas con la función *addFeatStatistics()*, tales como el número y porcentaje de *spots* en los que se detectó cada gen, su expresión promedio y su expresión total. Genes con bajos valores en estas métricas no aportan información al estar subrepresentados, por lo que pueden ser eliminados del conjunto de datos.

Como se ha mencionado, el filtrado permite eliminar *spots* y genes de baja calidad, que pueden influir negativamente en los análisis posteriores al introducir ruido técnico y ocultar las señales biológicas. Sin embargo, la eliminación de los *spots* de baja calidad podría afectar a pasos posteriores que implican la construcción de redes espaciales, lo que afectaría directamente a la comprensión del contexto espacial al generar huecos artificiales en el tejido. Para abordar esta situación es fundamental no aplicar criterios de filtrado excesivamente estrictos, asegurando el equilibrio entre la calidad de los datos y la preservación de la integridad espacial de las muestras. Por ello, adicionalmente, con la función *filterCombinations()* se representó de manera gráfica el número de genes y *spots* que se descartarían al aplicar distintas combinaciones de filtrado en base a los indicadores calculados previamente.

Tras la evaluación integral de las métricas de calidad calculadas y las visualizaciones exploratorias, se filtraron los *spots* con menos de 100 genes expresados y los genes que no se detectaron en al menos 10 *spots*, empleando para ello la función *filterGiotto()*. Cabe destacar que los umbrales dependen de la plataforma empleada, el tipo de tejido y su calidad. No hay un protocolo de filtrado estandarizado, ya que debe adaptarse al conjunto de datos que se está analizando y, por tanto, es habitual que cada estudio concrete sus propios criterios.

3.4. Normalización

Con objeto de contrarrestar el ruido técnico o el posible sesgo derivado de la profundidad de secuenciación (cantidad de veces que se secuencia una posición específica o nucleótido), tras el paso de filtrado, es necesario ajustar el total de conteos detectados en cada *spot*, proceso que recibe el nombre de normalización. Los métodos de normalización más extendidos en ST heredan de los algoritmos desarrollados para scRNA-Seq, de manera que cada *spot* se modela análogo a una célula única (78).

Para evaluar cuál era el mejor enfoque para normalizar los datos de ST, se llevó a cabo un análisis comparativo de varias técnicas de normalización implementadas en el paquete *Giotto*: log-normalización por tamaño de librería (69), normalización por residuos de Pearson (80) y normalización TF-IDF (81) seguida de una normalización L2 (82).

La metodología más habitual, y la que ofrece *Giotto* de forma predeterminada, es la log-normalización por tamaño de librería. En ella, se normalizan los conteos de todos los genes de cada uno de los *spots* ($x_{i,j}$) por el tamaño de librería de cada *spot* individual, que representa el sesgo técnico de su procesamiento, y se multiplica por un factor de escala (k). Al valor resultante se le suma una constante de pseudo-conteo (b) y se le aplica una transformación logarítmica en base 2 (Ecuación 1) (69,79). La constante de pseudo-conteo posibilita el cálculo del logaritmo en los valores de expresión igual a 0, que son la mayoría en este tipo de datos. Para la normalización con este método, se mantuvieron los valores por defecto: 6.000 como factor de escalado y 1 como pseudo-conteo (de este modo, los conteos con valor 0, seguirán siendo 0 tras la normalización).

$$x'_{i,j} = log_2\left(\frac{x_{i,j}}{\sum_i x_{i,j}} \times k \ + b\right) \qquad \text{Ecuación 1}$$

Otra normalización implementada en *Giotto* es la normalización por residuos de Pearson, un modelo estadístico que considera que los conteos siguen una distribución binomial negativa. Esta aproximación parte del supuesto de que un gen representa una fracción (p_i) del total de conteos de un *spot* (n_j), a partir de lo que calcula el valor de expresión esperado para cada posición de la matriz de conteos ($\mu_{i,j}$) (Ecuación 2). El residuo de Pearson para cada gen en cada *spot* ($z_{i,j}$) es la diferencia entre el valor de expresión observado ($x_{i,j}$) y el valor esperado, dividida por la desviación estándar de la diferencia, que es además ajustada por un parámetro de dispersión θ (Ecuación 3) (80). En este caso, como parámetro de dispersión se usó el valor por defecto de la función ($\theta \ = \ 100$). El resultado normalizado refleja cuánto se desvía la expresión real de lo esperado, por lo que es una métrica estadística. Por tanto, estos valores favorecen la detección de genes variables y la reducción de dimensiones (valores absolutos más altos equivalen a una mayor desviación), pero no se recomienda para los análisis de expresión diferencial al no ser valores de expresión.

$$\mu_{i,j} = p_i \cdot n_j = \frac{\sum_j x_{i,j}}{\sum_{i,j} x_{i,j}} \cdot \sum_i x_{i,j} \qquad \text{Ecuación 2}$$

$$z_{i,j} = \frac{x_{i,j} - \mu_{i,j}}{\sqrt{\mu_{i,j} + {\mu_{i,j}^2}/{\theta}}} \qquad \text{Ecuación 3}$$

En situaciones en las que los datos son muy ruidosos, se suele aplicar una normalización TF-IDF (del inglés *Term Frequency - Inverse Document Frequency*). Este método se basa en conceptos de procesamiento del lenguaje natural para identificar genes muy expresados en muestras específicas, pero que no se expresan en gran medida en todo el conjunto de datos. Esta normalización calcula, para cada entrada de la matriz de conteos, la frecuencia del término (TF, por sus siglas en inglés) como el valor de expresión sin normalizar ($x_{i,j}$) entre el tamaño de librería (Ecuación 4) y la frecuencia inversa del documento (IDF, por sus siglas en inglés) como el cociente entre el número total de *spots* (n_{spots}) y el total de conteos del gen (Ecuación 5). Finalmente, el valor normalizado TF-IDF se obtiene multiplicando la TF por el logaritmo de la IDF más 1 (para evitar valores indefinidos) (Ecuación 6) (81).

$$TF_{i,j} = \frac{x_{i,j}}{\sum_i x_{i,j}} \qquad \text{Ecuación 4}$$

$$IDF_{i,j} = \frac{n_{spots}}{\sum_j x_{i,j}} \qquad \text{Ecuación 5}$$

$$TFIDF_{i,j} = TF_{i,j} \times log\left(IDF_{i,j} + 1\right) \qquad \text{Ecuación 6}$$

Después de la normalización TF-IDF, es habitual realizar una normalización L2 (o euclidiana), en la que la expresión de cada gen en cada *spot* se escala de tal manera que la suma de los cuadrados de las expresiones en cada *spot* sea 1, facilitando así las comparaciones entre *spots* con diferente profundidad de secuenciación (Ecuación 7) (82).

$$x'_{i,j} = \frac{x_{i,j}}{\sqrt{\sum_i x_{i,j}^2}} \qquad \text{Ecuación 7}$$

Además de estos métodos, *Giotto* ofrece otros tres tipos de normalización. Uno de ellos es la transformación de Arcsinh con cofactor C, que se utiliza en proteómica espacial o en técnicas basadas en imagen (83). El segundo es la normalización empleada de forma específica para datos generados con osmFISH (técnica basada en imagen) (84). Puesto que se estaba trabajando con datos transcriptómicos de *Visium* (técnica basada en secuenciación), estos dos tipos de normalización no se exploraron. Por último, *Giotto* incluye una normalización por cuantiles, pero se recomienda realizar de forma previa una log-normalización por tamaño de librería (85). Para evitar normalizar demasiado los datos, lo que podría provocar la pérdida de variabilidad biológica, tampoco se realizó esta normalización.

La normalización de los datos se realizó con la función *processExpression()*, aplicando cada técnica de normalización descrita de manera independiente sobre la matriz de conteos cruda.

3.5. Selección de genes altamente variables

Dada la dimensión de la matriz de datos en ST, para un único estudio pueden evaluarse los niveles de expresión de varios miles de genes para decenas de miles de *spots*. Trabajar con tal magnitud de datos supone un elevado coste computacional y de almacenamiento, sumado además al hecho de que la mayoría de los genes no son biológicamente informativos para anotar los tipos celulares o identificar las regiones espaciales (análisis posteriores) (86).

A fin de reducir el ruido, se seleccionaron los genes cuyos niveles de expresión presentaban una mayor variabilidad biológica entre los *spots*. Estos genes se conocen como genes altamente variable (HVG, por sus siglas en inglés) y optimizan la relación señal-ruido al ser los más informativos de la variabilidad entre *spots* (87). *Giotto* ofrece tres métodos de detectar los HVG: grupos de covarianza, modelo basado en la predicción de regresión de LOESS (del inglés *Locally Weighted Scatterplot Smoothing*) y modelo basado en los residuos de Pearson.

En el método de grupos de covarianza, todos los genes se dividen en grupos del mismo tamaño en función de su expresión promedio y, dentro de cada grupo, se calcula para cada gen su coeficiente de variación (COV, por sus siglas en inglés). Los COV de cada grupo se convierten en un *z-score*, de modo que los genes con un *z-score* superior al umbral establecido, son considerados HVG. Se trata del método predeterminado de *Giotto* y se emplea cuando la variabilidad de la expresión génica difiere entre los distintos niveles de expresión o las regiones espaciales, sin asumir que hay una relación específica entre la expresión promedio y la varianza (69). La selección de HVG con este método se realizó con los valores por defecto: 20 grupos de covarianza y umbral de 1,5.

El segundo método alternativo identifica los HVG usando un modelo de regresión de LOESS, que predice el COV esperado para cada gen en base a su expresión log-normalizada. Este enfoque se utiliza en los casos en los que la relación media-varianza no es lineal o se puede describir mediante un modelo no paramétrico (69). Se consideraron HVG aquellos genes cuyo COV superó en 0,1 unidades al valor predicho por el modelo (valor por defecto).

El tercer y último método calcula los residuos de Pearson para estabilizar la varianza y explicar el ruido técnico, resaltando aquellos genes que muestran más dispersión de la esperada (80). En este caso, se seleccionaron como HVG los genes con una varianza superior a 1,5 (valor por defecto).

Para explorar las diferencias, se identificaron los HVG mediante estos tres métodos con la función *calculateHVF()*, utilizando como datos de entrada los valores de expresión previamente log-normalizados por tamaño de librería.

3.6. Reducción de la dimensionalidad

Una vez seleccionados los genes que representan la mayor parte de la variabilidad biológica, el siguiente paso es reducir la alta dimensionalidad intrínseca a los datos transcriptómicos espaciales. La reducción de la dimensionalidad, como su nombre indica, consiste en reducir el conjunto original de variables (en este caso, los niveles de expresión de los genes), combinando para ello la información más relevante de los HVG y descartando la información redundante o poco informativa. En ST, es común que la expresión de muchos genes esté correlacionada porque estos participan en los mismos procesos biológicos, o porque se expresan en localizaciones espaciales próximas. Reduciendo el número de características analizadas se consigue simplificar esta complejidad y reducir el tiempo de cómputo, pero sin alterar los patrones y tendencias biológicas. De este modo, se facilitan los análisis posteriores (como el agrupamiento de tipos celulares) y la visualización de los datos en un subespacio de menor dimensión (88,89).

Para esta tarea se realizó un Análisis de Componentes Principales (PCA, por sus siglas en inglés), una técnica de reducción de la dimensionalidad que convierte las variables originales, potencialmente correlacionadas, en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales. Cada componente principal se construye como una combinación lineal de los niveles de expresión génica, de tal manera que la máxima varianza posible se explique en el primer componente principal, la máxima varianza restante en el segundo, y así sucesivamente. Con esta aproximación se asume que la heterogeneidad debida a aspectos biológicos queda bien representada en los primeros componentes principales, mientras que los componentes restantes son una mezcla de ruido técnico y características poco relevantes (90).

Con la función *runPCA()* se calcularon los 50 primeros componentes principales para dos tipos de datos: valores de expresión log-normalizados por tamaño de librería y valores normalizados mediante residuos de Pearson. El PCA incluye un paso de centrado y escalado de los datos, siendo importante no realizar este proceso más de una vez. Por ello, para los valores de expresión log-normalizados por tamaño de librería fue necesario realizar también este procesamiento; mientras que los valores normalizados mediante residuos de Pearson estaban centrados y escalados, por lo que no requirieron este paso adicional. En ambos casos, para examinar la influencia de utilizar unos genes u otros en el tipo de información extraída, este análisis se realizó por triplicado, utilizando en cada caso los HVG previamente seleccionados con cada uno de los métodos descritos en el Apartado 3.5. La representación gráfica de los resultados se realizó con la función *plotPCA()*.

Posteriormente, con la función *screePlot()* se evaluó el número de componentes principales que contenían información relevante, con el objetivo de limitar los análisis posteriores únicamente a los componentes significativos, minimizando así el efecto de otros componentes menos significativos.

La suposición de linealidad subyacente al PCA a veces no representa la complejidad de los *spots*. Para superar esta limitación, se aplicó además un método de reducción de la dimensionalidad no lineal, conocido como UMAP (del inglés *Uniform Manifold Approximation and Projection*). En esta estrategia, se calcula la distancia que separa a los *spots* en el espacio de alta dimensionalidad, se conectan los *spots* cuya distancia sea menor a un umbral previamente establecido y, finalmente, las diferentes estructuras se conectan en función de la distancia relativa que las separa, proyectando los datos en un espacio de menor dimensión que mantiene la estructura global y local de los datos (91,92). Este método se realizó con la función *runUMAP(),* utilizando como datos de entrada los 20 componentes principales calculados previamente mediante PCA y 0,1 como umbral mínimo de distancia (valor por defecto). La representación gráfica de los resultados se realizó con la función *plotUMAP()*.

3.7. Agrupamiento

El agrupamiento se refiere a la clasificación de los *spots* en dominios con perfiles de expresión génica relativamente consistentes, de manera que queden categorizados en función de la similitud de sus patrones de expresión. Este tipo de agrupamiento solo considera la expresión génica y no tiene en cuenta las coordenadas espaciales, pero es esencial para que la posterior anotación de tipos celulares se realice sin sesgos (93).

Al trabajar con un conjunto de datos que incluye múltiples muestras procedentes del mismo tipo de tejido, los tipos celulares compartidos entre muestras deberían agruparse, independientemente de cuál es su muestra de origen. Sin embargo, debido a los efectos de lote derivados del procesamiento en diferentes días o a condiciones experimentales distintas, los *spots* pueden agruparse por muestra en lugar de por tipo celular. Esta variabilidad técnica podría ocultar la verdadera estructura biológica de los datos. Para solventar esta limitación, antes de realizar el agrupamiento, es necesario aplicar una técnica de integración entre muestras. El método implementado en *Giotto* para este fin es Harmony, que originalmente fue diseñado para integrar datos de scRNA-Seq (94).

La integración Harmony ajusta iterativamente las coordenadas de los *spots* en un espacio dimensional reducido, minimizando así el efecto de lote sin alterar la expresión génica. Este algoritmo comienza generando múltiples grupos de *spots*, asegurando que cada grupo tiene una mezcla de diferentes lotes[3] en una proporción basada en el número de *spots* de cada uno de ellos. Tras ello, asigna los *spots* a varios grupos y determina la probabilidad de que un *spot* particular pertenezca a una combinación de grupo-lote. Para cada una de estas combinaciones, se calculan unos factores de corrección que muestran cómo los *spots* de un lote deben desplazarse en el espacio dimensional para alinearse con los demás lotes del mismo grupo. Finalmente, se aplican a los *spots* estas correcciones basadas en su probabilidad de pertenecer a cada grupo. Este procedimiento se repite de forma iterativa hasta que el proceso converge y la distancia desplazada por los *spots* es inferior a un umbral previamente establecido (94) (Figura 7).

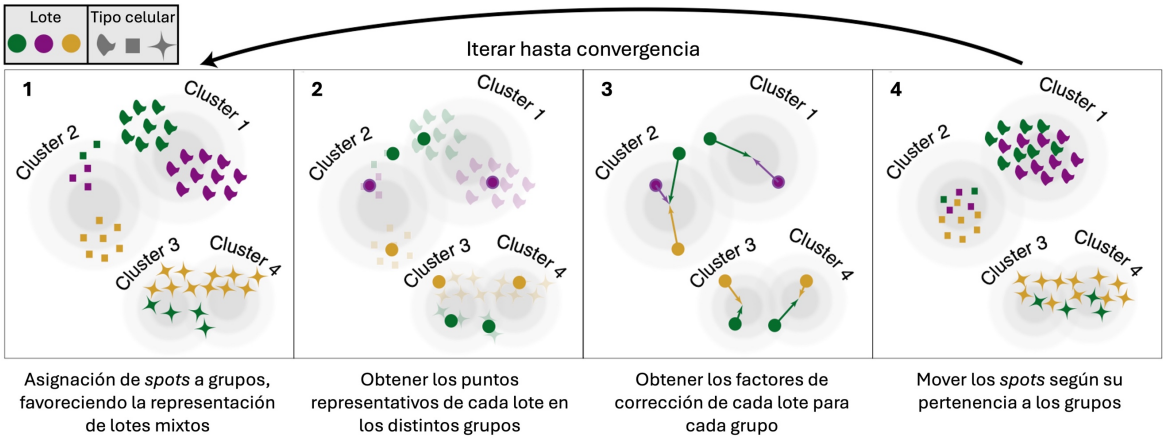

Figura 7. Esquema metodológico de la integración Harmony. La forma de los símbolos hace referencia a los distintos tipos celulares, y su color representa los diferentes lotes (por ejemplo, diferentes grupos experimentales). Figura adaptada de Korsunsky et al. (94).

La integración Harmony de los datos se llevó a cabo con la función *runGiottoHarmony()*, que a su vez recurre a la función *RunHarmony()* del paquete *harmony* (95). Se consideró el efecto de lote derivado del procesamiento en diferentes días (especificado por los autores en los metadatos de las muestras) y de los distintos grupos experimentales. Esta integración genera un espacio corregido a partir de un subespacio de menor dimensión, por lo que como datos de entrada en este caso se utilizaron los componentes principales calculados previamente (ver Apartado 3.6). Concretamente, se emplearon los 10 primeros componentes calculados para los valores log-normalizados por tamaño de librería a partir de los HVG seleccionados mediante el modelo de regresión de LOESS. Para comprobar la efectividad de esta integración, se recalculó el UMAP con la función *runUMAP()*, pero en este caso empleando el espacio de dimensiones generado con Harmony como datos de entrada.

Tras corregir los efectos de lote, y antes de ejecutar el algoritmo de agrupamiento, se requirió de la construcción de una red de vecinos más cercanos que facilitase la agrupación de características en el espacio de alta dimensión.

[3] Los lotes se refieren a los diferentes valores de la variable que se quiere corregir.

3.7.1. Red de vecinos más cercanos

Las redes de vecinos más cercanos modelan las relaciones de proximidad entre los *spots* sobre un espacio de dimensionalidad reducido, en función de sus similitudes en la expresión génica. Con *Giotto* se pueden generar redes de *k*-vecinos más cercanos (kNN, por sus siglas en inglés) o redes de vecinos más cercanos compartidos (sNN, por sus siglas en inglés).

El método kNN se fundamenta sobre el principio de que los datos similares tienden a agruparse cerca. Este algoritmo, basándose en una métrica de distancia (distancia euclidiana[4], por ejemplo), identifica los *k* puntos (en este caso, los *k spots*) más cercanos de cada uno de los puntos que forman el espacio de dimensiones. El resultado es un grafo dirigido, puesto que la relación de vecino más cercano no siempre es simétrica (96).

El algoritmo sNN establece la similitud entre dos puntos en función de los puntos vecinos que comparten. De igual modo que en las redes kNN, se comienza identificando los *k* puntos más cercanos para cada *spot*. Para cada par de puntos, calcula el número de vecinos compartidos y si es superior a un umbral establecido, conecta ese par de puntos por lo que se trata de un grafo no dirigido. Las redes sNN tienen más robustez ante variaciones de densidad local, lo que es crucial para grupos de diferente tamaño, además de que capturan agrupaciones más complejas (97).

Para evaluar la influencia del tipo de red en los resultados del agrupamiento posterior, se construyeron los dos tipos de redes para un número de vecinos más cercanos (*k*) igual a 15 con la función *createNearestNetwork()*, utilizando en ambos casos el espacio corregido de Harmony y los valores de expresión log-normalizados por tamaño de librería.

3.7.2. Agrupamiento de Leiden

Una vez creadas las redes de vecinos más cercanos, ya se pudo proceder al agrupamiento *per se*. *Giotto* implementa seis algoritmos de agrupamiento de *spots* basados en los perfiles de expresión de los genes: k-medias (98), agrupamiento jerárquico (99), paseo aleatorio (100), sNN (101), Louvain (102) y Leiden (103). De entre todos ellos, Louvain y Leiden, ambos basados en grafos, son los dos algoritmos más ampliamente empleados tanto en datos de célula/núcleo único como en datos de ST.

El algoritmo de Leiden consta de tres fases principales: 1) asignación de los nodos (en este caso, los *spots*) a una comunidad[5] y movimiento local de los nodos para tratar de maximizar la modularidad[6], 2) refinamiento de las divisiones para garantizar la correcta conexión interna y 3) creación de una red agregada basada en las divisiones refinadas para simplificar la estructura de la red generada. Estos tres pasos se repiten de forma iterativa hasta que no se observen más mejoras en la modularidad (103) (Figura 8).

[4] La distancia euclidiana representa el camino más corto entre dos puntos.
[5] Una comunidad es un grupo de nodos más conectados entre sí que con nodos de otros grupos.
[6] La modularidad cuantifica la calidad de división de una red en comunidades.

El método de Leiden fue propuesto como una mejora de Louvain, con capacidad de generar agrupaciones más precisas en tiempos de cómputo menores (103). Por ello, sumado a que fue el algoritmo aplicado por los autores de los datos empleados en este trabajo, el agrupamiento de tipos celulares se realizó con Leiden a partir de las redes sNN y kNN previamente generadas, utilizando la función *doLeidenCluster()*. La resolución del agrupamiento se puede regular, de manera que, a mayor resolución, más grupos de *spots* se definen. Para evitar la sobredetección de grupos y explorar el número de grupos más acorde a la distribución de los *spots*, en ambos casos se evaluaron varios valores de resolución: 0,15, 0,25 y 0,50.

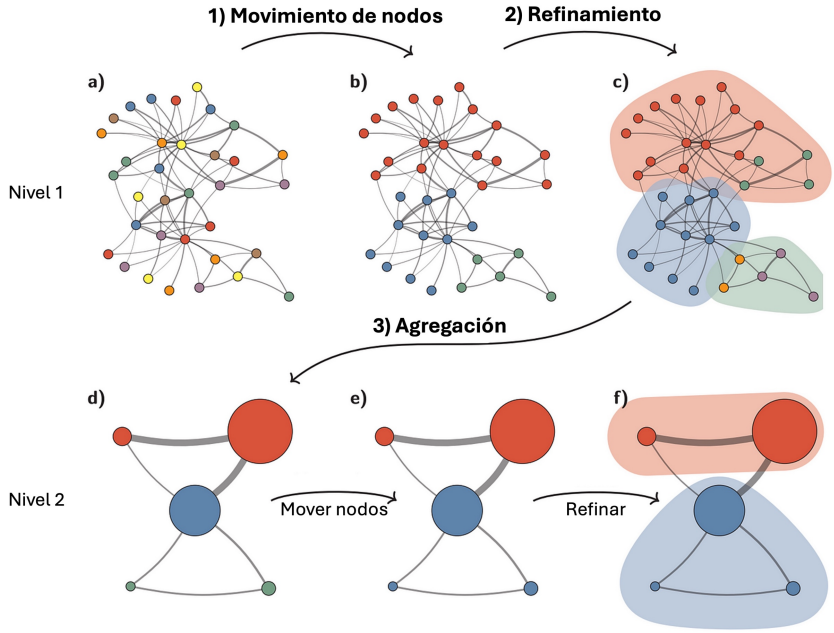

Figura 8. Esquema metodológico del algoritmo de Leiden. El algoritmo parte de los nodos conectados por una red de vecinos más cercanos (a), tras lo que mueve los nodos de un grupo a otro (b), refina las divisiones (c) y crea una red agregada (d). Todo este proceso se repite, de manera que el algoritmo traslada nodos individuales de la red agregada (e) y la refina de forma iterativa (f) hasta que no haya ninguna mejora. Figura adaptada de Traag et al. (103).

3.7.3. Genes marcadores

La detección de los genes más característicos de cada grupo de Leiden es un análisis que permite la posterior anotación de los tipos celulares que predominan en cada *spot*. Estos genes marcadores de grupo se identificaron mediante un análisis de expresión diferencial, para lo que *Giotto* ofrece tres algoritmos: MAST (104), método basado en los coeficientes de Gini (105), y Scran (106).

MAST implementa un modelo de obstáculos (del inglés *hurdle model*), el cual en una primera fase identifica si un gen se expresa (es decir, si "supera" el obstáculo de que su valor sea 0) y, de ser así, en una segunda fase evalúa si el nivel de expresión del gen supera un valor umbral definido previamente. De este modo, se pueden identificar todos aquellos genes diferencialmente expresados entre grupos (104).

El método de detección de genes marcadores mediante el coeficiente de Gini es un algoritmo diseñado por los propios desarolladores de *Giotto* (105). Este método permite identificar los genes que se expresan de manera muy selectiva en un grupo de Leiden, sin necesidad de que se expresen en todos los *spots* que componen dicho grupo.

Por último, el método implementado de Scran aplica la prueba *t* de Welch para cada gen entre pares de grupos de Leiden y, tras ello, combina y compara los resultados para determinar qué genes tienen una expresión significativamente superior en cada grupo respecto a los demás grupos (106). Es decir, los genes marcadores de un determinado grupo son los que presentan, en general, una mayor expresión diferencial en todas las comparaciones realizadas entre ese grupo y el resto de los grupos. Este algoritmo brinda un análisis más robusto de la expresión diferencial, especialmente si hay variabilidad técnica o diferencias en la profundidad de secuenciación entre *spots* (69).

Puesto que el método basado en los coeficientes de Gini era demasiado específico y que MAST tenía unos tiempos de ejecución muy largos, se decidió realizar el cálculo de los genes marcadores con el método de Scran, empleando para ello la función *findMarkers_one_vs_all()*. Este análisis de expresión diferencial se realizó entre los grupos de Leiden calculados a partir de la red kNN con una resolución de 0,50 y utilizando los valores log-normalizados por tamaño de librería. Para visualizar la correlación entre los grupos de Leiden identificados y los genes marcadores de cada uno de ellos, se generó un mapa de calor con la función *plotMetaDataHeatmap()*.

3.8. Anotación de tipos celulares

Los *spots* de *Visium* tienen un diámetro de 55 μM de diámetro, por lo que es probable que cada uno cubra múltiples células, las cuales a su vez pueden ser de diferentes tipos celulares. La falta de resolución espacial a nivel de célula única en los datos de ST de *Visium* puede provocar que los resultados de los genes marcadores de cada grupo de Leiden estén sesgados por el tipo celular dominante o por genes altamente expresados en un tipo celular concreto. Una solución para mejorar la resolución espacial y analizar la distribución espacial de las señales biológicas es inferir la composición celular de cada *spot* (deconvolución), integrando para ello la información externa de un atlas de snRNA-Seq emparejado como referencia. Concretamente, los datos de snRNA-Seq se emplearon para inferir las firmas genéticas de los distintos tipos celulares y, a partir de esta información, estimar la proporción de cada tipo de célula en las diferentes posiciones espaciales.

3.8.1. Conjunto de datos de snRNA-Seq de referencia

Los datos de snRNA-Seq se descargaron como un atlas curado en forma de fichero H5AD, formato que almacena objetos de clase *AnnData* generados a partir de datos de célula/núcleo único procesados en Python con paquetes como *Scanpy* (70). Al estar trabajando en R, con la función *readH5AD()* del paquete *zellkonverter* (107) se creó un objeto de clase *SingleCellExperiment* con el conjunto de datos de snRNA-Seq del estudio seleccionado tras la revisión sistemática. Para que hubiera correspondencia entre los datos de las dos tecnologías de transcriptómica y poder realizar la posterior anotación celular, el ID génico empleado en snRNA-Seq fue el GENE SYMBOL.

Los datos de núcleo único también son susceptibles a presentar artefactos técnicos relacionados con los procesos de disociación proteolítica de las células, encapsulación de los núcleos o secuenciación. Los datos de snRNA-Seq analizados en este trabajo emplearon un sistema de aislamiento de núcleos basado en encapsulación por microfluídica (*10x Genomics Chromium*). En este sistema, se generan gotas donde, idealmente, dentro de cada una se encuentra un único núcleo no dañado. Sin embargo, es habitual que se formen gotas sin ningún núcleo (gotas vacías), así como gotas que contienen dos o más núcleos (dobletes) y que, por tanto, muestran un perfil de expresión híbrido. En los estudios de snRNA-Seq también es frecuente la contaminación con ARNm ambiental, que es contabilizado junto con el ARNm nuclear (108) (Figura 9).

Para asegurar la fiabilidad de los resultados, los genes y núcleos de baja calidad deben ser descartados. Los criterios de filtrado se suelen adaptar a cada estudio, ya que la calidad de los datos depende en gran medida de cómo se ha realizado el diseño experimental. Para seleccionar los datos que proceden de núcleos individuales viables, se suelen filtrar los núcleos que no superan un umbral mínimo de genes detectados, lo que indica que se trata de una gota vacía o que ha habido algún error en la captura, amplificación y/o secuenciación; así como los que tienen un alto número de conteos, consecuencia de un posible doblete. Los genes expresados en pocas células no aportan información sobre la heterogeneidad celular, por lo que también suelen ser eliminados (51).

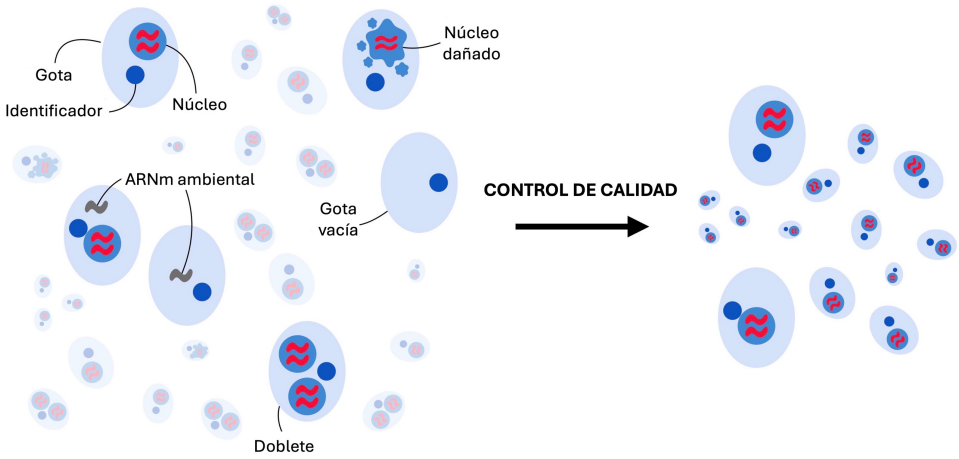

Figura 9. Esquema del control de calidad en datos de secuenciación de ARN de núcleo único (snRNA-Seq). En las tecnologías de secuenciación basadas en gotas (como *Chromium*), los datos de snRNA-Seq pueden contener núcleos de baja calidad, gotas vacías, dobletes y gotas que han incorporado ARNm ambiental. El control de calidad permite corregir todos estos factores para obtener un conjunto de datos que incluya únicamente las gotas con un único núcleo viable. Figura adaptada de Schaar et al. (109).

De igual modo que ocurre en ST, los datos de snRNA-Seq siempre son sometidos a un filtrado preliminar. *Cell Ranger* es un software desarrollado por *10x Genomics* para preprocesar los datos crudos de scRNA-Seq o snRNA-Seq generados con *10x Genomics Chromium*. Esta herramienta, antes de generar la matriz de expresión, elimina aquellas lecturas que no se pueden alinear con el genoma de referencia, filtra las gotas vacías y corrige la presencia de ARN ambiental. Los datos de snRNA-Seq empleados en este trabajo ya habían sido sujetos a este preprocesamiento con *Cell Ranger*. Además, puesto que se trataba de un atlas curado de snRNA-Seq, los datos también habían sido filtrados, analizados y anotados por los autores originales. No obstante, se comprobó la calidad de los datos, para lo que se emplearon varias métricas que permitieron identificar qué genes y núcleos no superaban los umbrales de calidad y que, por tanto, podían introducir sesgos en la anotación de tipos celulares de los datos espaciales.

Para el control de calidad a nivel de núcleo, se calculó el número de genes expresados y el tamaño de librería de cada núcleo con la función *addPerCellQC()* del paquete *scater* (110). La elevada presencia de transcritos de genes mitocondriales es un indicativo de mal aislamiento del núcleo y/o contaminación con ARN ambiental, por lo que también se calculó la proporción de ARN de genes mitocondriales. Por último, se evaluó el porcentaje de dobletes con la función *scDblFinder()* del paquete homónimo (111), que simula dobletes artificiales agrupando los núcleos en base a sus perfiles de expresión e identifica los dobletes reales comparando su similitud con los artificiales. Mediante la evaluación de todas las métricas de calidad, se valoró la necesidad de filtrar algún núcleo o gen. De manera paralela, se realizó el control de calidad a nivel de gen, donde se analizó el número de núcleos en los que se expresaba cada gen.

A continuación, los valores de expresión fueron log-normalizados por tamaño de librería con la función *normalizeGiotto()*, siendo necesaria la conversión previa del objeto *SingleCellExperiment* en un objeto *Giotto*. Para ello, se extrajeron tanto los metadatos de los núcleos como la matriz de expresión del objeto *SingleCellExperiment* y se pasaron como argumentos en la función *createGiottoObject()*. Cabe destacar que los datos de snRNA-Seq carecen de coordenadas espaciales, por lo que esta información se completó automáticamente con ubicaciones ficticias (*dummy values*).

La correcta identificación de los diferentes tipos celulares en el conjunto de datos de snRNA-Seq era clave para realizar la deconvolución de los datos de ST. Por ello, mediante un análisis de reducción de la dimensionalidad se comprobó que las poblaciones celulares identificadas en snRNA-Seq estaban bien agrupadas. Para ello, se identificaron los HVG entre los distintos núcleos con la función *calculateHVF()* a partir de la matriz de expresión log-normalizada por tamaño de librería. Puesto que los grupos de covarianza es el método predeterminado de esta función, la selección de los HVG en snRNA-Seq se realizó con este procedimiento. A partir de los HVG identificados, se calcularon los 50 primeros componentes principales con la función *runPCA()*. Para mejorar la visualización de los datos en un espacio de dimensiones reducido que fuera lo más similar posible al espacio de alta dimensionalidad, se realizó un UMAP con la función *runUMAP(),* utilizando como datos de entrada los 10 componentes principales (seleccionados a partir de la visualización de la varianza acumulada con la función *screePlot()*) y manteniendo el valor por defecto en el resto de parámetros. Con *plotUMAP()* se representaron gráficamente los resultados.

3.8.2. Deconvolución

Tomando como referencia la anotación celular de los datos de snRNA-Seq, se pueden seguir dos estrategias sobre los datos de ST: enriquecimiento de tipos celulares y deconvolución. El enriquecimiento de tipos celulares estima la probabilidad de presencia de cada tipo celular en los *spots*, evaluando la coincidencia entre las firmas genéticas de cada población celular (información extraída de los datos de snRNA-Seq) y los patrones de expresión de cada ubicación espacial. Sin embargo, este enfoque no ofrece una estimación cuantitativa de la proporción relativa de las diferentes células en cada *spot*, limitación abordada en los análisis de deconvolución. Puesto que el objetivo era caracterizar la distribución espacial y la abundancia relativa de los distintos tipos celulares en el tejido cerebral, se optó por seguir la estrategia de deconvolución de tipos celulares.

Para la deconvolución, *Giotto* incorpora el algoritmo SpatialDWLS (del inglés *Spatial Dampened Weighted Least Squares*) (112), cuya metodología se divide en dos fases. El algoritmo comienza con un análisis de enriquecimiento para identificar los tipos celulares que con mayor probabilidad están presentes en cada *spot*, tomando como referencia una matriz con las firmas genéticas de tipo celular de un conjunto de genes marcadores (m). Este enriquecimiento se realiza con el método PAGE (del inglés *Parametric Analysis of Gene Set Enrichment*) (113), el cual calcula la tasa de cambio[7] (FC, por sus siglas en inglés) de cada gen marcador como la relación entre su expresión en un *spot* y su expresión promedio en todos los *spots*, tras lo que computa la media de los FC del conjunto de genes marcadores en cada *spot* (S_m). Como control de fondo, calcula la media (μ) y la desviación estándar (δ) de los valores de FC considerando todos los genes en todos los *spots*. Por último, calcula una puntuación de enriquecimiento (ES) para cada tipo celular y *spot* (Ecuación 8) (112,113).

$$ES = \frac{(S_m - \mu) \cdot \sqrt{m}}{\delta} \qquad \text{Ecuación 8}$$

Solo los tipos celulares que presentan un ES mayor o igual a 2 en un *spot* dado se consideran para la segunda fase, en la que se infiere la composición celular de cada *spot* aplicando el algoritmo DWLS (114), desarrollado originalmente para la deconvolución de datos de *bulk* RNA-Seq. Este algoritmo corrige los sesgos hacia los tipos celulares que presentan genes altamente expresados o que son altamente prevalentes (112).

Para realizar la deconvolución con el algoritmo SpatialDWLS, era necesario transformar la matriz de expresión de snRNA-Seq a un formato compatible con la deconvolución. Por tanto, con la función *makeSignMatrixDWLSfromMatrix()* se generó una matriz con las firmas genéticas de los distintos tipos celulares a partir de la matriz de expresión normalizada de los datos de snRNA-Seq, un vector con los 50 principales genes marcadores de cada grupo de Leiden previamente identificados (ver Apartado 3.7.3) y un vector con todos los tipos celulares anotados en los datos de snRNA-Seq. Esta nueva matriz, que contenía la expresión promedio de todos los genes marcadores en cada tipo celular, fue empleada como dato de entrada para la deconvolución de los datos de ST mediante la función *runDWLSDeconv()*, donde además se especificó que se considerase un promedio de 5 células por *spot*, ya que cada *spot* de *Visium* suele tener entre 1 y 10 células. Con la función *spatDeconvPlot* se visualizaron espacialmente las proporciones de cada tipo celular en los *spots*. Esta función, en lugar de mostrar los *spots* como puntos, genera un gráfico de sectores para cada *spot* con su proporción de tipos celulares.

3.9. Patrones espaciales de expresión

Lo que distingue a la ST de otras técnicas transcriptómicas es que también incorpora información sobre la organización espacial de los perfiles transcripcionales dentro de un tejido, lo que permite identificar genes que siguen y comparten distribuciones espaciales en su expresión. Estos genes reciben el nombre de genes espacialmente variables (SVG, por sus siglas en inglés) y, analizando cómo se correlacionan sus patrones de expresión con la arquitectura tisular específica de cada una de las muestras, facilitan la comprensión del microambiente espacial del tejido en la EM.

[7] La tasa de cambio es una medida que describe la magnitud de cambio en la expresión de los genes.

3.9.1. Red espacial de Delaunay

Una manera de identificar las relaciones espaciales entre genes y *spots* (patrones espaciales de expresión) es construyendo una red espacial basada en la proximidad espacial entre los *spots*. Por tanto, en este tipo de redes los nodos son los *spots* y las aristas son las conexiones entre pares de *spots* vecinos. Existen múltiples tipos de redes espaciales y métodos para generarlas, siendo la red de Delaunay la estrategia implementada en *Giotto* de forma predeterminada.

La red de Delaunay se basa en la triangulación de Delaunay y se construye conectando los *spots* en un plano, de tal manera que ningún *spot* quede dentro de la circunferencia circunscrita en un triángulo formado por otros tres *spots* (115). De este modo, el ángulo mínimo de todos los triángulos queda maximizado, evitando triángulos con ángulos muy pequeños que generan lados muy desequilibrados al ser extremadamente estrechos. Además, se puede establecer una distancia física máxima a partir de la cual ya no se considera la vecindad entre *spots*. La ventaja de estas redes es que proporcionan una búsqueda eficiente del vecino más cercano (69).

Giotto ofrece a su vez varias implementaciones para generar las redes de Delaunay: *deldir* (para dos dimensiones), *delaunayn_geometry* (para múltiples dimensiones) y *RTriangle* (para geometrías complejas o con bordes irregulares). Puesto que los datos estaban en dos dimensiones y las secciones de tejido no presentaban ninguna geometría compleja o irregular, la red de Delaunay de cada muestra se construyó con el método *deldir* utilizando la función *createSpatialDelaunayNetwork()*. Como límite de distancia espacial entre los vecinos de Delaunay se utilizó el valor del bigote superior del vector de distancia entre vecinos (valor por defecto). Por último, dado que los *spots* de *Visium* tienen una geometría hexagonal, y teniendo en cuenta que los *spots* del borde de la muestra de tejido o los que están próximos a una rotura carecen de vecinos en todos sus lados, se especificó que al menos se considerara un mínimo de tres vecinos más cercanos para cada *spot*.

3.9.2. Genes espacialmente variables

Los genes con patrones espaciales de expresión se detectaron integrando los valores de expresión génica con las coordenadas espaciales de los *spots*, lo que permitió encontrar los genes más informativos de la estructura biológica dentro de la sección de tejido. Para la identificación de los SVG, *Giotto* ofrece cuatro estrategias diferentes, entre las que se encuentran tres métodos previamente publicados: SpatialDE (116), Trendsceek (117) y SPARK (del inglés *Spatial Pattern Recognition via Kernels*) (118).

SpatialDE es un método estadístico basado en la regresión del proceso gaussiano, que permite identificar genes cuya expresión génica depende de forma significativa de las coordenadas espaciales de forma no lineal y no paramétrica (116). Trendsceek es un método no paramétrico basado en procesos de puntos marcados que clasifica y evalúa la importancia de las tendencias de expresión espacial de cada gen individual, identificando gradientes espaciales de expresión significativos (117). SPARK evalúa los genes y modela su nivel de expresión a través de los distintos *spots* utilizando un modelo espacial lineal generalizado (118).

El cuarto método implementado en *Giotto* es una metodología novedosa desarrollada por los propios autores del paquete, denominada BinSpect (del inglés *Binary Spatial Extraction*) (69). BinSpect, a diferencia de los otros tres métodos mencionados, requiere de una red espacial generada previamente para clasificar los genes en función de si presentan una localización de patrón espacial o no. Se trata de un método computacional significativamente más rápido que los otros métodos y capaz de detectar los SVG de manera robusta, aunque se introduzca ruido adicional (69). Por ello, fue el método de elección para la identificación de los SVG de los datos analizados.

Para cada uno de los genes incluidos en los datos espaciales, el algoritmo de BinSpect realiza tres pasos de forma independiente que permiten identificar si se trata de un SVG (Figura 10). En primer lugar, asigna una etiqueta binaria a cada *spot* en función de si el gen evaluado presenta niveles de expresión bajos o altos. Esta binarización se puede realizar mediante dos estrategias: *k-means* y *rank*. El método de *k-means* clasifica los *spots* en dos grupos según el nivel de expresión del gen, de manera que los *spots* del grupo con niveles de expresión altos del gen reciben un 1, y los *spots* del otro grupo un 0. Alternativamente, el método de *rank* ordena los *spots* en orden decreciente según el nivel de expresión que presentan para el gen, asignando un 1 a los *spots* del percentil superior (30 % por defecto) y un 0 al resto. Tras ello, se crea una tabla de contingencia basada en el número de conexiones de los *spots* vecinos de una red espacial y la expresión binarizada de los genes, a partir de la cual realiza una prueba exacta de Fisher para analizar si los *spots* con la misma etiqueta binaria tienden a estar conectados. Un gen es un SVG si, en general, está altamente expresado en *spots* cercanos (69).

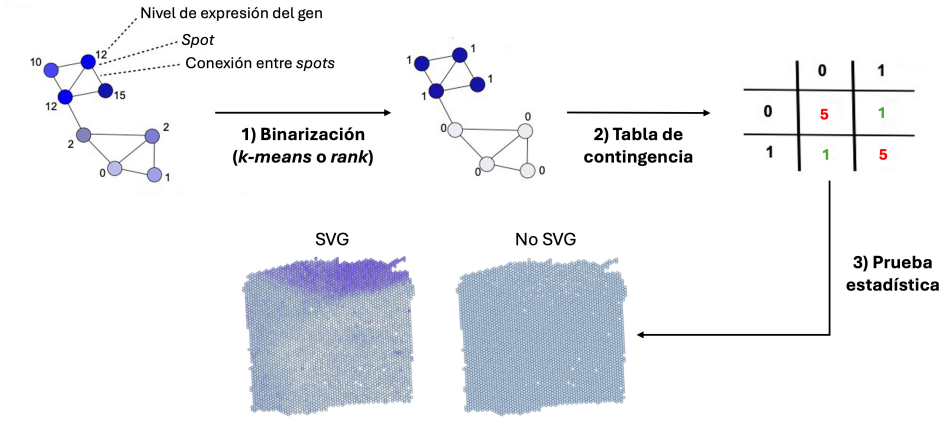

Figura 10. Esquema metodológico del algoritmo BinSpect para la detección de genes espacialmente variables (SVG). 1) Binarización de los *spots* según el nivel de expresión del gen analizado, 2) creación de una tabla de contingencia con las etiquetas (0,0), (0,1), (1,0) o (1,1) en base al número de conexiones entre los *spots* y 3) prueba estadística para determinar si se trata de un SVG, es decir, si está altamente expresado en una región espacial. Un gen no se considera espacialmente variable (no SVG) si presenta una abundancia espacialmente uniforme. Figura modificada de Dries et al. (69).

La identificación de los SVG se realizó con la función *binSpect()*, empleando tanto la binarización con el método de *k-means* como el de *rank* para explorar las diferencias. Como datos de entrada se utilizaron los valores de expresión log-normalizados por tamaño de librería y la red de Delaunay generada previamente. Los p-valores obtenidos tras la prueba de Fisher se ajustaron automáticamente con Benjamini-Hochberg, método que controla la tasa de falso descubrimiento (FDR, por sus siglas en inglés)[8] (119). Para el resto de los parámetros, se mantuvieron los valores por defecto.

[8] El FDR es una medida estadística que evalúa la proporción de resultados significativos que son falsos positivos entre todos los resultados positivos.

3.9.3. Dominios espaciales

A partir del conjunto de SVG identificados, se procedió a definir los dominios espaciales, es decir, las regiones del tejido formadas por *spots* próximos que presentaban patrones de expresión génica similares. Mientras los métodos de agrupación no espacial (como el algoritmo de Leiden) agrupan los *spots* en el espacio de expresión, y después mapean los resultados en sus localizaciones en el tejido; los métodos de agrupación espacial combinan los perfiles de expresión génica de los *spots* con su ubicación espacial, ofreciendo un resultado más continuo y homogéneo que permite caracterizar la organización histológica dentro del tejido.

El método para la detección de dominios espaciales implementado en *Giotto* utiliza un campo aleatorio oculto de Markov (HMRF, por sus siglas en inglés) (120). HMRF se trata de un modelo basado en grafos que infiere el dominio espacial al que pertenece cada *spot* comparando su firma genética con la de sus *spots* más cercanos, tratando de buscar patrones de expresión comunes. Esta inferencia se basa en la probabilidad conjunta del estado intrínseco del *spot*, determinado por sus patrones de expresión génica, y de su estado extrínseco, basado en la distribución de los dominios espaciales de sus *spots* vecinos. Para ello, el modelo necesita una red espacial que defina la vecindad entre los *spots* y un parámetro que regule la fuerza de interacción entre *spots* (β). Valores altos de β implican un mayor peso en el estado extrínseco, de manera que la asignación de los *spots* a un dominio espacial u otro está fuertemente influenciada por el estado de dominio de los *spots* vecinos, favoreciendo así la formación de dominios espacialmente más homogéneos, incluso aunque exista variabilidad en la expresión dentro de dicho dominio.

La identificación de los dominios espaciales con HMRF se divide en dos fases: 1) inicialización de la agrupación de dominios y 2) inferencia del modelo. La primera fase se realizó con la función *initHMRF_V2()*, utilizando como datos de entrada los valores de expresión log-normalizados por tamaño de librería, la red de Delaunay previamente generada y los SVG identificados con el método *k-means* de BinSpect. Esta inicialización consistió en la reducción de la dimensión de los SVG (se seleccionaron los 500 SVG más significativos, es decir, con menor p-valor ajustado) y en la agrupación de los *spots* en nueve dominios según sus perfiles de expresión utilizando el algoritmo de k-medias. Para refinar la asignación de los *spots* a los distintos dominios espaciales, considerando también la información espacial (segunda fase), se utilizó la función *doHMRF_V2()*, especificando un rango de valores de β de 0 a 45 con incrementos de 5. Los resultados se incorporaron al objeto *Giotto* con la función *addHMRF_V2()*.

Finalmente, los dominios espaciales identificados para β igual a 15 fueron anotados manualmente en las diferentes áreas tisulares características de las lesiones de EM en base a la presencia de los distintos tipos celulares (resultados de la deconvolución). De este modo, los dominios espaciales que correspondían con regiones con una elevada proporción de oligodendrocitos fueron anotados como sustancia blanca, distinguiendo entre sustancia blanca sana en las muestras de control y sustancia blanca perilesional (adyacente a la lesión) en las muestras de EM. Los dominios con baja densidad de oligodendrocitos y alta de astrocitos se asignaron al núcleo de la lesión de la EM. Los dominios que presentaban una alta concentración de microglía se establecieron como el borde de la lesión. Por último, los dominios correspondientes a las regiones espaciales en las que predominaban las neuronas fueron categorizados como sustancia gris.

3.10. Análisis de expresión diferencial

Las diferencias asociadas al sexo en la EM se caracterizaron mediante un enfoque que consideró la resolución espacial de los datos de ST, lo que posibilitó la realización de comparaciones entre las distintas regiones tisulares previamente anotadas.

Con el objetivo de evaluar las diferencias en la expresión génica relacionadas con el sexo dentro del foco patológico de la EM, se comparó la expresión génica en el núcleo desmielinizado de las lesiones de EM mediante un análisis de expresión diferencial entre hombres y mujeres. Para poder atribuir las diferencias observadas a la enfermedad y no a variaciones biológicas inherentes al sexo, se comparó además el nivel de expresión en la sustancia blanca sana entre hombres y mujeres. Este segundo contraste que utilizó tejido sano sirvió como referencia para distinguir los cambios basales entre sexos no asociados a la EM.

Adicionalmente, se exploraron los cambios transcripcionales debidos a la progresión de la EM mediante comparaciones intrasexo entre las zonas afectadas y no afectadas por la enfermedad. Para ello, se comparó el nivel de expresión génica entre la sustancia blanca perilesional y el núcleo de la lesión (Figura 11), realizando el análisis en hombres y mujeres por separado. De este modo, se pudo evaluar si la organización de la lesión de EM variaba en función del sexo.

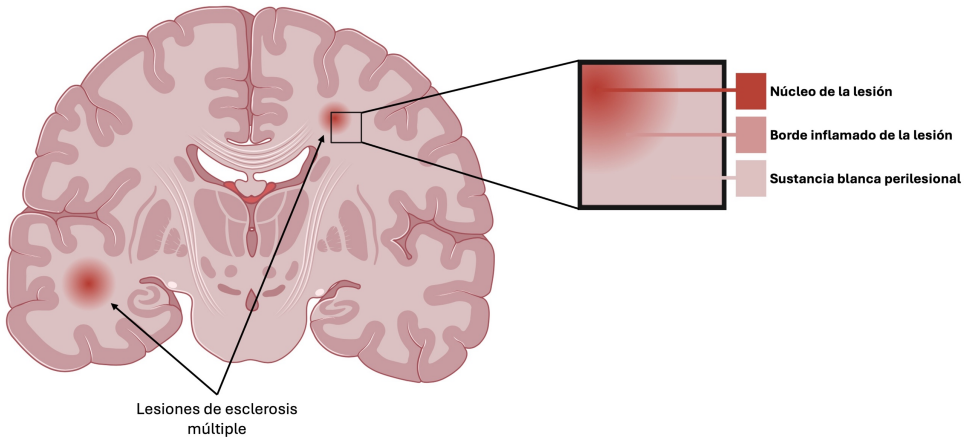

Figura 11. Representación esquemática de las lesiones de esclerosis múltiple en un corte coronal de cerebro. Se muestra la localización de dos lesiones de esclerosis múltiple y un acercamiento que ilustra su organización tisular: núcleo de la lesión, borde inflamado de la lesión y sustancia blanca perilesional. Figura generada con BioRender.

Todas las comparaciones de expresión diferencial (Tabla 3) se llevaron a cabo con la función *findScranMarkers()*, la cual a su vez emplea internamente la función *findMarkers()* del paquete *scran* (106). Esta función aplica la prueba *t* de Welch para cada gen entre los dos grupos definidos en cada contraste y ajusta automáticamente los p-valores obtenidos mediante el procedimiento de Benjamini-Hochberg (119). Como niveles de expresión se utilizaron los valores log-normalizados por tamaño de librería. Se consideraron diferencialmente expresados los genes con un FDR inferior a 0,05 y un valor absoluto del logaritmo del FC (|logFC|) superior a 0,5. El signo del logFC se interpretó en relación con el grupo de referencia utilizado en cada caso, de manera que un valor positivo para un gen indicó sobreexpresión y un valor negativo señaló subexpresión en el grupo comparado.

Tabla 3. Descripción de las comparaciones realizadas para caracterizar las diferencias de sexo en esclerosis múltiple mediante un análisis de expresión diferencial. Cada contraste compara los niveles de expresión asociados al sexo y a la localización tisular. Los contrastes 1 y 2 evalúan las diferencias entre hombres y mujeres en el núcleo de la lesión de EM y en la sustancia blanca sana, respectivamente. Los contrastes 3 y 4 analizan las diferencias entre la sustancia blanca perilesional (adyacente a la lesión) y el núcleo de la lesión en muestras de mujeres y de hombres con EM, respectivamente. EM: esclerosis múltiple. CTRL: control.

Contraste	Subconjunto	Grupo 1 (referencia)	Grupo 2
1	Núcleo de la lesión de esclerosis múltiple	Mujeres EM	Hombres EM
2	Sustancia blanca sana	Mujeres CTRL	Hombres CTRL
3	Mujeres EM	Sustancia blanca perilesional	Núcleo de la lesión de esclerosis múltiple
4	Hombres EM	Sustancia blanca perilesional	Núcleo de la lesión de esclerosis múltiple

4. Resultados

4.1. Revisión sistemática

En la revisión sistemática de las distintas bases de datos y repositorios, aplicando los criterios de inclusión especificados en la Tabla 1 del Apartado 3.2 se identificaron un total de 17 estudios únicos de ST en EM[9]. Tras revisar cada estudio de forma individual, y atendiendo a los criterios de exclusión establecidos (Tabla 2), 5 de los 17 estudios se descartaron por carecer de información sobre el sexo de los participantes. De los 12 estudios restantes, 2 de ellos fueron rechazados por no tener disponibles los datos en un repositorio público. Además, se identificaron 8 estudios que, aunque tenían información sobre el sexo de los individuos y los datos públicos, carecían de al menos 3 muestras por sexo y condición, por lo que también fueron excluidos.

En este punto, tan solo 2 estudios superaron los criterios de exclusión anteriores. Ambos disponían de un conjunto de datos de snRNA-Seq generado a partir de los mismos individuos (necesario para el análisis de deconvolución), por lo que este requisito no permitió discriminar uno de ellos. Sin embargo, para la generación de los datos de ST, uno de ellos empleó la tecnología *smFISH* (técnica basada en imágenes) y otro la tecnología de *Visium* (técnica basada en secuenciación). Puesto que *Visium* es la plataforma más extendida, se eligió el estudio que utilizó esta última tecnología.

4.1.1. Descripción del estudio seleccionado

Los datos seleccionados provienen del trabajo de investigación de Lerma-Martin et al. (121). Este estudio tenía como objetivo generar un atlas de la patología de las lesiones subcorticales de la EM integrando los datos de dos tecnologías ómicas: snRNA-Seq y ST. Para el análisis de ST con *Visium*, emplearon un total de 19 muestras humanas de tejido cerebral *post mortem* congeladas: 12 de lesiones de EM y 7 de sustancia blanca subcortical de control (CTRL). Cada muestra CTRL procedía de un donante sano diferente sin alteraciones neuropatológicas reconocidas, mientras que algunas muestras de EM provenían del mismo donante (réplica técnica). Todos los donantes de EM (7 en total) habían sido diagnosticados con el subtipo de EMSP.

Para caracterizar el tipo de lesión de las muestras de EM, los autores realizaron una evaluación histopatológica mediante inmunohistoquímica y varias tinciones de las secciones de tejido, identificando así las células mieloides activas, el núcleo de la lesión desmielinizado y el borde de la lesión. De este modo, 8 muestras fueron clasificadas como lesiones crónicas activas (EM-CA), caracterizadas por la presencia de un núcleo completamente desmielinizado, un borde inflamado bien definido y células mieloides activadas. Las 4 muestras restantes correspondían a lesiones crónicas inactivas (EM-CI), ya que presentaban una cicatriz glial en el núcleo desmielinizado y bajo nivel de inflamación en el borde de la lesión.

[9] Los resultados de la revisión sistemática fueron de febrero de 2025. Una revisión más actualizada con los mismos criterios podría incluir un mayor número de estudios.

De manera complementaria, 16 de estas muestras fueron utilizadas para generar un atlas de snRNA-Seq emparejado con los datos de ST. Este conjunto de datos se obtuvo a partir de los núcleos aislados y procesados mediante el sistema de microfluídica *Chromium* de *10x Genomics*. El *dataset* resultante incluyó un total de 6 muestras CTRL, 6 de EM-CA y 4 muestras de EM-CI.

Todos los datos brutos generados por los autores del estudio original están disponibles en el repositorio GEO bajo el identificador GSE279183. Este identificador corresponde a una "SuperSerie" (conjunto de experimentos relacionados con una misma publicación), que a su vez está compuesta por dos conjuntos de datos: snRNA-Seq (GSE279180) y ST (GSE279181).

Aunque en la publicación original los autores señalan que generaron datos de ST a partir de 7 muestras CTRL (4 de mujeres y 3 de hombres), al descargar la información de GEO, solo fue posible acceder a los datos espaciales de 6 muestras CTRL (4 de mujeres y 2 de hombres). A pesar de esta limitación, y dado que ningún otro estudio cumplía con todos los criterios de elegibilidad establecidos, se optó por continuar el análisis utilizando los datos disponibles del estudio de Lerma-Martin et al. (121). Las características más relevantes de las muestras de este estudio se resumen en la Figura 12.

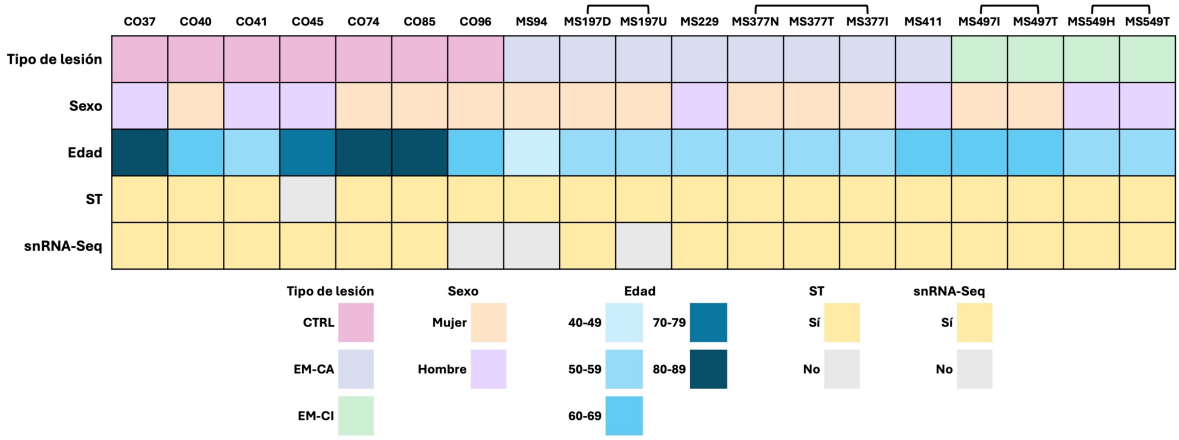

Figura 12. Descripción del diseño experimental y de las muestras incluidas en el estudio de Lerma-Martin et al. (121). Se representa la distribución de los tipos de lesiones, el sexo y rango de edad de los donantes, así como la disponibilidad de datos de transcriptómica espacial (ST) y de secuenciación de ARN de núcleo único (snRNA-Seq) para cada una de las 19 muestras analizadas. Los colores muestran la clasificación de cada variable según la leyenda de la parte inferior y las conexiones entre las muestras indican que provienen del mismo donante. CTRL: Control. EM-CA: Lesiones crónicas activas de esclerosis múltiple. EM-CI: Lesiones crónicas inactivas de esclerosis múltiple.

Los datos de ST se descargaron del material suplementario de GEO, consistiendo en los ficheros generados por el *software* de análisis *Space Ranger* para cada muestra. Tras organizar los archivos en la estructura de directorios adecuada, se creó un objeto de clase *Giotto* por muestra. Tras filtrar los *spots* vacíos, se mantuvieron, en promedio, aproximadamente 4.000 *spots* en cada muestra. El número de posiciones espaciales sin tejido depende en primera instancia de la morfología y el tamaño de la sección de tejido. Por ejemplo, la muestra MS497T fue aquella en la que se descartaron menos *spots*, en contraste con MS549T, que presentaba únicamente 2.811 localizaciones cubiertas por tejido de las casi 5.000 posibles (Figura 13).

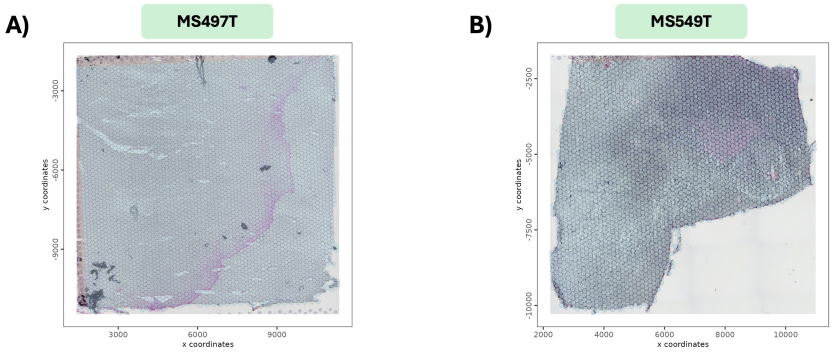

Figura 13. Comparación visual del grado de cobertura tisular entre áreas de captura de *Visium*. A) Muestra MS497T con alta cobertura tisular de *spots*. B) Muestra MS549T con cobertura tisular parcial. Imágenes derivadas de la superposición del área de captura con la imagen histológica de H&E. Ambas muestras son de lesiones crónicas inactivas de esclerosis múltiple. H&E: Hematoxilina y eosina.

4.2. Control de calidad

El control de calidad en ST tiene como principal objetivo identificar aquellos genes poco informativos y *spots* de baja calidad que puedan comprometer el análisis posterior. En este contexto, se calcularon una serie de métricas o indicadores de calidad, cuya distribución y variabilidad entre muestras se representa en la Figura 14.

El análisis de la expresión génica a lo largo de los *spots* reveló que algunas muestras presentaban *spots* con un menor número de genes detectados (Figura 14A) y con un tamaño de librería reducido en comparación con el resto de muestras (Figura 14B). Además, se observó una correlación entre ambas métricas, ya que los *spots* de las muestras con pocos genes detectados también presentaban un tamaño de librería inferior. Destacan las muestras MS197D, MS229, MS377I y MS94, todas del grupo EM-CA, en las que la mayoría de los *spots* de baja calidad se localizan en regiones espaciales específicas (Figura 15A).

Filtrando todos los *spots* con menos de 200 genes, criterio que aplicaron los autores del artículo original, la muestra MS229 conservó únicamente poco más de la mitad del total de sus posiciones iniciales, siendo la muestra con mayor descarte de *spots*, seguida por las muestras MS377I y MS94 (Figura 15B). Para minimizar la pérdida de buena parte de las localizaciones espaciales en algunas muestras, especialmente en las del grupo EM-CA, lo que podría enmascarar señales biológicas, se valoraron otros umbrales menos restrictivos. Finalmente, como criterio de filtrado se estableció la eliminación de los *spots* con menos de 100 genes detectados, lo que disminuyó significativamente la pérdida de posiciones espaciales en la muestra MS229 (Figura 15C). El resto de las muestras conservaron prácticamente todos sus *spots* (Figura A1 del Anexo).

En cuanto al porcentaje de lecturas mitocondriales, aunque generalmente se asocia a daño celular, en este estudio los mayores porcentajes se observaron en las muestras de EM (Figura 14C) y distribuidas en regiones espaciales concretas (Figura 15D). Con objeto de evaluar la posible implicación de este tipo de genes en la fisiopatología de la EM, esta métrica no se incluyó como un criterio para el filtrado, conservándose todos los *spots* con más de 100 genes, independientemente de su contenido mitocondrial. Finalmente, puesto que la muestra MS94 presentó una proporción de genes mitocondriales extremadamente elevada en la mayoría de sus *spots*, aunque no se eliminó del estudio, este hecho se consideró para la interpretación de los análisis posteriores.

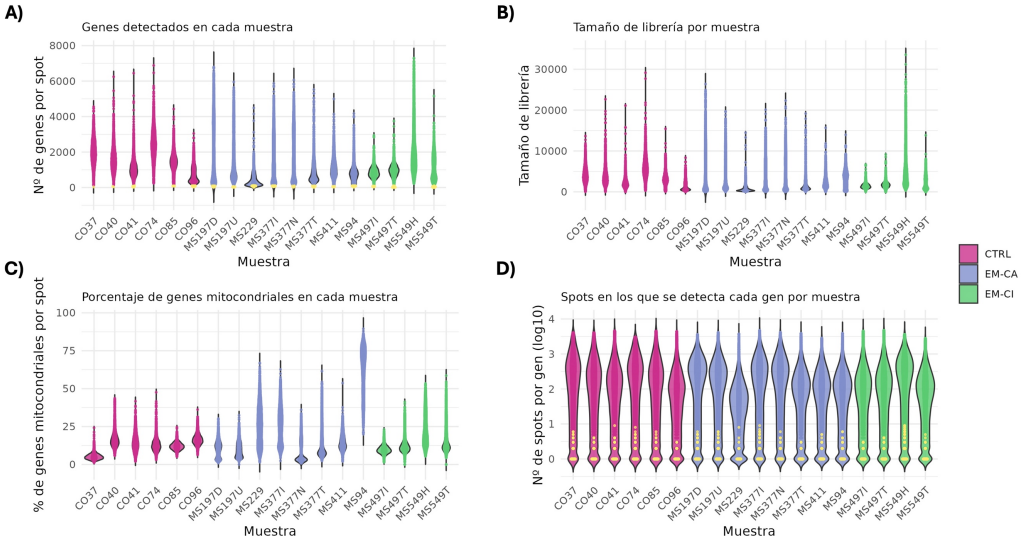

Figura 14. Distribución de los indicadores de calidad calculados para los datos de transcriptómica espacial. A) Número de genes detectados por *spot*. Los puntos amarillos representan los *spots* con menos de 100 genes. B) Tamaño de librería (total de conteos por *spot*). C) Porcentaje de genes mitocondriales por *spot*. D) Número de *spots* en los que se expresa cada gen (en escala logarítmica). Los puntos amarillos representan los genes no expresados en al menos 10 *spots*. Los gráficos de violín muestran la distribución de los parámetros en cada una de las muestras, diferenciando entre los tres grupos experimentales: Control (CTRL, rosa), lesiones crónicas activas de esclerosis múltiple (EM-CA, azul) y lesiones crónicas inactivas de esclerosis múltiple (EM-CI, verde).

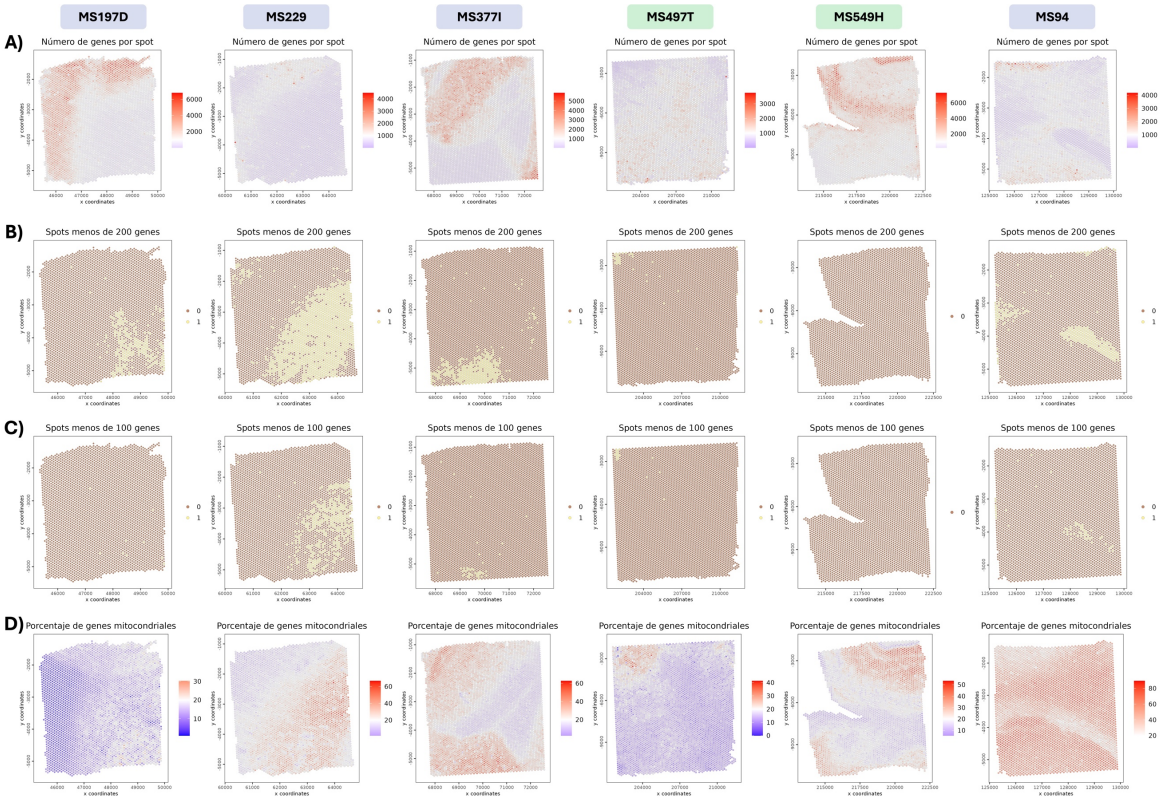

Figura 15. Evaluación espacial de la calidad de los datos de transcriptómica espacial en seis muestras ejemplo. Las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos azules (MS197D, MS229, MS377I y MS94), y las de lesiones crónicas inactivas de esclerosis múltiple con marcos verdes (MS497T y MS549H). A) Distribución espacial del número de genes detectados por *spot* (escala de color centrada en 1.000 genes). B) Localización de los *spots* con menos de 200 genes detectados (en amarillo). C) Localización de los *spots* con menos de 100 genes detectados (en amarillo). D) Porcentaje de genes mitocondriales por *spot* (escala de color centrada en el 20 %).

En el conjunto de muestras se detectaron aproximadamente 28.000 genes únicos, estando muchos de ellos únicamente presentes en una fracción ínfima de *spots* (Figura 14D), lo que supuso que gran parte de la matriz de expresión fueran valores de 0 (matriz dispersa). Estos genes, además de aumentar considerablemente la dimensionalidad de los datos, no son representativos de la actividad biológica bajo estudio. Por ello, todos aquellos genes sin ningún conteo en al menos 10 *spots* de entre todos las muestras fueron filtrados, lo que supuso la eliminación de un total de 5.666 genes. Tras el filtrado completo, las dimensiones del conjunto de datos resultaron en 22.671 genes y 70.237 *spots* provenientes de todas las muestras.

4.3. Normalización

La comparación de los métodos implementados en *Giotto* para normalizar los datos transcriptómicos espaciales incluyó la log-normalización por tamaño de librería, el método basado en los residuos de Pearson y la normalización TF-IDF seguida de la normalización L2 (Figura 16).

Respecto a los datos sin normalizar (Figura 14B), la log-normalización por tamaño de librería, método predeterminado en *Giotto*, supuso una distribución más homogénea del tamaño de librería entre muestras (Figura 16A). Los residuos de Pearson revelaron que, en la mayoría de las muestras, los genes se expresaban conforme a lo esperado (valores cercanos a 0), a excepción de la muestra MS94, donde el total de conteos fue menor de lo esperado (valores negativos) (Figura 16B), posiblemente debido a la alta proporción de genes mitocondriales detectada en esta muestra. Por último, la combinación de las normalizaciones TF-IDF y L2 fue la que más homogeneizó el tamaño de librería entre las muestras, lo que podría ocasionar una corrección excesiva de la variabilidad biológica entre muestras (Figura 16C).

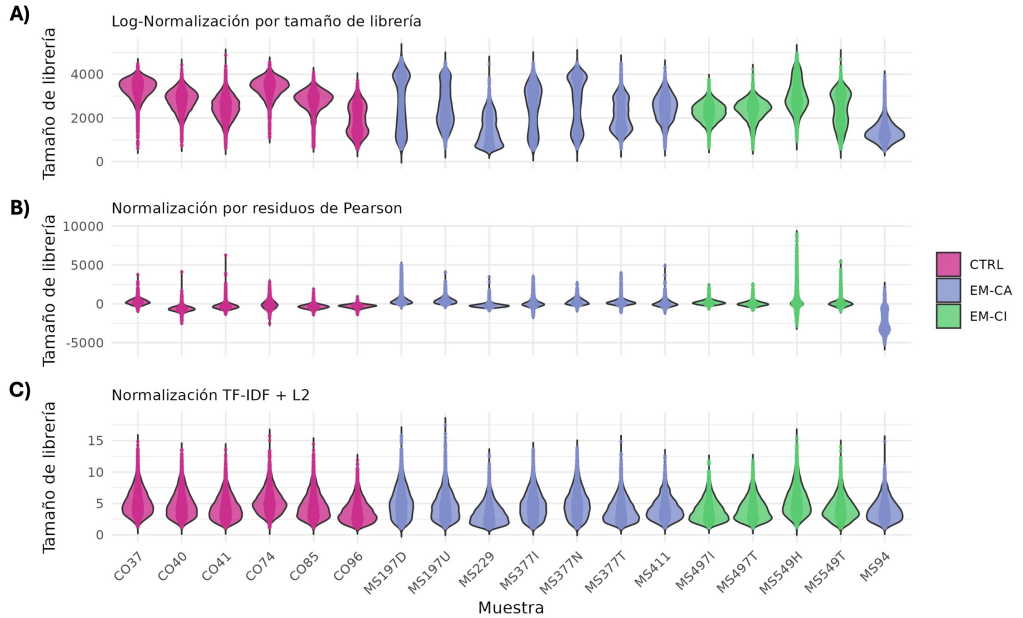

Figura 16. Distribución del tamaño de librería tras aplicar diferentes metodologías de normalización para datos de transcriptómica espacial. A) Log-normalización por tamaño de librería, B) normalización por residuos de Pearson y C) normalización TF-IDF seguida de L2. Cada gráfico de violín representa el tamaño de librería (número total de conteos) de las diferentes muestras tras filtrar y normalizar, diferenciando entre muestras control (CTRL, rosa), lesiones crónicas activas de esclerosis múltiple (EM-CA, azul) y lesiones crónicas inactivas de esclerosis múltiple (EM-CI, verde).

4.4. Selección de genes altamente variables

El número de genes identificados como altamente variables (HVG) por el método de grupos de covarianza en los datos de ST filtrados y log-normalizados por tamaño de librería fue de 1.441, lo que supone aproximadamente un 6,4 % del total de genes originales (Figura 17A). Con el modelo de regresión de LOESS se seleccionaron un total de 11.274 HVG, lo cual representa aproximadamente un 49,7 % del total de genes (Figura 17B). Por último, con el método de residuos de Pearson se detectaron 313 HVG, que es tan solo un 1,4 % de los genes totales (Figura 17C).

Únicamente 3 genes fueron clasificados como HVG por sendos métodos: IGHG3, IGHG4, GPNMB. IGHG3 e IGHG4 son dos genes que codifican parte de la estructura de dos subtipos de inmunoglobulinas G (IgG), un tipo de anticuerpo del sistema inmune, por lo que ambos podrían estar relacionados con la patogénesis de la EM (122). GPNMB codifica una glucoproteína transmembrana, que se expresa altamente en macrófagos y células de la microglía, ambas implicadas en la respuesta inmune innata (123).

No obstante, prácticamente la totalidad de los genes seleccionados como HVG por el método de grupo de covarianzas también lo fueron al aplicar el modelo de regresión de LOESS (1.423 genes comunes), lo que muestra una alta coincidencia entre ambos métodos. En contraste, el método basado en los residuos de Pearson no solo identificó el menor número de HVG, sino que además mostró una escasa intersección con los otros dos métodos evaluados. Esta limitada coincidencia, sumado a que un porcentaje demasiado pequeño de genes no siempre resulta adecuado para capturar bien toda la variabilidad biológica de las muestras, podría afectar directamente a los análisis posteriores de reducción de la dimensionalidad y agrupamiento de tipos celulares.

Finalmente, cabe destacar que todos estos resultados están condicionados por los umbrales preestablecidos en cada uno de los métodos empleados, de manera que el uso de unos valores más permisivos conllevaría a la detección de un mayor número de HVG, mientras que unos umbrales más restrictivos la limitarían.

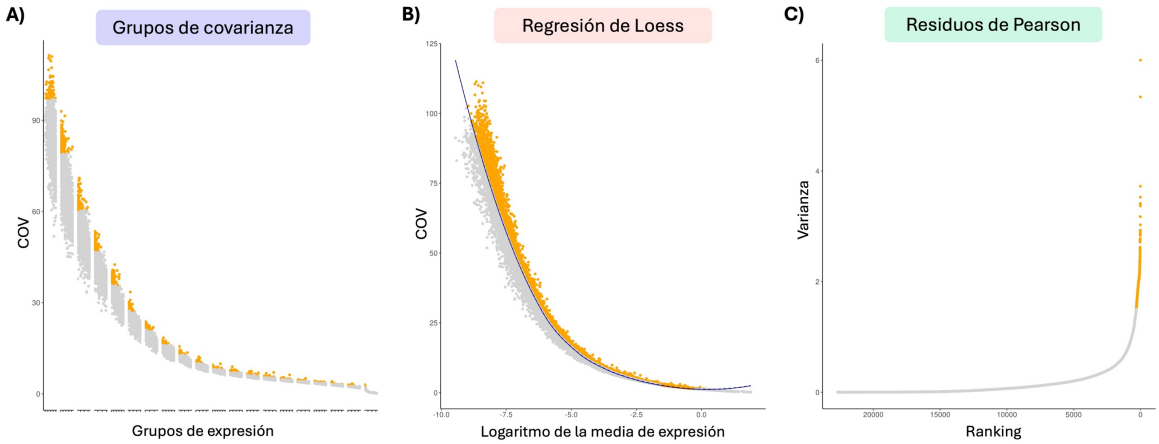

Figura 17. Identificación de los genes altamente variables en los datos de transcriptómica espacial filtrados y log-normalizados por tamaño de librería. Los puntos de color naranja representan los genes altamente variables (HVG, por sus siglas en inglés) seleccionados por cada método; y los puntos grises, los no seleccionados. A) Grupos de covarianza. Los HVG tienen un *z-score* del coeficiente de variación (COV, por sus siglas en inglés) superior a 1,5 dentro de cada uno de los 20 grupos de expresión generados. B) Modelo de regresión de LOESS. Los HVG tienen un COV que supera en 0,1 unidades al COV esperado. En azul oscuro se muestra la línea de tendencia. C) Método de residuos de Pearson. Los HVG tienen una varianza superior a 1,5.

4.5. Reducción de la dimensionalidad

El PCA se realizó calculando los 50 componentes principales a partir de los tres conjuntos de HVG previamente seleccionados, empleando tanto los valores log-normalizados por tamaño de librería como los valores normalizados con los residuos de Pearson. A pesar de que la normalización de Pearson ha sido señalada como una metodología que favorece la captura de variación biológica significativa en la reducción de la dimensionalidad (80), el porcentaje de variabilidad explicado por los primeros componentes principales en este conjunto de datos fue, en general, menor que cuando se utilizaron los valores log-normalizados por tamaño de librería (Figura 18A). Este resultado sugiere que, para estos datos en concreto, con la log-normalización por tamaño de librería se detecta de manera más eficiente la variabilidad global.

El objetivo de este análisis era poder agrupar posteriormente los *spots* por tipos celulares, independientemente de la condición experimental (CTRL, EM-CA o EM-CI), observándose que el método empleado para la selección de los HVG tenía una gran influencia en esta separación de grupos basada en la expresión.

Empleando los HVG identificados con el método de grupos de covarianza, se logró una separación de *spots* que parecía independiente del grupo experimental, de manera que los *spots* del mismo tipo celular podrían haberse agrupados juntos (Figura 18B). Esta observación se respalda también con el hecho de que este método es el implementado por defecto en *Giotto* y, por tanto, el recomendado por los desarrolladores de este paquete para realizar la reducción de la dimensionalidad. No obstante, considerando que los autores del estudio del que proceden los datos analizados en este trabajo anotaron hasta nueve tipos celulares distintos, requeridos para el posterior análisis de deconvolución, los grupos resultantes no mostraron una separación suficientemente definida, lo que sugirió un posible solapamiento entre tipos celulares diferentes.

En contraste, los HVG identificados con el modelo de regresión de LOESS o el método de los residuos de Pearson generaron grupos bien definidos y con una clara continuidad, siendo especialmente evidente en el modelo de LOESS (Figura 18B). Sin embargo, en este caso, los *spots* de cada una de las muestras se agruparon principalmente de acuerdo con su tipo de lesión, indicando que había más diferencias entre los grupos de pacientes que dentro de cada uno de ellos.

Puesto que todas las muestras analizadas procedían del mismo tipo de tejido (cerebro), lo esperable sería encontrar similitudes entre tipos celulares equivalentes entre las diferentes condiciones. Por ello, esta agrupación debida a grupo experimental podría deberse a un efecto de lote entre las muestras, que estaría enmascarando la heterogeneidad celular y que por tanto debe corregirse. Estos resultados pusieron de manifiesto la necesidad de aplicar alguna técnica de integración más compleja que permitiera alinear correctamente las muestras y mitigar los efectos de lote o condición experimental de forma previa al agrupamiento.

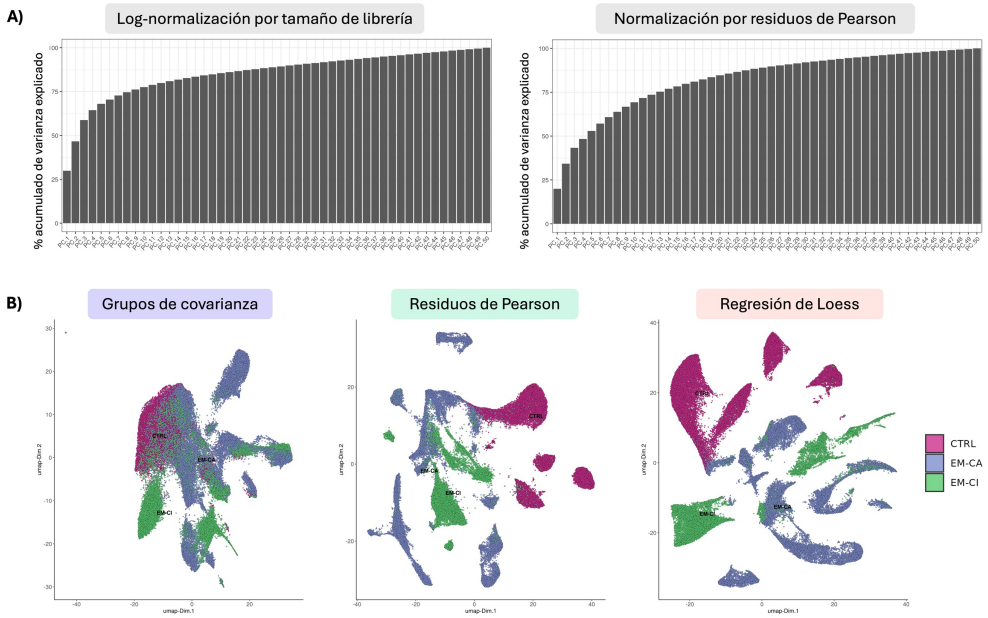

Figura 18. Reducción de la dimensionalidad de los datos de transcriptómica espacial filtrados y normalizados. A) Porcentaje de varianza acumulada a lo largo de los 50 componentes principales calculados mediante un Análisis de Componentes Principales (PCA, por su siglas en inglés) a partir de los genes altamente variables (HVG, por sus siglas en inglés) seleccionados con el modelo de regresión de LOESS, utilizando los valores log-normalizados por tamaño de librería (izquierda) y los valores normalizados con los residuos de Pearson (derecha). B) Representación UMAP (del inglés *Uniform Manifold Approximation and Projection*) de los *spots* del conjunto de datos filtrado y log-normalizado por tamaño de librería a partir de los HVG identificados con el método basado en grupos de covarianza (izquierda), el método de residuos de Pearson (centro) y el modelo de regresión de LOESS (derecha). Los *spots* están coloreados por grupo experimental: control (CTRL, rosa), lesiones crónicas activas de esclerosis múltiple (EM-CA, azul) y lesiones crónicas inactivas de esclerosis múltiple (EM-CI, verde).

4.6. Agrupamiento

En vista de los resultados de la reducción de la dimensionalidad previa, para asegurar que el agrupamiento se realizaba en base a los diferentes tipos de células, y no a efectos de lote derivados del procesamiento en diferentes días o a grupos experimentales distintos, se aplicó la técnica de integración Harmony. A pesar de que el método de grupos de covarianza para seleccionar HVG, como se ha visto en el apartado anterior, demostró ser el más eficaz para identificar la heterogeneidad celular en este conjunto de datos, como subespacio de entrada para la integración Harmony se prefirieron utilizar los 10 primeros componentes principales calculados a partir de los HVG seleccionados por el modelo de regresión de LOESS. Esta decisión se fundamentó en que el mayor número de HVG identificados por este método ofrece una representación más completa de la variabilidad celular de las muestras.

La integración Harmony convergió eficientemente tras 6 iteraciones, generando un subespacio dimensional corregido sobre el que se construyeron los dos tipos de redes de vecinos más cercanos (redes sNN y kNN). Ambas redes fueron la base para realizar el agrupamiento de Leiden con varios valores de resolución, obteniendo como resultado distinto número de grupos en función de la red empleada y la resolución seleccionada (Tabla 4). El algoritmo de Leiden tiene en cuenta la expresión génica para la agrupación, de manera que los *spots* con perfiles de expresión similares quedaron clasificados dentro del mismo grupo.

Tabla 4. Número de grupos obtenidos en el agrupamiento de Leiden para cada valor de resolución y tipo de red. En todos los casos se aplicó una *k* (número de vecinos más cercanos) igual a 15.

Resolución	Red kNN	Red sNN
0,15	4	10
0,25	7	17
0,50	12	24

sNN: *Shared Nearest Neighbor* (vecino más cercano compartido).
kNN: *k-Nearest Neighbors* (*k*-vecinos más cercanos).

Para valores equivalentes de resolución, la utilización de redes sNN como situación inicial del algoritmo de Leiden originó un mayor número de grupos en comparación con las redes kNN (Figura 19). Esta diferencia se debe principalmente a que las redes sNN tienen mayor sensibilidad de detección de subpoblaciones, ya que para determinar que dos *spots* pertenecen al mismo grupo, consideran tanto su proximidad como la cantidad de vecinos que comparten, lo que conduce a la formación de grupos de Leiden más pequeños (es decir, con menos *spots*) (Figura 19A). En contraposición, los grupos calculados a partir de las redes kNN mostraron una organización más definida a lo largo de las ubicaciones espaciales (Figura 19B).

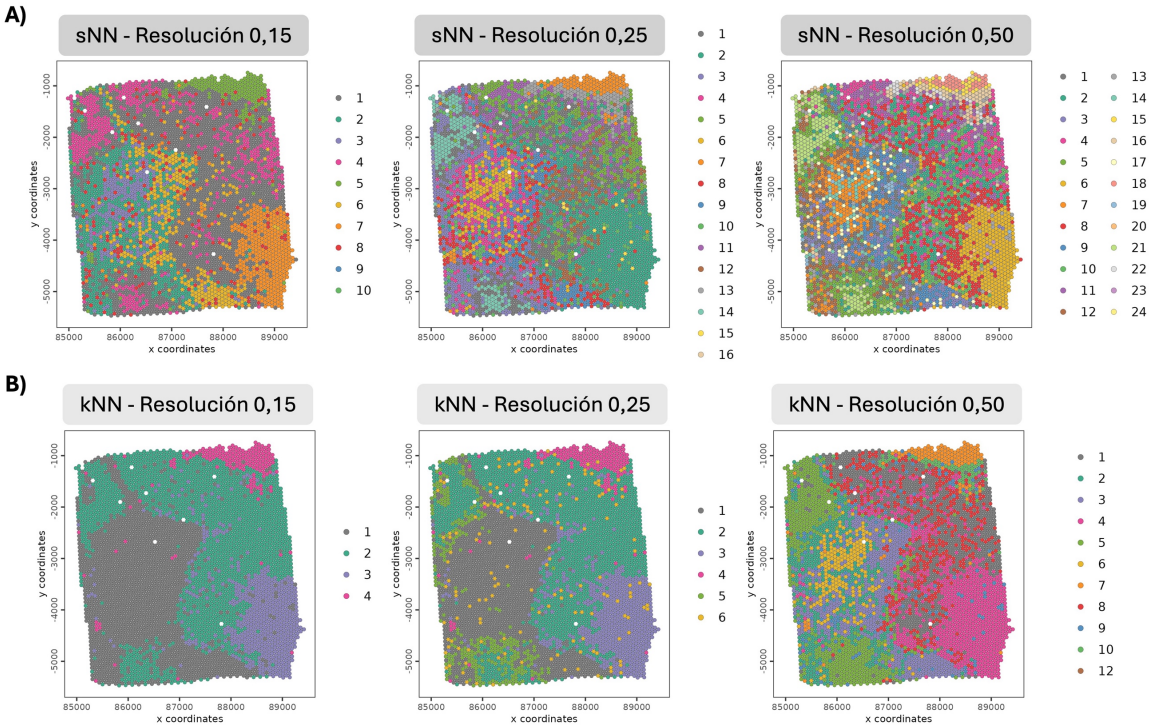

Figura 19. Visualización espacial de los grupos de Leiden generados para una muestra ejemplo del grupo de lesiones crónicas activas de esclerosis múltiple (MS377T). Se muestran los resultados al utilizar como grafo de entrada A) redes de *k*-vecinos más cercanos (kNN, por sus siglas en inglés) y B) redes de vecinos más cercanos compartidos (sNN, por sus siglas en inglés) para diferentes resoluciones de agrupamiento: 0,15 (izquierda), 0,25 (centro) y 0,50 (derecha). Cada punto representa un *spot*, y los colores indican su asignación a los distintos grupos de Leiden. Los puntos blancos representan los *spots* filtrados, es decir, que tenían menos de 100 genes detectados.

En cualquier caso, el aumento del nivel de resolución se tradujo en una subdivisión progresiva de los grupos identificados, formando grupos de menor tamaño. Esto es lo que ocurre al calcular los grupos de Leiden a partir de una red kNN (Figura 20). Los grupos 7 y 10 obtenidos para una resolución de 0,50 correspondían al grupo 4 identificado para una resolución de 0,15 y 0,25, mientras que el grupo 1 de las resoluciones más bajas se dividió en tres nuevos subgrupos cuando se utilizó una resolución de 0,50 (subgrupos 2, 3 y 6). Sin embargo, valores de resolución demasiado altos pueden dar lugar a la sobreestimación de grupos, es decir, a la aparición de múltiples grupos independientes que, en realidad, están estrechamente relacionados a nivel biológico y, por tanto, podrían considerarse como una única población.

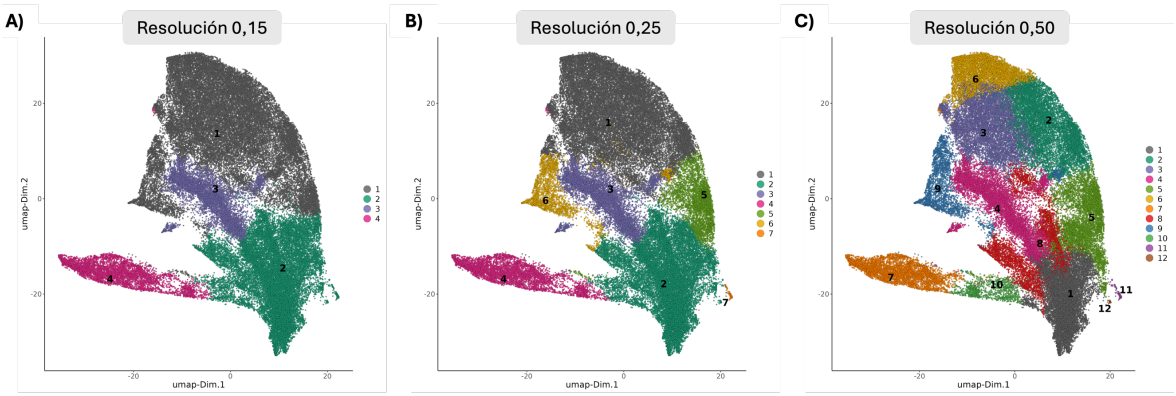

Figura 20. Representación UMAP calculada a partir del subespacio dimensional resultante de la integración Harmony. Los *spots* están coloreados según el grupo de Leiden asignado a partir de una red de *k*-vecinos más cercanos (kNN, por sus siglas en inglés) para diferentes valores de resolución: A) 0,15, B) 0,25 y D) 0,50.

No existe un consenso establecido sobre cuál es el número óptimo de grupos que se deben definir, sino que este se debe ajustar a los objetivos concretos del estudio, al contexto biológico y al nivel de resolución que se desea conseguir. Por ejemplo, en este trabajo, la determinación del número de grupos más adecuado podría basarse tanto en el número de tipos celulares identificados en el cerebro (o subtipos para mayor detalle) como en las diferentes características tisulares de las muestras de EM respecto a los controles.

En este caso concreto, dado que los autores originales anotaron un total de nueve tipos celulares en sus datos de snRNA-Seq (información requerida para la posterior deconvolución de tipos celulares), para los análisis posteriores se optó por utilizar los grupos de Leiden calculados a partir de la red kNN con una resolución de 0,50. Esta elección se basó en que con esta configuración se identificaron una cantidad de grupos similar al total de tipos celulares y en que se obtuvo una definición de las regiones espaciales superior a la lograda con las redes sNN.

En general, los *spots* de un mismo grupo correspondían a regiones tisulares cercanas, lo que sugirió que los patrones de expresión génica también seguían un patrón espacial que podría estar determinado por los tipos celulares mayoritarios de cada región tisular. Las muestras CTRL mostraron una distribución relativamente homogénea de los grupos de Leiden, sin regiones con una clara predominancia de un único grupo (Figura 21A).

Sin embargo, en las muestras procedentes de pacientes con EM sí que se observaron zonas dominadas por uno o dos grupos (Figura 21B y Figura 21C), siendo especialmente notorio en las muestras del grupo EM-CA. Las muestras de las lesiones de EM, según indicaron los autores de los datos en su publicación, capturaron tanto el núcleo desmielinizado como el borde inflamado de la lesión, por lo que los grupos específicos de algunas regiones (como los grupos 1, 4 y 8, o los grupos 2, 3 y 6) podrían estar relacionados con la propia patología de la EM y las consecuentes lesiones crónicas, tanto activas como inactivas. La visualización espacial de los grupos de Leiden definidos para cada una de las muestras se muestra en la Figura A2 del Anexo.

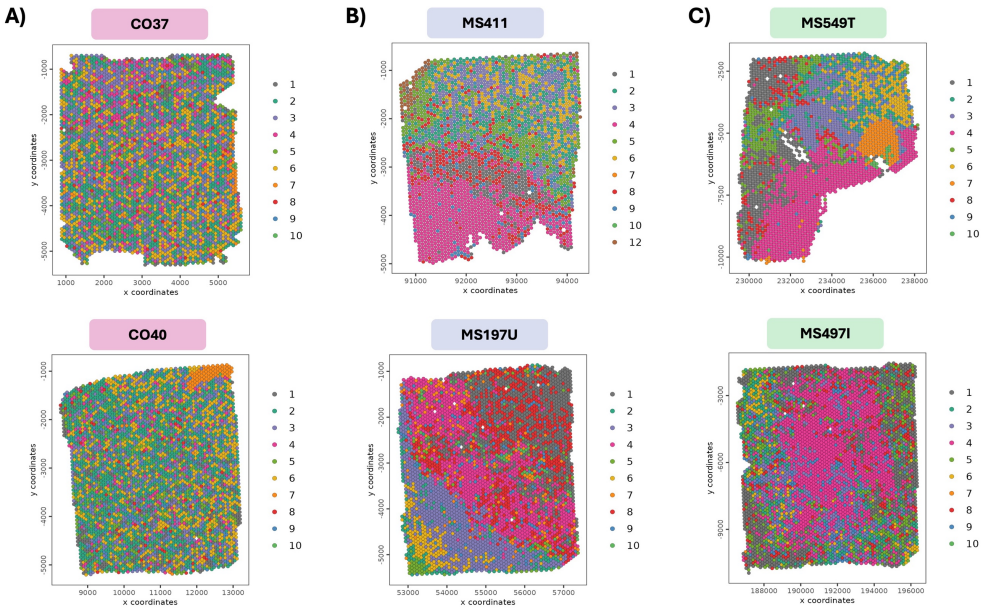

Figura 21. Visualización espacial de los grupos de Leiden obtenidos en seis muestras ejemplo de diferente grupo experimental. A) Control (CO37 y CO40), B) lesiones crónicas activas de esclerosis múltiple (MS411 y MS197U) y C) lesiones crónicas inactivas de esclerosis múltiple (MS549T y MS497I). Los grupos fueron definidos a partir de una red de *k*-vecinos más cercanos y con una resolución de 0,50. La fila superior corresponde a muestras de hombres y la fila inferior a muestras de mujeres. Cada punto representa un *spot*, coloreado según su asignación a los distintos grupos de Leiden. Los puntos de color blanco son los *spots* filtrados por tener menos de 100 genes detectados.

Una vez definidos los grupos de Leiden, se calcularon los genes diferencialmente expresados entre cada uno de los grupos. La correcta identificación de estos genes marcadores es un paso esencial para la posterior anotación de los tipos celulares que predominan en los *spots* de cada grupo (deconvolución). Cada grupo de Leiden presentó una firma genética distinta, mostrando en general una elevada expresión de genes específicos de algún tipo celular concreto. Por ejemplo, los grupos 2, 3, 5 y 6 mostraron una alta expresión de genes característicos de los oligodendrocitos y las vainas de mielina (PLP1 y MBP); mientras que el grupo 7 tenía una sobreexpresión de genes relacionados con las sinapsis neuronales (SNAP25 y GPM6A) (Figura 22). La especificidad de la expresión génica en cada grupo de Leiden confirmó que este tipo de agrupamiento era capaz de reflejar correctamente la diversidad celular del tejido cerebral.

El hecho de que algunos grupos presentan patrones de expresión muy similares, como sucede en los grupos 2 y 6, por ejemplo, sugiere que estos grupos podrían estar formados por los mismos tipos de células o estar implicados en las mismas funciones biológicas. Además, se observó que estos grupos correlacionados tendían a agruparse próximos en el subespacio dimensional (Figura 20C), lo que refuerza la idea de que comparten características transcriptómicas.

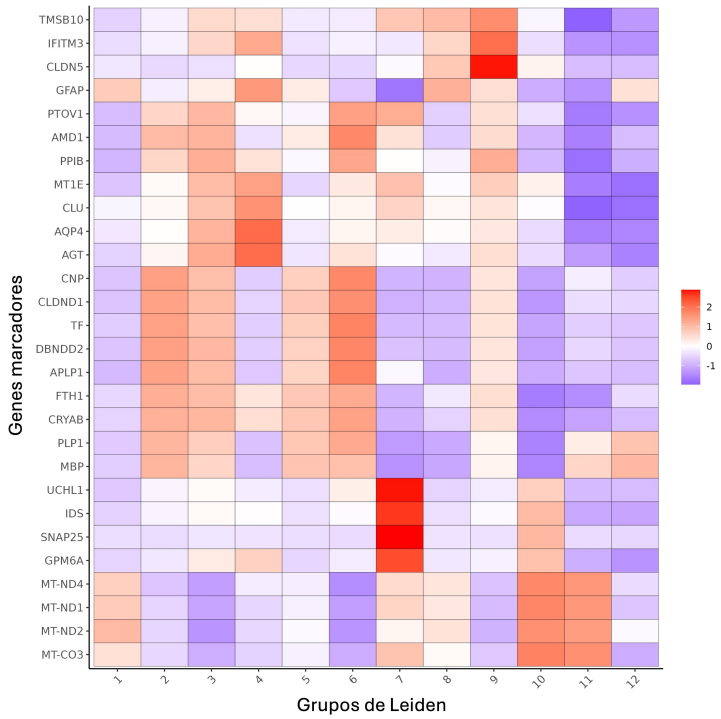

Figura 22. Mapa de calor que muestra la expresión normalizada de los dos genes con mayor expresión diferencial de cada grupo de Leiden. Para amplificar visualmente las diferencias, la expresión de cada gen se transformó en un *z-score* a partir de la media y la desviación estándar de la expresión de cada grupo de Leiden. Los tonos rojos reflejan la sobreexpresión del gen en ese grupo en comparación con el resto de grupos, y los tonos azules indican subexpresión del gen en ese grupo respecto a los demás.

4.7. Anotación celular

4.7.1. Conjunto de datos de snRNA-Seq de referencia

El atlas completo de snRNA-Seq estaba disponible como un fichero H5AD, a partir del cual se generó en R un objeto de clase *SingleCellExperiment*. Según lo especificado en la publicación original, los investigadores que realizaron el procesamiento de los datos de snRNA-Seq empleados en este trabajo eliminaron todos los núcleos que expresaban menos de 200 genes, tenían un tamaño de librería superior al percentil 99 o presentaban un porcentaje de genes mitocondriales superior al 5 %. Asimismo, eliminaron los genes que no presentaban algún conteo en al menos 3 núcleos por muestra, asegurando la relevancia biológica de los genes conservados (121).

El control de calidad realizado a los datos de snRNA-Seq reveló que los datos descargados ya habían sido filtrados previamente, ya que todos los núcleos incluidos presentaban más de 200 genes (Figura 23A), un tamaño de librería en general sin valores extremos que pudieran corresponder a dobletes (Figura 23B), y un porcentaje de genes mitocondriales inferior al 5 % en todas las muestras (Figura 23C). Además, el bajo número de dobletes identificados (8 % del total de núcleos) sugirió que estos también habían sido descartados por los autores. En conjunto, los resultados confirmaron que el atlas de snRNA-Seq ya había sido sometido a un control de calidad y filtrado de genes y núcleos de baja calidad, gotas vacías y dobletes, por lo que no fue necesario realizar un filtrado adicional previo a la deconvolución. En total, este estudio abarcó un total de 32.115 genes y 103.794 núcleos procedentes de todas las muestras.

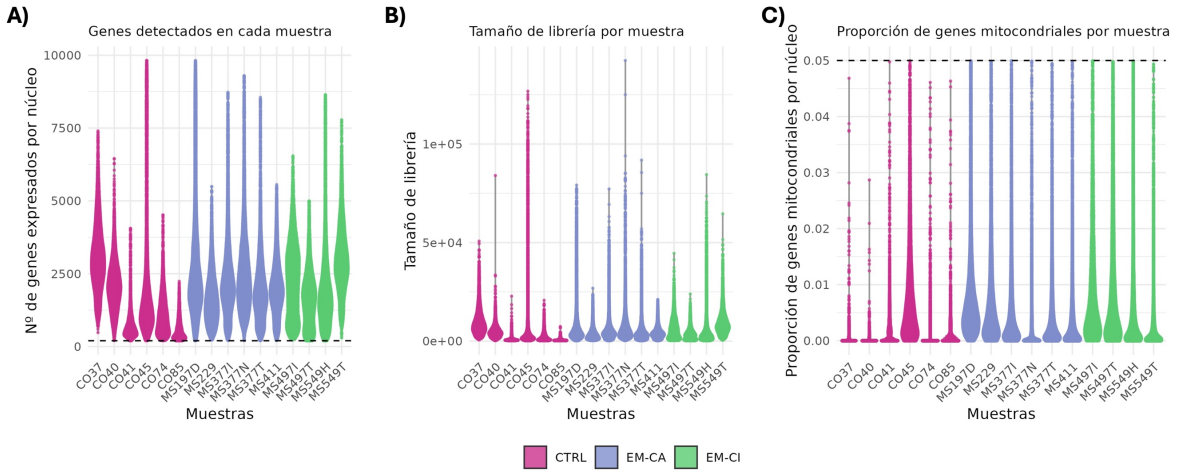

Figura 23. Distribución de los indicadores de calidad calculados para los datos de secuenciación de ARN de núcleo único (snRNA-Seq). A) Número total de genes detectados por núcleo en cada muestra. La línea negra discontinua señala el umbral de filtrado establecido por los autores de estos datos (> 200 genes). B) Tamaño de librería (conteos totales en cada núcleo) por muestra. C) Proporción de genes mitocondriales por núcleo en cada muestra (rango de 0 a 1). La línea negra discontinua muestra el umbral de filtrado aplicado por los autores (< 5 % de genes mitocondriales). Los gráficos de violín representan la distribución de cada parámetro en las distintas muestras, diferenciándose entre los tres grupos experimentales: control (CTRL, rosa), lesiones crónicas activas de esclerosis múltiple (EM-CA, azul) y lesiones crónicas inactivas de esclerosis múltiple (EM-CI, verde).

Para contrarrestar el ruido técnico o el posible sesgo derivado de la profundidad de secuenciación, la matriz de conteos fue log-normalizada por tamaño de librería (método por defecto de *Giotto*). A fin de reducir la dimensionalidad, se realizó una selección de los HVG con el método de los grupos de covarianza, identificando un total de 2.448 genes cuyos niveles de expresión presentaban una gran variabilidad entre los núcleos. A partir de estos genes, se calcularon los 50 primeros componentes principales mediante PCA, explicando casi el 75 % de la variabilidad con los 10 primeros componentes principales (Figura 24A).

La mayor fuente de variabilidad de estos datos se debía a los distintos tipos celulares anotados, lo cual se evidenció en la separación de grupos basada en la expresión debida principalmente a esta variable (Figura 24B). Todos los núcleos de un mismo tipo celular se agruparon correctamente, ya que no se observó ningún grupo con varios tipos celulares mezclados, lo que podría indicar una anotación celular deficiente. Por tanto, los datos de snRNA-Seq analizados eran un buen atlas de referencia para la deconvolución de tipos celulares en los datos de ST.

Finalmente, cabe resaltar que en todas las muestras que conformaban este *dataset*, los oligodendrocitos constituían el tipo celular predominante, lo que concuerda con su abundancia mayoritaria en la sustancia blanca del cerebro. Por otro lado, mientras que las muestras procedentes de individuos sanos (CTRL) carecían de células B, todas las correspondientes a pacientes con EM presentaron cierto porcentaje de este tipo celular, lo que puede asociarse a la infiltración de células inmunes en el cerebro característica de esta patología (Figura 24C).

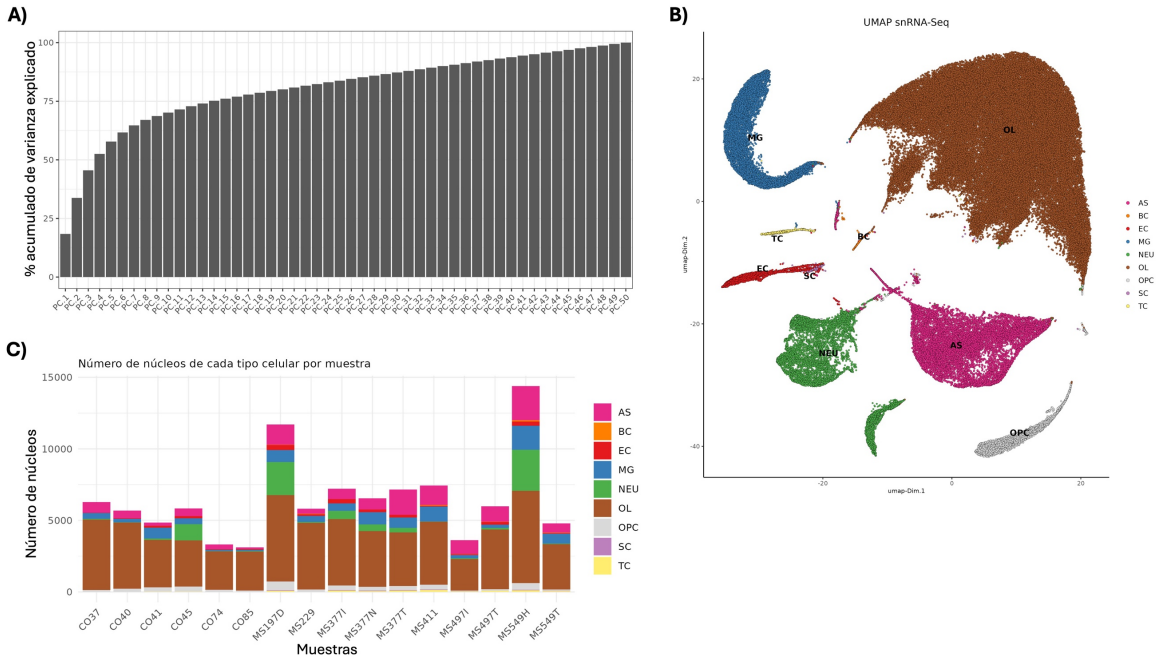

Figura 24. Resultados del análisis de los datos de secuenciación de ARN de núcleo único. A) Porcentaje de varianza acumulada explicada a lo largo de los 50 componentes principales calculados a partir de los genes altamente variables (HVG) seleccionados con el método de grupos de covarianza. B) Representación UMAP generada a partir de los 10 primeros componentes principales coloreada por tipo celular. C) Distribución de los núcleos de cada tipo celular para cada muestra. Los colores de los gráficos B) y C) representan los distintos tipos celulares anotados: astrocitos (AS), células B (BC, por sus siglas en inglés), células endoteliales (EC, por sus siglas en inglés), microglía (MG), neuronas (NEU), oligodendrocitos (OL), células progenitoras de oligodendrocitos (OPC, por sus siglas en inglés), células del estroma (SC, por sus siglas en inglés) y células T (TC, por sus siglas en inglés).

4.7.2. Deconvolución

Utilizando el atlas de snRNA-Seq emparejado como referencia, se deconvolucionaron los datos espaciales con el algoritmo SpatialDWLS implementado en *Giotto*. Este análisis reveló diferencias en la composición celular entre las secciones de tejido CTRL y las muestras de pacientes con EM (Figura 25).

Las áreas de tejido de las muestras CTRL estaban compuestas uniformemente por oligodendrocitos (Figura 25A y Figura 25B), un tipo de célula glial encargada de la formación y mantenimiento de las vainas de mielina en el SNC (124).

Sin embargo, en las muestras de tejido correspondientes a las lesiones de EM se observó una mayor diversidad celular. En este caso, los oligodendrocitos se encontraban en una menor proporción en algunas regiones específicas (Figura 25B), posiblemente debido a la desmielinización consecuente de los ataques autoinmunes característicos de esta patología. En su lugar, estas zonas presentaban una alta abundancia de astrocitos (Figura 25C), un tipo celular relacionado con la patogénesis de la EM al propagar la inflamación. Puesto que la activación de astrocitos reactivos es una característica de las lesiones tisulares activas (muestras EM-CA) y de la formación de las cicatrices gliales propias de las lesiones inactivas (muestras EM-CI) (28), las regiones tisulares con pocos oligodendrocitos y alta proporción de astrocitos podrían tratarse del núcleo desmielinizado de las lesiones de EM.

Algunas muestras de EM también presentaron cierto porcentaje de células B en el núcleo de la lesión (Figura 25A), que podrían corresponder con los espacios perivasculares asociados a las lesiones.

Asimismo, algunas muestras EM-CA mostraron una región alrededor del núcleo desmielinizado formada principalmente por microglía (Figura 25D), pudiendo tratarse en este caso del borde inflamado de la lesión (28). En las muestras de EM-CI este borde no estaba tan bien definido, y presentaban una menor infiltración de células inmunes, lo que refleja una disminución de la actividad inflamatoria. La mayor parte del área de tejido de las muestras EM-CI se correspondió con las cicatrices gliales compuestas por astrocitos (Figura 25C), aunque también presentaron algunas regiones que estaban comenzado a remielinizarse como resultado de la diferenciación de las células progenitoras de oligodendrocitos, evidenciada por la presencia de oligodendrocitos (Figura 25B).

Finalmente, cabe destacar que algunas secciones de tejido presentaban zonas con alta abundancia en neuronas (Figura 25A), lo que se puede atribuir a áreas de sustancia gris capturadas al seccionar el tejido y obtener la muestra. Este hecho se respalda con las descripciones de los autores de los datos en su publicación sobre cómo tomaron las muestras de cerebro (121).

En la Figura A3 del Anexo se muestran los resultados de la deconvolución obtenidos para cada una de las muestras. Los diferentes tipos celulares se alinearon espacialmente con la organización tisular esperada tanto en las muestras CTRL como en las de EM. Estos resultados en conjunto ponen de manifiesto la pérdida de la estructura normal dominada por los oligodendrocitos en la EM, reemplazándose principalmente por astrocitos y microglía.

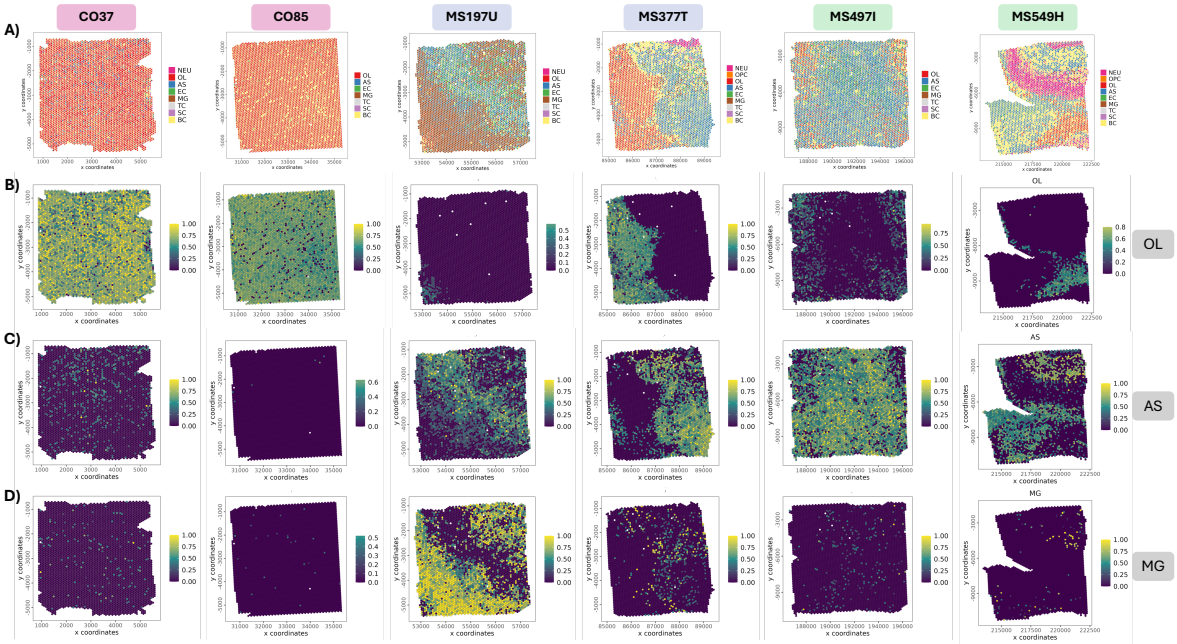

Figura 25. Visualización espacial de la deconvolución en seis muestras ejemplo de diferente grupo experimental. Las muestras control se resaltan con marcos rosas (CO37 y CO85), las de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos azules (MS197U y MS377T) y las muestras de lesiones crónicas inactivas de esclerosis múltiple con marcos verdes (MS497I y MS549H). A) Resultados de la deconvolución, donde cada gráfico de sectores muestra el porcentaje de cada tipo celular por *spot*. B) Proporción de oligodendrocitos por *spot*. C) Proporción de astrocitos por *spot*. D) Proporción de microglía por *spot*. AS: astrocitos. BC: Células B. EC: células endoteliales. MG: microglía. NEU: neuronas. OL: oligodendrocitos. OPC: células progenitoras de oligodendrocitos. SC: células del estroma. TC: células T.

4.8. Patrones espaciales de expresión

La identificación de patrones espaciales de expresión requirió de la construcción previa de un grafo no dirigido que representase la relación espacial entre *spots*, para lo que se construyó una red espacial de Delaunay. Aproximadamente el 70 % de los *spots* estaban conectados a 6 vecinos directos, lo cual es coherente con la geometría hexagonal de las *spots* de *Visium*. El resto de los *spots* presentaron entre 3 y 5 conexiones, lo que corresponde con los bordes de las secciones de tejido o con huecos internos debidos a roturas en el tejido o a los *spots* filtrados.

La detección de los SVG se realizó a partir de la red espacial con el algoritmo BinSpect, aplicando los dos métodos de binarización implementados en *Giotto*. Con la binarización con *k-means* se detectaron un total de 14.953 SVG significativos con un FDR de 0,01 y con la binarización con el método *rank* se identificaron 16.692 SVG para este mismo FDR. No obstante, un gran subconjunto de estos genes fue identificado por ambos métodos (14.837 SVG), incluyendo genes específicos de los tipos celulares característicos de las lesiones de EM.

Por ejemplo, uno de los genes con un patrón espacial de expresión claramente definido fue MBP, un gen que codifica uno de los componentes proteicos más abundantes de la vaina de mielina de los oligodendrocitos. El gen MBP presentó niveles de expresión bajos en las regiones correspondientes a las lesiones de las muestras de EM, mientras que su expresión fue notablemente más alta en las zonas con una mayor proporción de oligodendrocitos (Figura 26A). El gen GFAP, que codifica uno de los filamentos proteicos de los astrocitos, mostró patrones espaciales de expresión opuestos a MBP: niveles de expresión altos en las lesiones de EM y bajos en el resto de las regiones (Figura 26B). Además, un gran número de genes que codifican las IgG también fueron considerados SVG, observándose una alta expresión de estos genes en las lesiones de EM, como se ejemplifica con el patrón de expresión del gen IGHG4 (Figura 26C).

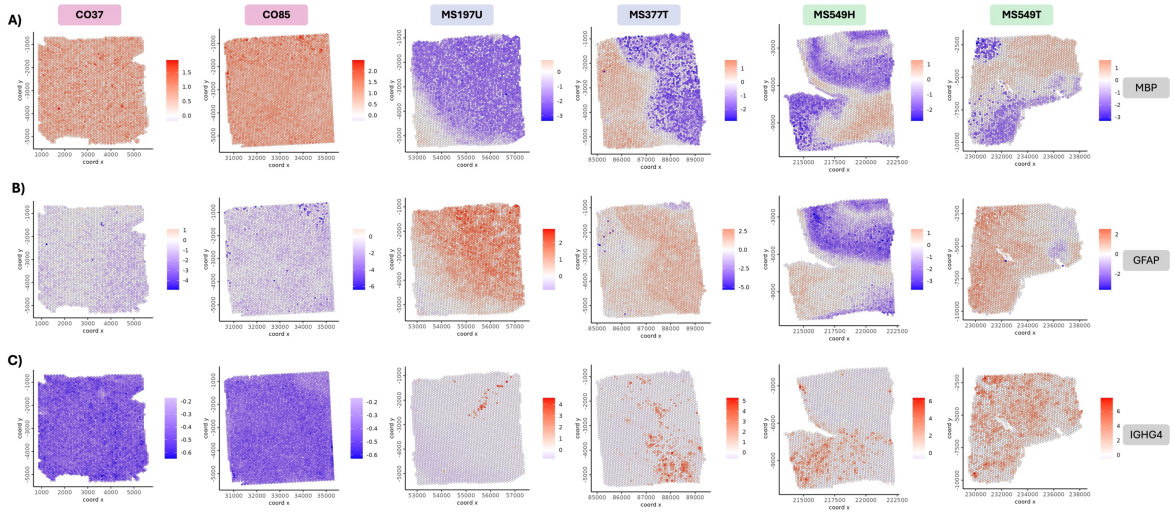

Figura 26. Visualización espacial de los niveles de expresión de tres genes espacialmente variables identificados con BinSpect en seis muestras ejemplo de diferente grupo experimental. Cada punto representa un *spot*, y el gradiente de colores muestran el nivel de expresión log-normalizado por tamaño de librería y escalado en forma de *z-score* de los genes: A) MBP (específico de oligodendrocitos), B) GFAP (específico de astrocitos) y C) IGHG4 (inmunoglobulina). Los tonos rojos y azules reflejan altos y bajos niveles de expresión, respectivamente. La escala de color está centrada en 0. Las muestras control se resaltan en rosa (CO37 y CO85), las de lesiones crónicas activas de esclerosis múltiple en azul (MS197U y MS377T) y las de lesiones crónicas inactivas de esclerosis múltiple en verde (MS549H y MS549T).

A partir de los 500 genes con mayor probabilidad de ser SVG (menor p-valor ajustado), se realizó la agrupación espacial de los *spots* en nueve dominios con el algoritmo de HMRF. Este algoritmo define los dominios espaciales considerando tanto los perfiles de expresión génica de los *spots* como sus coordenadas espaciales, a diferencia del algoritmo de Leiden que agrupa únicamente en función de las similitudes en el nivel de expresión. Esta inferencia del estado de dominio de cada *spot* se aplicó para diferentes valores del factor de suavizado β, el cual modula la influencia de la información espacial en el refinamiento del agrupamiento.

El resultado esperado para β igual a 5 (baja influencia espacial) era la formación de dominios espaciales altamente fragmentados. Sin embargo, en algunas muestras de EM se observó que el modelo no lograba distinguir correctamente los diferentes patrones espaciales, asignando la mayoría de los *spots* a un único dominio (Figura 27A). El aumento del valor de β (mayor influencia espacial) permitió resolver los patrones que estaban "enmascarados" para un valor de β bajo, resultando en la obtención de grupos espaciales más definidos y coherentes con la anotación de los tipos celulares mayoritarios de cada región tisular (descrito en el Apartado 4.7.2).

Los incrementos en el parámetro β a partir de un valor 15 no produjeron cambios significativos en la distribución de los patrones espaciales (Figura 27B, Figura 27C y Figura 27D). Este hecho indicó que el modelo de HMRF había alcanzado un punto de convergencia en el suavizado espacial. Puesto que la estructura espacial subyacente de los datos de ST quedaba completamente capturada para β igual a 15, la caracterización de los dominios se realizó utilizando estos últimos resultados.

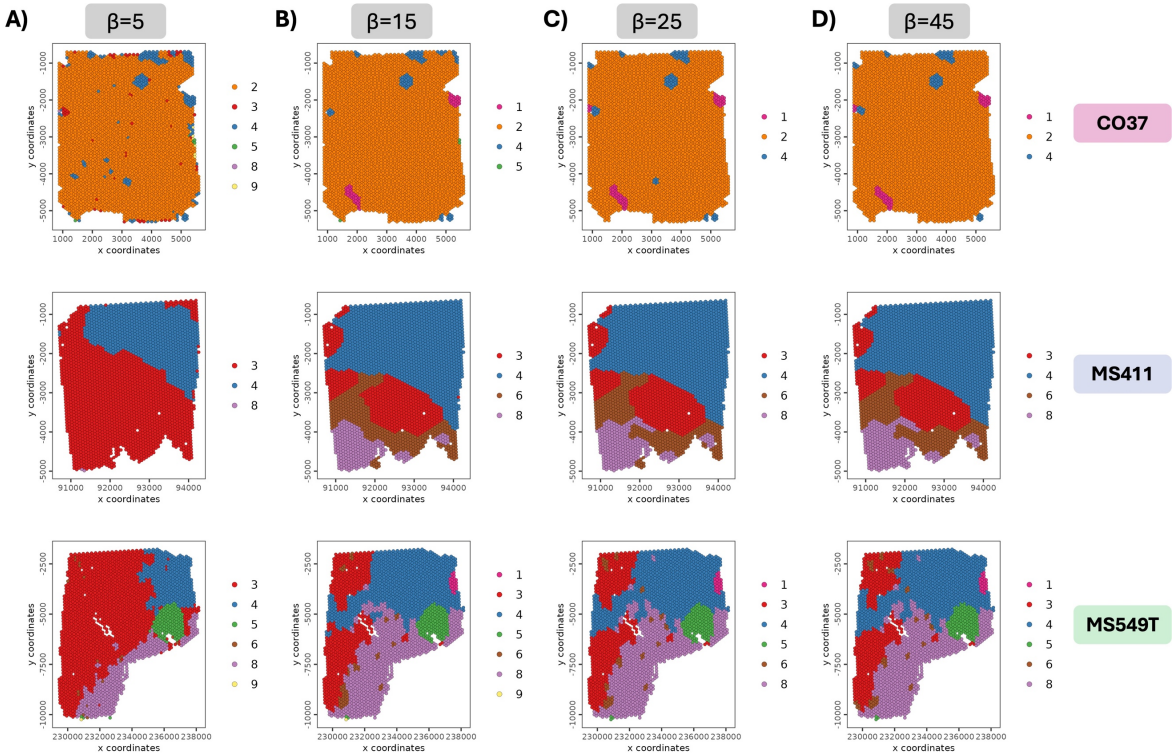

Figura 27. Visualización espacial de los dominios espaciales identificados con el modelo de Campo Aleatorio Oculto de Markov en tres muestras ejemplo para diferentes valores de β. A) $\beta = 5$, B) $\beta = 15$, $\beta = 25$, $\beta = 45$. El parámetro β regula la fuerza de interacción entre *spots*, de manera que valores mayores dan más peso a la información espacial en el refinamiento del agrupamiento respecto al perfil de expresión. Cada punto representa un *spot*, y los colores indican su asignación a los distintos dominios espaciales identificados. La muestra control (CO37) se resalta con un marco de color rosa, la de la lesión crónica activa de esclerosis múltiple en azul (MS411) y la de lesión crónica inactiva en verde (MS549T).

Para realizar la anotación de los dominios espaciales en las distintas áreas tisulares descritas para las lesiones de EM (Figura 28), se compararon los dominios espaciales identificados mediante el modelo HMRF con el patrón de los tipos celulares deconvolucionados (Figura 25 y Figura A3) y la expresión de sus principales genes marcadores (Figura 26).

Los *spots* de las muestras CTRL se agruparon de forma más uniforme y en un menor número de dominios espaciales diferentes, siendo algunos de ellos prácticamente exclusivos de esta condición experimental (dominios 1 y 2). Puesto que este grupo de muestras estaban compuestas mayoritariamente por oligodendrocitos y presentaban una alta expresión de genes implicados en el mantenimiento y la formación de la mielina (como MBP), estos dos dominios fueron anotados como sustancia blanca sana (Figura 28A).

El dominio 4 fue identificado en algunas muestras CTRL y en localizaciones específicas de las muestras de EM, coincidiendo en ambos casos con zonas con una elevada proporción de oligodendrocitos. Por ello, este dominio también se clasificó como sustancia blanca, diferenciando en este caso entre sustancia blanca sana en las muestras CTRL y sustancia blanca perilesional (adyacente a la lesión) en las muestras de EM (Figura 28B).

Los dominios 3, 6 y 8 se capturaron principalmente en las regiones espaciales de las muestras de EM caracterizadas por una alta densidad de astrocitos y alta expresión de genes asociados a su activación (como GFAP), lo que permitió la anotación de estos tres dominios espaciales como el núcleo desmielinizado de las lesiones de EM (Figura 28C).

El dominio 7 se atribuyó al borde inflamado de la lesión de EM, distinguiéndose en este caso por la abundante presencia microglía activada (Figura 28D). Por último, los dominios 5 y 9 se anotaron como sustancia gris, capturada al seccionar las muestras de tejido cerebrales junto con la sustancia blanca, basándose en su localización en regiones espaciales con una significativa concentración de neuronas y una alta expresión de genes implicados en las funciones neuronales (Figura 28E).

En la Figura 28F se muestra la integración final de todas las áreas tisulares anotadas, observándose que los dominios asignados a cada región tisular reflejan la heterogeneidad y arquitectura espacial de las lesiones de EM. Los dominios espaciales identificados en cada una de las muestras mediante el modelo de HMRF para β igual a 15 y su anotación manual se pueden consultar en la Figura A4 y Figura A5 del Anexo, respectivamente.

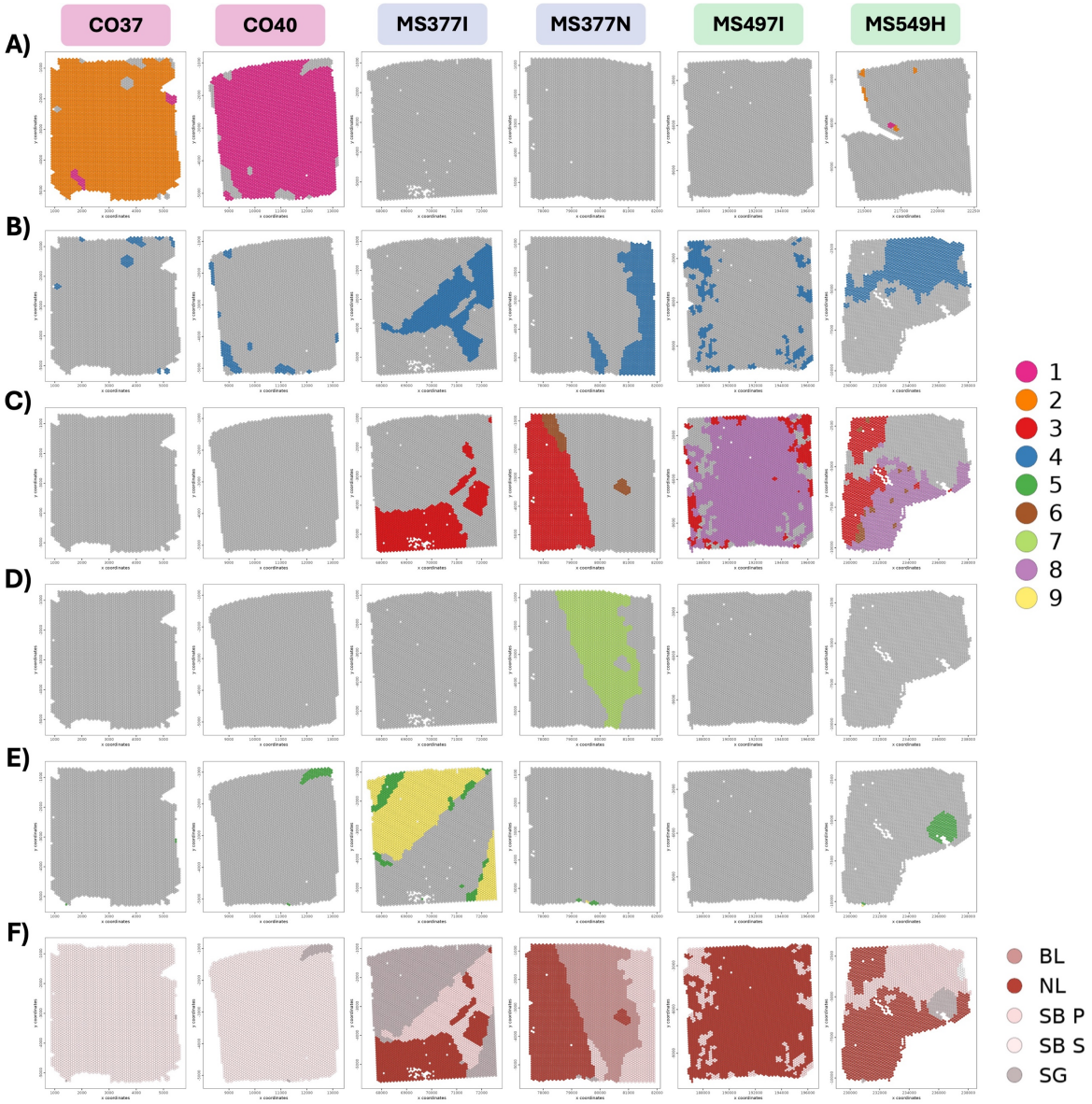

Figura 28. Visualización espacial de la anotación manual de los dominios identificados con el modelo de Campo Aleatorio Oculto de Markov (HMRF) en seis muestras ejemplo para $\beta = 15$. A) Sustancia blanca sana (dominios 1 y 2). B) Sustancia blanca sana en las muestras control y sustancia blanca perilesional en EM (dominio 4). C) Núcleo de la lesión de EM (dominios 3, 6 y 8). D) Borde inflamado de la lesión (dominio 7). E) Sustancia gris (dominios 5 y 9). F) Integración de las cinco regiones tisulares anotadas. El color de cada *spot* indica su asignación a los dominios espaciales (A-E) o a las áreas tisulares anotadas (F). Las muestras control se resaltan con marcos rosas (CO37 y CO40), las de lesiones crónicas activas de EM con marcos azules (MS377I y MS377N) y las de lesiones crónicas inactivas con marcos verdes (MS497I y MS549H). BL: borde de la lesión. NL: núcleo de la lesión. SB P: sustancia blanca perilesional (adyacente a las lesión) de las muestras de EM. SB S: sustancia blanca sana de las muestras control. SG: sustancia gris capturada al seccionar el tejido cerebral. EM: esclerosis múltiple.

4.9. Análisis de expresión diferencial

Para caracterizar los cambios en la expresión génica asociados al sexo entre las distintas áreas tisulares características de la EM, se llevó a cabo un análisis de expresión diferencial. En total, se realizaron cuatro contrastes diferentes (Tabla 3), considerando como diferencialmente expresados aquellos genes con un p-valor ajustado menor o igual a 0,05 y una magnitud de cambio absoluta (|logFC|) superior a 0,5.

La comparación de la expresión génica del núcleo desmielinizado de las lesiones de EM entre hombres y mujeres reveló un perfil de expresión diferencial entre ambos grupos de 47 genes. En general, en el núcleo de la lesión de los hombres se observó una sobreexpresión de algunos genes mitocondriales (MT-ATP8, MT-ND2, MT-ND1, MT-ND4L y MT-CO1), así como de genes relacionados con el sistema inmune (IGKC, IGHG4, IGHG3 e IGHG1) y la respuesta al estrés celular (HSPA1A, HSPA1B y HSP90AA1), entre otros. Por el contrario, en las mujeres, los genes sobreexpresados estaban principalmente vinculados a procesos inflamatorios (CHIT1, SERPINA3, APOE, HLA-C, S100A6 o SPP1) y a la remodelación del citoesqueleto (VIM, ACTB, TMSB4X o TMSB10) (Figura 29) (125).

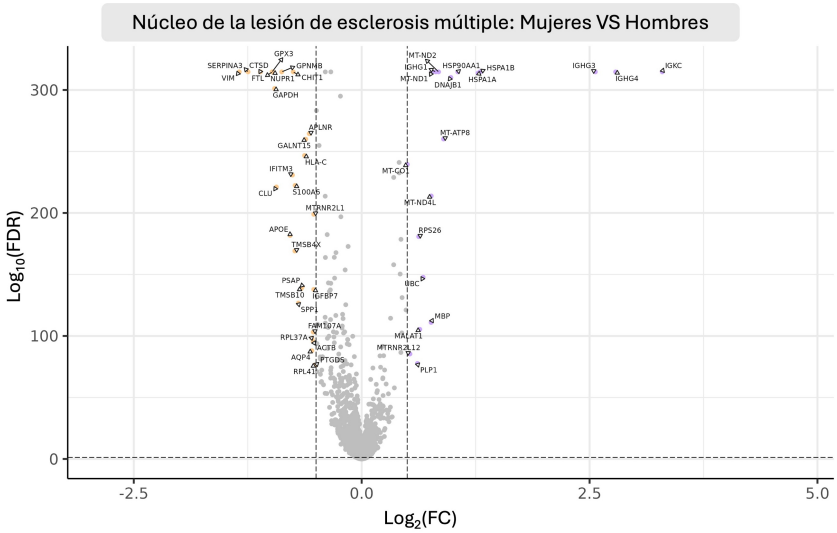

Figura 29. Gráfico de volcán con los genes diferencialmente expresados en el núcleo de la lesión de esclerósis multiple entre mujeres y hombres. El eje Y representa la significancia estadística y el eje X la magnitud de cambio. Los genes etiquetados corresponden a aquellos que presentan un FDR $\leq$ 0,05 y una magnitud de cambio absoluta ($|\log_2 FC|$) superior a 0,5. Los genes sobreexpresados en mujeres se muestran en color naranja y los sobreexpresados en hombres en color morado.

En el análisis de la sustancia blanca de las muestras CTRL (tejido sano) se detectaron 67 genes diferencialmente expresados entre hombre y mujeres, de los cuales 12 también habían sido identificados como significativos en la comparación anterior (Figura 30A). Este resultado mostró la existencia de diferencias transcriptómicas basales en ausencia de enfermedad entre hombres y mujeres, asociadas a dimorfismos sexuales del cerebro y no a la fisiopatología de la EM.

Por otro lado, las comparaciones entre las sustancia blanca perilesional y el núcleo de la lesión de EM evidenciaron la existencia de cambios asociados a la progresión de la enfermedad en ambos sexos. En el caso de los hombres se detectaron un total de 331 genes diferencialmente expresados entre ambas localizaciones tisulares, mientras que en las mujeres se identificaron 250 genes. De entre ellos, 217 genes estaban diferencialmente expresados en ambos grupos (Figura 30B). El alto grado de solapamiento entre mujeres y hombres sugirió que gran parte de los mecanismos moleculares que conllevan a la desmielinización son independientes de la variable sexo. Muchos de los genes comunes estaban relacionados con la pérdida de mielina en el núcleo de la lesión, evidenciada por la subexpresión de genes como MBP, PLP1, CNP, MOBP, MAG y MOG, entre otros. Asimismo, en ambos grupos se observó la sobreexpresión de algunos genes mitocondriales y de genes relacionados con la activación de los astrocitos reactivos, como es el caso del gen GFAP.

Sin embargo, la expresión de 114 genes cambiaba entre estas dos regiones únicamente en hombres, mientras que 33 genes lo hacían de forma específica en mujeres, lo que señaló la existencia de rutas diferenciales en función del sexo (Figura 30B). Entre los genes sobreexpresados en el núcleo de la lesión exclusivamente en hombres se encontraron los genes de la familia HSP, mientras que la mayoría de los genes subexpresados correspondían a genes que codifican las proteínas ribosomales (RPLP1, RPS28, RPS14, RPL41, RPS19, RPL39 y RPS7). Por otro lado, algunos de los genes sobreexpresados solo en las mujeres también presentaban una expresión diferencial al comparar el núcleo de la lesión entre los dos sexos, como es el caso de CHIT1, S100A6, VIM, FAM107A, NUPR1 o AQP4. La expresión de estos genes sexo-específicos se muestra en la Figura A6 del Anexo.

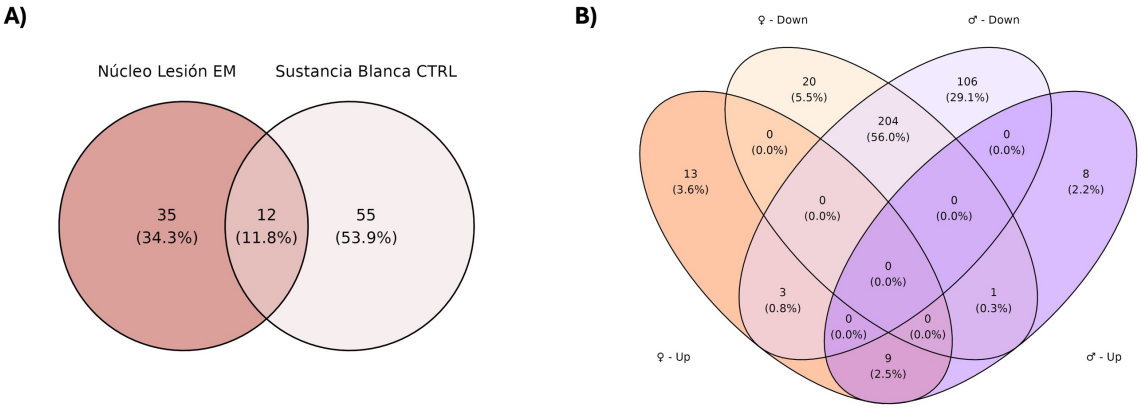

Figura 30. Diagrama de Venn de los genes diferencialmente expresados en el análisis de expresión diferencial. A) Genes diferencialmente expresados entre hombres y mujeres en el núcleo de la lesión de esclerosis múltiple (EM) y en la sustancia blanca sana de control (CTRL). Se observa que hay 12 genes que han sido identificados como significativos en ambas comparaciones. B) Genes sobreexpresados (*Up*) y subexpresados (*Down*) entre la sustancia blanca perilesional y el núcleo de la lesión de esclerosis múltiple, analizando por separado hombres (♂) y mujeres (♀).

5. Discusión

La organización espacial de los componentes celulares es crucial para el correcto funcionamiento de los sistemas biológicos, creándose microambientes a lo largo de los tejidos con funciones biológicas y composiciones celulares bien diferenciadas. Por ello, conocer no solo qué genes se expresan, sino también cuánto, cuándo y dónde lo hacen, ofrece una visión mucho más profunda de la organización y variabilidad de las distintas regiones tisulares, facilitando enormemente la comprensión de procesos fisiopatológicos complejos (126). El requerimiento de contexto espacial para comprender aspectos clave de la biología molecular ha impulsado el rápido desarrollo de las tecnologías de transcriptómica espacial (ST) en los últimos años. Gracias a la capacidad de la ST de estudiar los tipos celulares de un tejido en su entorno nativo al mismo tiempo que conserva la información espacial de sus perfiles de expresión, fue reconocida como "Método del Año" por la revista *Nature* en 2020 (127). Sin embargo, dada la novedad de este tipo de tecnología, aún presenta numerosos desafíos técnicos y computacionales.

El número de metodologías desarrolladas para el análisis de sus datos ha aumentado exponencialmente, derivando la mayoría de ellas de los algoritmos implementados originalmente para el análisis de datos de scRNA-Seq/snRNA-Seq. Sin embargo, la distribución y dispersión de los datos espaciales difiere de la de los datos unicelulares, y la asunción de que un *spot* es equivalente a una célula individual es conceptualmente incorrecta, especialmente en tecnologías como *Visium* donde cada *spot* puede contener múltiples células (87). Para superar estas limitaciones, se han desarrollado herramientas específicas para datos espaciales, como *Giotto*, una *suite* de análisis computacional diseñada específicamente para el procesamiento, análisis e integración de datos de ST. *Giotto* incluye un flujo de trabajo completo para el análisis de ST con algoritmos propios como BinSpect para la identificación de SVG, o SpatialDWLS para la deconvolución de tipos celulares, entre otros (69).

Se trata de un paquete altamente documentado, que dispone de una web propia con información sobre todas las funciones implementadas y con diversos tutoriales de cómo y cuándo utilizarlas. Sin embargo, prácticamente la totalidad de los tutoriales y ejemplos están destinados al análisis de una única muestra, careciendo de información sobre cómo tratar los estudios compuestos por más de una muestra para análisis más complejos como la deconvolución, la formación de redes espaciales o la detección de dominios espaciales.

El elevado número de argumentos y opciones que ofrecen sus funciones proporciona gran flexibilidad para adaptar los análisis a los datos concretos que se están analizando, pero a su vez incrementa significativamente la complejidad analítica asociada. Esta dificultad fue especialmente notoria en la identificación de los dominios espaciales, donde la función de inicialización del modelo HMRF admitía hasta 30 argumentos diferentes, cuyo cambio afectaba de forma importante a los resultados obtenidos. Aunque se pueden mantener los valores por defecto de todos estos parámetros, en ocasiones estos no son los más adecuados para el tipo de datos concretos que se están analizando, como se ha visto a lo largo de trabajo al ir comparando las diferentes opciones que ofrecía *Giotto* para cada paso del análisis. Esto se evidenció, por ejemplo, en el proceso de agrupamiento celular, donde la elección del tipo de red de vecinos más cercanos mostró claras diferencias en la definición de los grupos de Leiden. Mientras que el algoritmo sNN era el método por defecto para generar la red, los resultados obtenidos a partir de las redes kNN se adecuaban mejor al contexto espacial de las muestras.

Otro aspecto técnico relevante para considerar en el uso de *Giotto* son sus dependencias externas. Este paquete emplea internamente otros paquetes y funciones, tanto de R como de Python, lo que en ocasiones resulta en errores de ejecución derivados de incompatibilidades entre versiones que en su mayoría no se pueden solucionar sin modificar directamente el código fuente o el entorno de ejecución. Aunque los desarrolladores mantienen un foro activo en GitHub para reportar o resolver este tipo de incidencias, así como para consultas específicas de los usuarios, estas dificultades evidencian que los flujos de trabajo para ST se encuentran en constante actualización, reflejo del carácter emergente de esta tecnología.

En lo referente al desarrollo del presente trabajo, la selección de un conjunto de datos adecuado fue un paso crítico. Debido al objetivo propuesto consistente en caracterizar las diferencias de sexo en esclerosis múltiple (EM) a nivel transcriptómico con resolución espacial, se requería de un estudio que incluyera información explícita sobre el sexo del donante de cada sección de tejido, así como de un número suficiente de muestras de cada sexo y grupo experimental para poder realizar las comparaciones.

Sin embargo, a pesar de que las diferencias de sexo en la fisiopatología de la EM están bien documentadas (36,37), varios estudios encontrados no incluían esta información en sus metadatos o, si la incluían, no la consideraban en sus análisis o únicamente analizaban muestras procedentes de uno de los dos sexos, considerando que los hallazgos para mujeres eran equivalentes para hombres, y viceversa. Este aspecto motivó especialmente el enfoque de este trabajo, tratando de caracterizar a nivel molecular y espacial las diferencias de sexo en la enfermedad de EM.

Con este objetivo, se llevó a cabo un abordaje *in silico* a partir de los datos disponibles en una base de datos pública. En este contexto, se detectaron discrepancias entre la información disponible en la base de datos y la descrita en la publicación original, como la ausencia de algunas muestras control del estudio analizado. Este tipo de inconsistencias, junto con la falta de acceso público a los datos de algunos trabajos, limitación identificada durante la revisión sistemática, ponen de manifiesto la necesidad de fomentar el uso de datos que sigan los principios FAIR (del inglés *Findable, Accessible, Interoperable and Reusable*) (128), unas bases propuestas para estandarizar el formato de los datos depositados en los repositorios públicos. De este modo, se facilita que otros investigadores puedan explorar nuevos métodos, así como reanalizar los datos disponibles para contestar nuevas preguntas de investigación.

Durante el preprocesamiento de los datos, uno de los desafíos fue establecer los criterios de calidad óptimos para el filtrado de *spots* y genes. Los umbrales dependen intrínsicamente de los datos, influyendo directamente en la calidad de los análisis posteriores, ya que un filtrado demasiado estricto puede sesgar las conexiones espaciales, mientras que uno demasiado laxo puede introducir ruido. Por ello, es importante valorar diferentes combinaciones de los umbrales de criterios y seleccionar las que mejor se ajustan a la calidad de los datos que se están analizando. Además, el número de alternativas de cada etapa del análisis bioinformático fue considerable, lo que demuestra la necesidad de evaluarlas todas y escoger la opción más adecuada en función de las características del conjunto de datos.

Para la normalización se compararon tres metodologías distintas implementadas en *Giotto*, cada una de las cuales ofreció resultados diferentes. Con la log-normalización por tamaño de librería, método recomendado por *Giotto*, se conservó la variabilidad biológica entre las muestras y permitió realizar el análisis de expresión diferencial, a diferencia de la normalización basada en residuos de Pearson o la normalización TF-IDF seguida de la L2. No obstante, ningún método de normalización actual aborda el hecho de que un mismo gen puede variar su expresión en función de su localización espacial, lo que pone de manifiesto la necesidad de desarrollar nuevos algoritmos que consideren tanto la expresión de los genes como las coordenadas espaciales (78).

De forma similar, la elección del método de selección de HVG también condujo a resultados significativamente diferentes en la posterior identificación de grupos celulares. Mientras que el método de grupos de covarianza conducía a solapamientos entre tipos celulares distintos, los métodos basados en los residuos de Pearson o la regresión de LOESS suponían un agrupamiento preferente por condición experimental. Este aspecto demuestra cómo la decisión sobre el método a emplear influye directamente en la interpretación biológica de los resultados y pone de manifiesto, de nuevo, la necesidad de adaptar cada etapa del análisis al conjunto de datos.

La deconvolución de tipos celulares y el análisis de los patrones espaciales de expresión revelaron la heterogeneidad en la composición de los tipos celulares en función de la zona de la lesión de EM (núcleo, borde inflamado o sustancia blanca adyacente). De este modo, se observaron patrones espaciales ligados a la diversidad de los tipos celulares mayoritarios de cada región, superponiéndose a la organización tisular descrita en la literatura para las lesiones de EM (26–28). Además, no solo se anotaron tipos celulares propios del tejido cerebral, sino que en las muestras de EM también se detectaron células inmunes y vasculares. El deterioro de la BHE y la inflamación del parénquima podrían ser los responsables de la infiltración de este tipo de células en el tejido cerebral (129). Asimismo, se observó que las muestras correspondientes a tejido sano (CTRL) estaban compuestas principalmente por oligodendrocitos, lo que representa la arquitectura fisiológica de la sustancia blanca, donde este tipo de célula es abundante para asegurar la rápida y eficiente transmisión eléctrica (124).

Por otro lado, mediante un análisis de expresión diferencial se estudiaron las diferencias transcriptómicas entre el tejido cerebral sano y las lesiones de EM, así como entre hombres y mujeres para una región anatómica concreta. Independientemente de la variable sexo, las lesiones de EM se caracterizaron principalmente por la destrucción de oligodendrocitos y la activación de astrocitos, como consecuencia de la desmielinización; mientras que las regiones adyacentes sanas presentaron una mayor expresión de genes propios de los oligodendrocitos.

Finalmente, se detectaron diferencias de expresión génica entre hombres y mujeres, particularmente en la región correspondiente al núcleo de la lesión. Los hombres muestran, en general, una neurodegeneración más rápida y severa que las mujeres, lo que conduce a una mayor atrofia cerebral y al aumento de la pérdida neuronal (130). Estas evidencias concuerdan con los hallazgos de este trabajo, donde los hombres mostraron una sobreexpresión de genes asociados al daño y estrés celular. Entre ellos, se observaron algunos genes mitocondriales, cuyo aumento podría ser la consecuencia de los ataques autoinmunes a las vainas de mielina, así como genes que codifican las proteínas de la familia HSP (del inglés *Heat Shock Proteins*), las cuales se producen en respuesta a situaciones de estrés celular como podría ser la neuroinflamación (131).

Por el contrario, el perfil de genes diferencialmente expresados en las mujeres indicó una respuesta inflamatoria más exacerbada, posiblemente mediada por microglía y astrocitos reactivos. Asimismo, en este grupo también se observó una mayor expresión de genes implicados en la remodelación del citoesqueleto, en concordancia con algunos estudios previos que sugieren que la remielinización tiene mayor efectividad en las mujeres que en los hombres (132).

Por tanto, todos estos hallazgos pueden correlacionarse con las diferencias a nivel de susceptibilidad y progresión de la enfermedad descritas en la literatura, reforzando además la importancia de considerar el sexo como una variable biológica relevante en los estudios biomédicos, especialmente de enfermedades neurodegenerativas como la EM, tanto para una mejor caracterización de las enfermedades humanas como para desarrollar terapias personalizadas.

6. Conclusiones

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

1. La transcriptómica espacial es una tecnología capaz de dilucidar las alteraciones entre un estado homeostático y uno patológico subyacente a la progresión de una enfermedad con resolución espacial, enfoque que no proporcionan el resto de las tecnologías transcriptómicas existentes.
2. El resultado de la revisión sistemática ha puesto de manifiesto la necesidad de incluir información sobre el sexo de los donantes en los estudios de esclerosis múltiple para poder caracterizar las diferencias de sexo en esta enfermedad y avanzar hacia una medicina personalizada.
3. Las múltiples alternativas para realizar cada paso del análisis bioinformático revelan la falta de estandarización en el procesamiento de los datos de transcriptómica espacial.
4. Los diferentes resultados obtenidos según la metodología empleada demuestran la necesidad de adaptar cada paso del análisis a los datos específicos con los que se está trabajando.
5. Los tipos celulares del cerebro presentan perfiles de expresión específicos que siguen además un patrón espacial determinado por los tipos celulares mayoritarios de cada región tisular.
6. Se han identificado patrones de expresión diferenciales entre sexos en la desmielinización del núcleo de la lesión de esclerosis múltiple, con respuestas inflamatorias más exacerbadas en mujeres y mayor muerte celular en hombres.
7. El presente trabajo ha permitido identificar las diferencias de sexo en esclerosis múltiple con resolución espacial a partir de unos datos generados con anterioridad, lo que realza la potencia y el valor científico de este tipo de aproximaciones *in silico*.

7. Perspectivas futuras

El presente trabajo proporciona un flujo de trabajo completo para analizar datos transcriptómicos espaciales empleando el paquete *Giotto*, que en última instancia pretende caracterizar las diferencias de sexo en EM. No obstante, este abordaje representa una línea de investigación que puede extenderse mediante múltiples estrategias que mejoren y complementen los hallazgos descritos.

La EM se trata de una enfermedad compleja, donde no solamente afecta el sexo, sino también otras variables como la edad, el tipo de lesión o la duración de la enfermedad, entre otras. Por ello, la inclusión de estas covariables en el modelo de expresión diferencial permitiría obtener resultados mucho más precisos.

Asimismo, la integración de varios estudios a través de un enfoque de metaanálisis permitiría incluir una mayor diversidad de muestras, mejorando así la generalización de los resultados obtenidos. Además, gracias a la capacidad de *Giotto* para integrar datos multiómicos espaciales, la combinación de ST con otras capas moleculares, como la proteómica espacial, ofrecería una visión más completa de los mecanismos patológicos subyacentes a la EM.

La caracterización funcional de los resultados obtenidos en la expresión diferencial, con aproximaciones como el análisis de sobrerrepresentación (ORA, por sus siglas en inglés) o el análisis de enriquecimiento de conjuntos de genes (GSEA, por sus siglas en inglés), permitiría relacionar los genes diferencialmente expresados con procesos biológicos o celulares alterados en la EM.

Finalmente, los resultados computacionales derivados de este estudio podrían ser validados funcional y experimentalmente, mediante técnicas como smFISH (técnica de ST basada en imagen) o a través de modelos animales o cultivos *ex vivo*.

8. Bibliografía

1. Li S jie, Wu Y li, Chen J hua, Shen S yi, Duan J, Xu HE. Autoimmune diseases: targets, biology, and drug discovery. Acta Pharmacol Sin. 2024 Apr 14;45(4):674–85.
2. Theofilopoulos AN, Kono DH, Baccala R. The multiple pathways to autoimmunity. Nat Immunol. 2017 Jun 20;18(7):716–24.
3. Raval P. Systemic (non-Organ Specific) Autoimmune Disorders. In: Enna SJ, Bylund DB, editors. xPharm: The Comprehensive Pharmacology Reference. New York: Elsevier; 2007. p. 1–6.
4. Rezaei N, Yazdanpanah N. Autoimmune diseases in different organs. In: Rezaei N, editor. Translational Autoimmunity. 1st ed. Academic Press; 2022. p. 1–13.
5. Kono DH, Theofilopoulos AN. Autoimmunity. In: Firestein GS, Budd RC, Gabriel SE, McInnes IB, O'Dell JR, editors. Kelley and Firestein's Textbook of Rheumatology (Tenth Edition). Tenth Edition. Elsevier; 2017. p. 301-317.e5.
6. Pisetsky DS. Pathogenesis of autoimmune disease. Nat Rev Nephrol. 2023 Aug;19(8):509–24.
7. Lassmann H, Brück W, Lucchinetti CF. The Immunopathology of Multiple Sclerosis: An Overview. Brain Pathology. 2007 Apr 2;17(2):210–8.
8. Haider L, Zrzavy T, Hametner S, Höftberger R, Bagnato F, Grabner G, et al. The topograpy of demyelination and neurodegeneration in the multiple sclerosis brain. Brain. 2016 Mar;139(3):807–15.
9. Pericot I, Montalban X. Esclerosis múltiple. Medicina Integral. 2001;38(1):18–24.
10. Multiple Sclerosis International Federation (MSIF). Number of people with MS. 2023 [cited 2025 May 7]. Atlas of MS. Available from: https://www.atlasofms.org/map/spain/epidemiology/number-of-people-with-ms
11. Steinmetz JD, Seeher KM, Schiess N, Nichols E, Cao B, Servili C, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol. 2024 Apr;23(4):344–81.
12. Jakimovski D, Bittner S, Zivadinov R, Morrow SA, Benedict RHB, Zipp F, et al. Multiple sclerosis. The Lancet. 2024 Jan;403(10422):183–202.
13. Ward M, Goldman MD. Epidemiology and Pathophysiology of Multiple Sclerosis. CONTINUUM: Lifelong Learning in Neurology. 2022 Aug;28(4):988–1005.
14. Patsopoulos NA, Baranzini SE, Santaniello A, Shoostari P, Cotsapas C, Wong G, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science (1979). 2019 Sep 27;365(6460).
15. Baranzini SE, Oksenberg JR. The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. Trends in Genetics. 2017 Dec;33(12):960–70.
16. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med. 2007 Aug 30;357(9):851–62.
17. Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. The Lancet. 2018 Apr;391(10130):1622–36.
18. Tizaoui K. Multiple sclerosis genetics: Results from meta-analyses of candidate-gene association studies. Cytokine. 2018 Jun;106:154–64.

19. Simpson S, Wang W, Otahal P, Blizzard L, van der Mei IAF, Taylor B V. Latitude continues to be significantly associated with the prevalence of multiple sclerosis: an updated meta-analysis. J Neurol Neurosurg Psychiatry. 2019 Nov;90(11):1193–200.
20. Bjornevik K, Cortese M, Healy BC, Kuhle J, Mina MJ, Leng Y, et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science (1979). 2022 Jan 21;375(6578):296–301.
21. Lanz T V, Brewer RC, Ho PP, Moon JS, Jude KM, Fernandez D, et al. Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. Nature. 2022 Mar;603(7900):321–7.
22. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017 Jan 9;13(1):25–36.
23. Bar-Or A, Li R. Cellular immunology of relapsing multiple sclerosis: interactions, checks, and balances. Lancet Neurol. 2021 Jun;20(6):470–83.
24. Haki M, AL-Biati HA, Al-Tameemi ZS, Ali IS, Al-hussaniy HA. Review of multiple sclerosis: Epidemiology, etiology, pathophysiology, and treatment. Medicine. 2024 Feb 23;103(8):e37297.
25. van Langelaar J, Rijvers L, Smolders J, van Luijn MM. B and T Cells Driving Multiple Sclerosis: Identity, Mechanisms and Potential Triggers. Front Immunol. 2020 May 8;11.
26. Schirmer L, Velmeshev D, Holmqvist S, Kaufmann M, Werneburg S, Jung D, et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. Nature. 2019 Sep 17;573(7772):75–82.
27. Morgan BP, Gommerman JL, Ramaglia V. An "Outside-In" and "Inside-Out" Consideration of Complement in the Multiple Sclerosis Brain: Lessons From Development and Neurodegenerative Diseases. Front Cell Neurosci. 2021 Jan 7;14.
28. Lassmann H. Multiple Sclerosis Pathology. Cold Spring Harb Perspect Med. 2018 Mar 1;8(3):a028936.
29. Ford H. Clinical presentation and diagnosis of multiple sclerosis. Clinical Medicine. 2020 Jul;20(4):380–3.
30. Klineova S, Lublin FD. Clinical Course of Multiple Sclerosis. Cold Spring Harb Perspect Med. 2018 Sep 4;8(9).
31. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. 2018 Feb;17(2):162–73.
32. Claflin SB, Broadley S, Taylor B V. The Effect of Disease Modifying Therapies on Disability Progression in Multiple Sclerosis: A Systematic Overview of Meta-Analyses. Front Neurol. 2019 Jan 10;9.
33. DeLuca J, Chiaravalloti ND, Sandroff BM. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. Nat Rev Neurol. 2020 Jun 5;16(6):319–32.
34. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis. Neurology. 1996 Apr;46(4):907–11.
35. Lublin FD, Reingold SC, Cohen JA, Cutter GR, Sørensen PS, Thompson AJ, et al. Defining the clinical course of multiple sclerosis. Neurology. 2014 Jul 15;83(3):278–86.
36. Voskuhl RR. The effect of sex on multiple sclerosis risk and disease progression. Multiple Sclerosis Journal. 2020 Apr 22;26(5):554–60.

37. Ryan L, Mills KHG. Sex differences regulate immune responses in experimental autoimmune encephalomyelitis and multiple sclerosis. Eur J Immunol. 2022 Jan 13;52(1):24–33.
38. Voskuhl RR, Sawalha AH, Itoh Y. Sex chromosome contributions to sex differences in multiple sclerosis susceptibility and progression. Multiple Sclerosis Journal. 2018 Jan 8;24(1):22–31.
39. MacKenzie-Graham A, Brook J, Kurth F, Itoh Y, Meyer C, Montag MJ, et al. Estriol-mediated neuroprotection in multiple sclerosis localized by voxel-based morphometry. Brain Behav. 2018 Sep 24;8(9).
40. Bove R, Musallam A, Healy B, Raghavan K, Glanz B, Bakshi R, et al. Low testosterone is associated with disability in men with multiple sclerosis. Multiple Sclerosis Journal. 2014 Oct 7;20(12):1584–92.
41. Raja K, Patrick M, Gao Y, Madu D, Yang Y, Tsoi LC. A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. Int J Genomics. 2017;2017:1–10.
42. Vailati-Riboni M, Palombo V, Loor JJ. What Are Omics Sciences? In: Ametaj BN, editor. Periparturient Diseases of Dairy Cows: A Systems Biology Approach. Cham: Springer International Publishing; 2017. p. 1–7.
43. Khodadadian A, Darzi S, Haghi-Daredeh S, sadat Eshaghi F, Babakhanzadeh E, Mirabutalebi SH, et al. Genomics and Transcriptomics: The Powerful Technologies in Precision Medicine. Int J Gen Med. 2020 Sep;Volume 13:627–40.
44. Xu C, Shao J. High-throughput omics technologies in inflammatory bowel disease. Clinica Chimica Acta. 2024 Mar;555:117828.
45. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci. 2017 Jul 29;18(8):1652.
46. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 18;13(5):e1005457.
47. Dai X, Shen L. Advances and Trends in Omics Technology Development. Front Med (Lausanne). 2022;9:911861.
48. Conesa Cegarra A. Transcriptómica computacional: La biología y la ingeniería van de la mano en un viaje alucinante por la vida a nivel molecular. In: Real Academia de Ingeniería, editor. Real Academia de Ingeniería; 2022 [cited 2025 May 14]. Available from: https://raing.es/pdf/publicaciones/discursos_de_ingreso/transcriptomica_computacional.pdf
49. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021 Dec 15;13(1):36.
50. Cha J, Lee I. Single-cell network biology for resolving cellular heterogeneity in human diseases. Exp Mol Med. 2020 Nov 26;52(11):1798–808.
51. Walter TJ, Suter RK, Ayad NG. An overview of human single-cell RNA sequencing studies in neurobiological disease. Neurobiol Dis. 2023 Aug;184:106201.
52. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One. 2018 Dec 26;13(12):e0209648.
53. Kersey HN, Acri DJ, Dabin LC, Hartigan K, Mustaklem R, Park JH, et al. Comparative analysis of nuclei isolation methods for brain single-nucleus RNA sequencing. bioRxiv. 2025 Mar 26;

54. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020 Jun 1;38(6):737–46.
55. Maitra M, Nagy C, Chawla A, Wang YC, Nascimento C, Suderman M, et al. Extraction of nuclei from archived postmortem tissues for single-nucleus sequencing applications. Nat Protoc. 2021 Jun 10;16(6):2788–801.
56. Grindberg R V., Yee-Greenbaum JL, McConnell MJ, Novotny M, O'Shaughnessy AL, Lambert GM, et al. RNA-sequencing from single nuclei. Proceedings of the National Academy of Sciences. 2013 Dec 3;110(49):19802–7.
57. He J, Lin L, Chen J. Practical bioinformatics pipelines for single-cell RNA-seq data analysis. Biophys Rep. 2022 Jun 30;8(3):158–69.
58. Jung N, Kim TK. Spatial transcriptomics in neuroscience. Exp Mol Med. 2023 Oct;55(10):2105–15.
59. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021 Aug 12;596(7871):211–20.
60. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science (1979). 2016 Jul;353(6294):78–82.
61. Wang Y, Liu B, Zhao G, Lee Y, Buzdin A, Mu X, et al. Spatial transcriptomics: Technologies, applications and experimental considerations. Genomics. 2023 Sep;115(5):110671.
62. Lim HJ, Wang Y, Buzdin A, Li X. A practical guide for choosing an optimal spatial transcriptomics technology from seven major commercially available options. BMC Genomics. 2025 Jan 20;26(1):47.
63. Yue L, Liu F, Hu J, Yang P, Wang Y, Dong J, et al. A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. Comput Struct Biotechnol J. 2023;21:940–55.
64. Isnard P, Humphreys BD. Spatial Transcriptomics. Am J Pathol. 2025 Jan;195(1):23–39.
65. BMKGENE. 10x Genomics Visium Spatial Transcriptome [Internet]. 2025 [cited 2025 May 17]. Available from: https://www.bmkgene.com/10x-genomics-visium-spatial-transcriptome-product/
66. 10x Genomics. Visium Spatial Gene Expression Reagent Kits - User Guide [Internet]. 2022 Jan [cited 2025 May 17]. Report No.: CG000239. Available from: https://cdn.10xgenomics.com/image/upload/v1660261286/support-documents/CG000239_Visium_Spatial_Gene_Expression_User_Guide_Rev_F.pdf
67. Liu B, Li Y, Zhang L. Analysis and Visualization of Spatial Transcriptomic Data. Front Genet. 2022 Jan 27;12.
68. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024 Feb 25;42(2):293–304.
69. Dries R, Zhu Q, Dong R, Eng CHL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 2021 Dec 8;22(1):78.
70. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018 Dec 6;19(1):15.
71. Roselli CE. Neurobiology of gender identity and sexual orientation. J Neuroendocrinol. 2018 Jul 11;30(7).

72. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2024. Available from: https://www.R-project.org/
73. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul 21;6(7):e1000097.
74. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;n71.
75. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207–10.
76. Li Y, Dennis S, Hutch MR, Ding Y, Zhou Y, Li Y, et al. SOAR elucidates disease mechanisms and empowers drug discovery through spatial transcriptomics. 2022.
77. Xu Z, Wang W, Yang T, Li L, Ma X, Chen J, et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. Nucleic Acids Res. 2024 Jan 5;52(D1):D1053–61.
78. Fang S, Chen B, Zhang Y, Sun H, Liu L, Liu S, et al. Computational Approaches and Challenges in Spatial Transcriptomics. Genomics Proteomics Bioinformatics. 2023 Feb;21(1):24–47.
79. Del Rossi N, Chen JG, Yuan GC, Dries R. Analyzing Spatial Transcriptomics Data Using Giotto. Curr Protoc. 2022 Apr;2(4):e405.
80. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021 Sep 6;22(1):258.
81. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 1972 Jan;28(1):11–21.
82. Maliamanis T V., Papakostas GA. Machine learning vulnerability in medical imaging. Machine Learning, Big Data, and IoT for Medical Informatics. 2021 Jan 1;53–70.
83. Finak G, Perez JM, Weng A, Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics. 2010 Dec 4;11(1):546.
84. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018 Nov 30;15(11):932–5.
85. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185–93.
86. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020 Feb 2;17(2):137–45.
87. Weber LM. Orchestrating Spatial Transcriptomics Analysis with Bioconductor [Internet]. 2025 [cited 2025 May 31]. Available from: https://lmweber.org/OSTA/
88. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746.
89. Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun. 2022 Nov 23;13(1):1–22.
90. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics. 2015 Feb 15;31(4):545–54.

91. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019 Jan 3;37(1):38–44.
92. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018 Sep 2;3(29):861.
93. Yuan Z, Zhao F, Lin S, Zhao Y, Yao J, Cui Y, et al. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. Nat Methods. 2024 Mar 15;21(4):712–22.
94. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Nov 18;16(12):1289–96.
95. Korsunsky I, Hemberg M, Patikas N, Yao H, Millard N, Fan J, et al. harmony: Fast, Sensitive, and Accurate Integration of Single Cell Data [Internet]. CRAN: Contributed Packages. 2021 [cited 2025 Jun 5]. Available from: https://CRAN.R-project.org/package=harmony
96. Halder RK, Uddin MN, Uddin MA, Aryal S, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. J Big Data. 2024 Aug 11;11(1):1–55.
97. Schmidt ST, Akhave N, Knightly RE, Reuben A, Vokes N, Zhang J, et al. Shared Nearest Neighbors Approach and Interactive Browser for Network Analysis of a Comprehensive Non-Small-Cell Lung Cancer Data Set. JCO Clin Cancer Inform. 2022 Jul;6(6).
98. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat. 1979;28(1):100.
99. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif. 2014 Oct 18;31(3):274–95.
100. Pons Pascal and Latapy M. Computing Communities in Large Networks Using Random Walks. In: Yolum pInarand G, Gürgen T, Özturan Can F, editors. Computer and Information Sciences - ISCIS 2005. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 284–93.
101. Ertöz L, Steinbach M, Kumar V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: Proceedings of the 2003 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003. p. 47–58.
102. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008 Oct 1;2008(10):P10008.
103. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):1–12.
104. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015 Dec 10;16(1):278.
105. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016 Dec 1;17(1):144.
106. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016 Oct 31;5:2122.

107. Zappia L, Lun A. zellkonverter: Conversion Between scRNA-seq Objects [Internet]. 2024. Available from: https://bioconductor.org/packages/zellkonverter
108. Hong R, Koga Y, Bandyadka S, Leshchyk A, Wang Y, Akavoor V, et al. Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data. Nat Commun. 2022 Mar 30;13(1):1688.
109. Schaar A, Heumos L, Zappia L, single-cell best practices consortium. Quality Control. 2023 [cited 2025 May 23]. Single-cell best practices. Available from: https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html
110. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017 Apr 15;33(8):1179–86.
111. Germain PL, Lun A, Garcia Meixide C, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. 2022 May 16;10:979.
112. Dong R, Yuan GC. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. 2021 Dec 10;22(1):145.
113. Kim SY, Volsky DJ. PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics. 2005 Jun 8;6(1):144.
114. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan GC. Accurate estimation of cell-type composition from gene expression data. Nat Commun. 2019 Jul 5;10(1):2975.
115. Lee DT, Schachter BJ. Two algorithms for constructing a Delaunay triangulation. International Journal of Computer & Information Sciences. 1980 Jun;9(3):219–42.
116. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018 May 19;15(5):343–6.
117. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods. 2018 Mar 19;15(5):339–42.
118. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020 Feb 27;17(2):193–200.
119. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995 Jan 1;57(1):289–300.
120. Zhu Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat Biotechnol. 2018 Dec 29;36(12):1183–90.
121. Lerma-Martin C, Badia-i-Mompel P, Ramirez Flores RO, Sekol P, Schäfer PSL, Riedl CJ, et al. Cell type mapping reveals tissue niches and interactions in subcortical multiple sclerosis lesions. Nat Neurosci. 2024 Dec 5;27(12):2354–65.
122. Marsh-Wakefield F, Ashhurst T, Trend S, McGuire HM, Juillard P, Zinger A, et al. IgG3 + B cells are associated with the development of multiple sclerosis. Clin Transl Immunology. 2020 May 1;9(5):e1133.
123. Saade M, Araujo de Souza G, Scavone C, Kinoshita PF. The Role of GPNMB in Inflammation. Front Immunol. 2021 May 12;12:674739.
124. Emery B. Regulation of oligodendrocyte differentiation and myelination. Science (1979). 2010 Nov 5;330(6005):779–82.

125. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics. 2016 Jun 20;54:1.30.1-1.30.33.
126. Wang Y, Liu B, Zhao G, Lee Y, Buzdin A, Mu X, et al. Spatial transcriptomics: Technologies, applications and experimental considerations. Genomics. 2023 Sep;115(5):110671.
127. Method of the Year 2020: spatially resolved transcriptomics. Nat Methods. 2021 Jan 6;18(1):1–1.
128. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data . 2016 Mar 15;3(1):1–9.
129. Mohan H, Krumbholz M, Sharma R, Eisele S, Junker A, Sixt M, et al. Extracellular matrix in multiple sclerosis lesions: Fibrillar collagens, biglycan and decorin are upregulated and associated with infiltrating immune cells. Brain pathology. 2010 Sep;20(5):966–75.
130. Voskuhl RR, Patel K, Paul F, Gold SM, Scheel M, Kuchling J, et al. Sex differences in brain atrophy in multiple sclerosis. Biol Sex Differ. 2020 Aug 28;11(1):49.
131. Georgopoulos C, McFarland H. Heat shock proteins in multiple sclerosis and other autoimmune diseases. Immunol Today. 1993 Aug;14(8):373–5.
132. Li WW, Penderis J, Zhao C, Schumacher M, Franklin RJM. Females remyelinate more efficiently than males following demyelination in the aged but not young adult CNS. Exp Neurol. 2006 Nov;202(1):250–4.

ANEXO

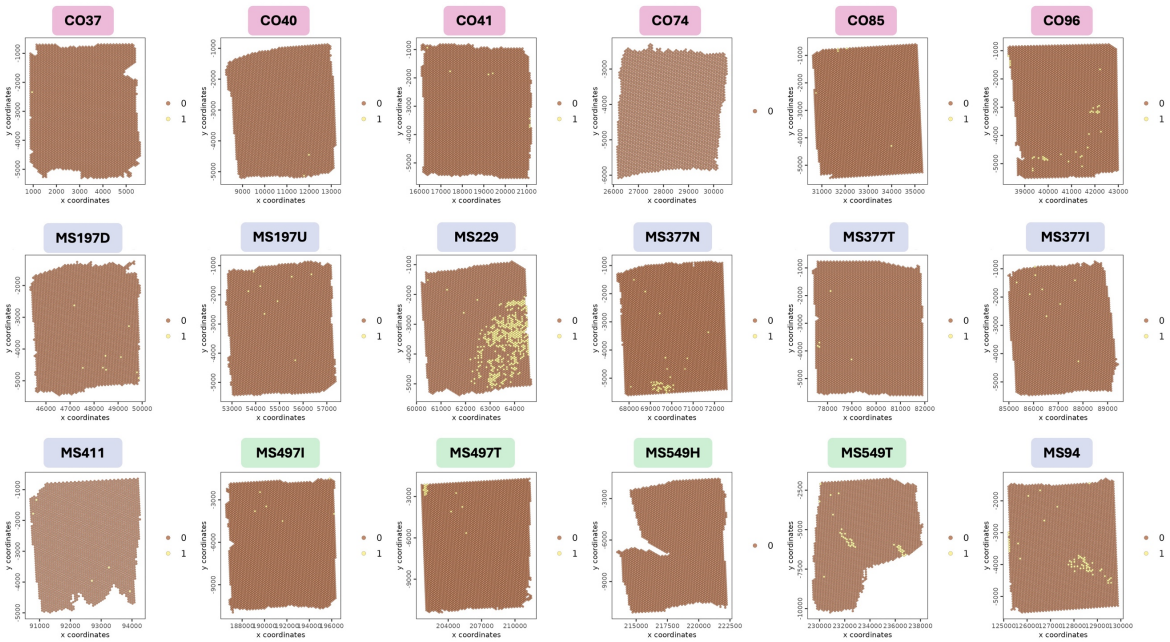

Figura A1. Evaluación espacial de la calidad de los datos de transcriptómica espacial. Se muestra la localización de los *spots* con menos de 100 genes detectados (en amarillo) para cada una de las muestras. Estos *spots* no superaron el umbral de calidad establecido, por lo que fueron eliminados del conjunto de datos. Las muestras control se resaltan con marcos en rosa, las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos en color azul y las de lesiones crónicas inactivas de esclerosis múltiple se resaltan con marcos verdes.

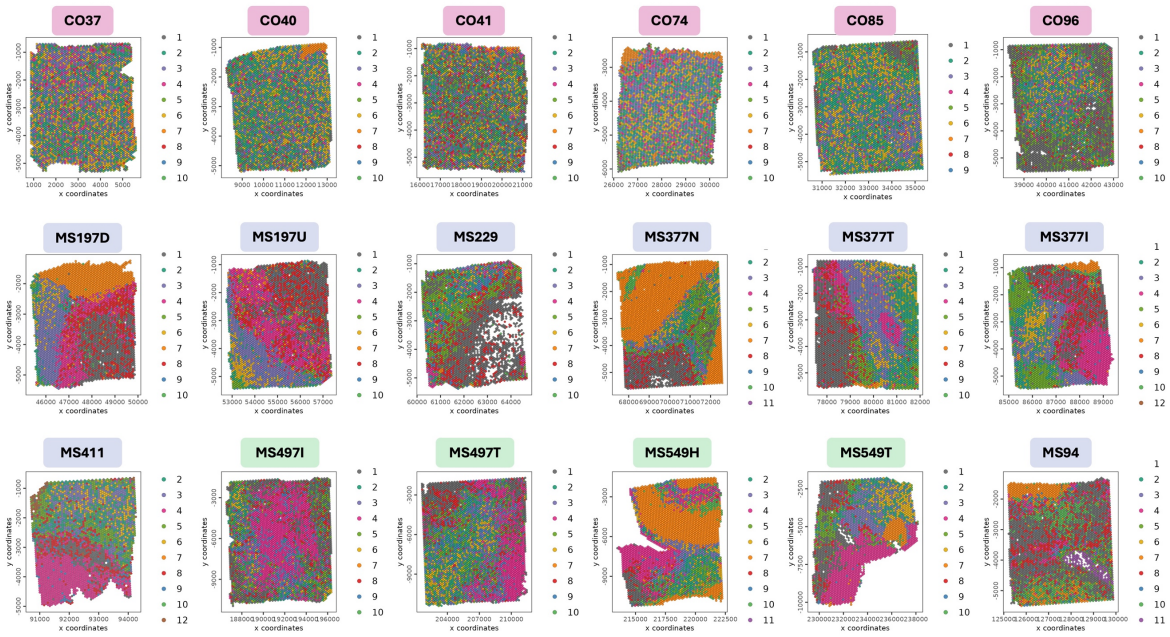

Figura A2. Visualización espacial de los grupos de Leiden obtenidos. Los grupos fueron definidos a partir de una red de *k*-vecinos más cercanos y con una resolución de 0,50. Las muestras control se resaltan con marcos rosas, las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos azules y las muestras de lesiones crónicas inactivas de esclerosis múltiple se resaltan con marcos verdes. Cada punto representa un *spot*, coloreado según su asignación a los distintos grupos de Leiden. Los puntos de color blanco son los *spots* filtrados por tener menos de 100 genes detectados.

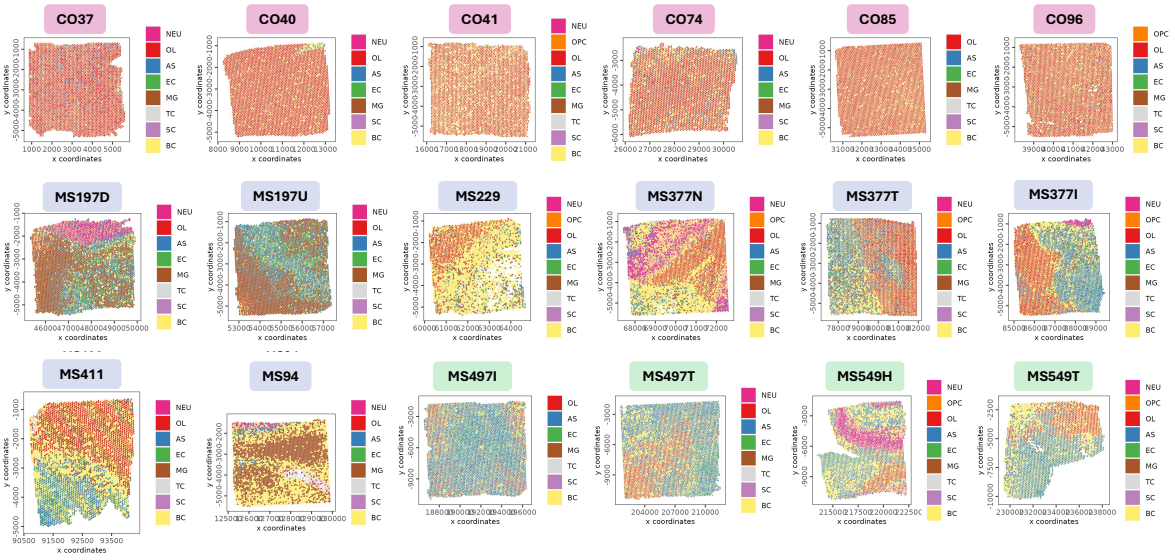

Figura A3. Visualización espacial de los resultados de la deconvolución de tipos celulares. Cada *spot* está representando por un gráfico de sectores que muestra el porcentaje de cada tipo celular anotado. Las muestras control se resaltan con marcos de color rosa, las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos de color azul y las muestras de lesiones crónicas inactivas de esclerosis múltiple se resaltan con marcos verdes. AS: astrocitos. BC: Células B. EC: células endoteliales. MG: microglía. NEU: neuronas. OL: oligodendrocitos. OPC: células progenitoras de oligodendrocitos. SC: células del estroma. TC: células T.

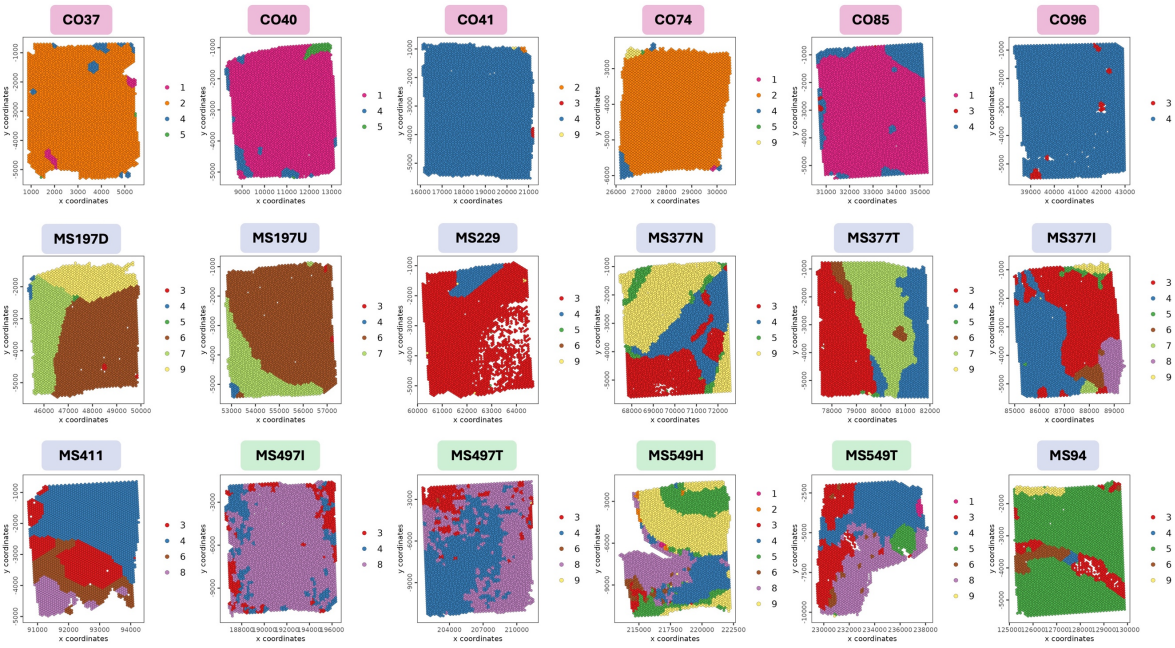

Figura A4. Visualización espacial de los dominios espaciales identificados con el modelo de Campo Aleatorio Oculto de Markov (HMRF) para $\beta = 15$. El parámetro β regula la fuerza de interacción entre *spots*, de manera que valores mayores dan más peso a la información espacial en el refinamiento del agrupamiento respecto al perfil de expresión. Cada punto representa un *spot*, y los colores indican su asignación a los distintos dominios espaciales identificados. Las muestras control se resaltan con marcos rosas, las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos azules y las muestras de lesiones crónicas inactivas de esclerosis múltiple se resaltan con marcos verdes.

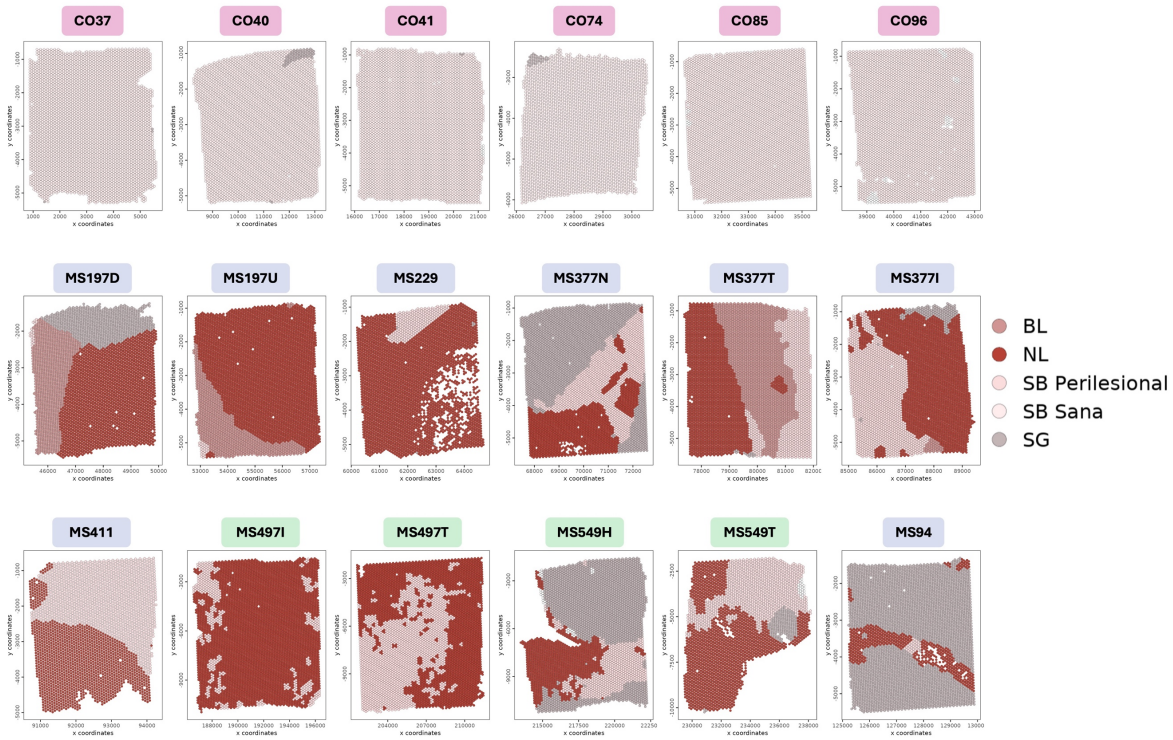

Figura A5. Visualización espacial de la anotación manual de los dominios identificados con el modelo de Campo Aleatorio Oculto de Markov (HMRF) para $\beta = 15$. El color de cada *spot* indica su asignación a cada una de las áreas tisulares anotadas. Las muestras control se resaltan con marcos de color rosa, las muestras de lesiones crónicas activas de esclerosis múltiple se resaltan con marcos azules y las muestras de lesiones crónicas inactivas de esclerosis múltiple se resaltan con marcos verdes. BL: borde de la lesión de esclerosis múltiple. NL: núcleo de la lesión de esclerosis múltiple. SB Perilesional: sustancia blanca perilesional (es decir, adyacente a las lesión) de las muestras de esclerosis múltiple. SB Sana: sustancia blanca sana de las muestras control. SG: sustancia gris capturada al seccionar el tejido cerebral.

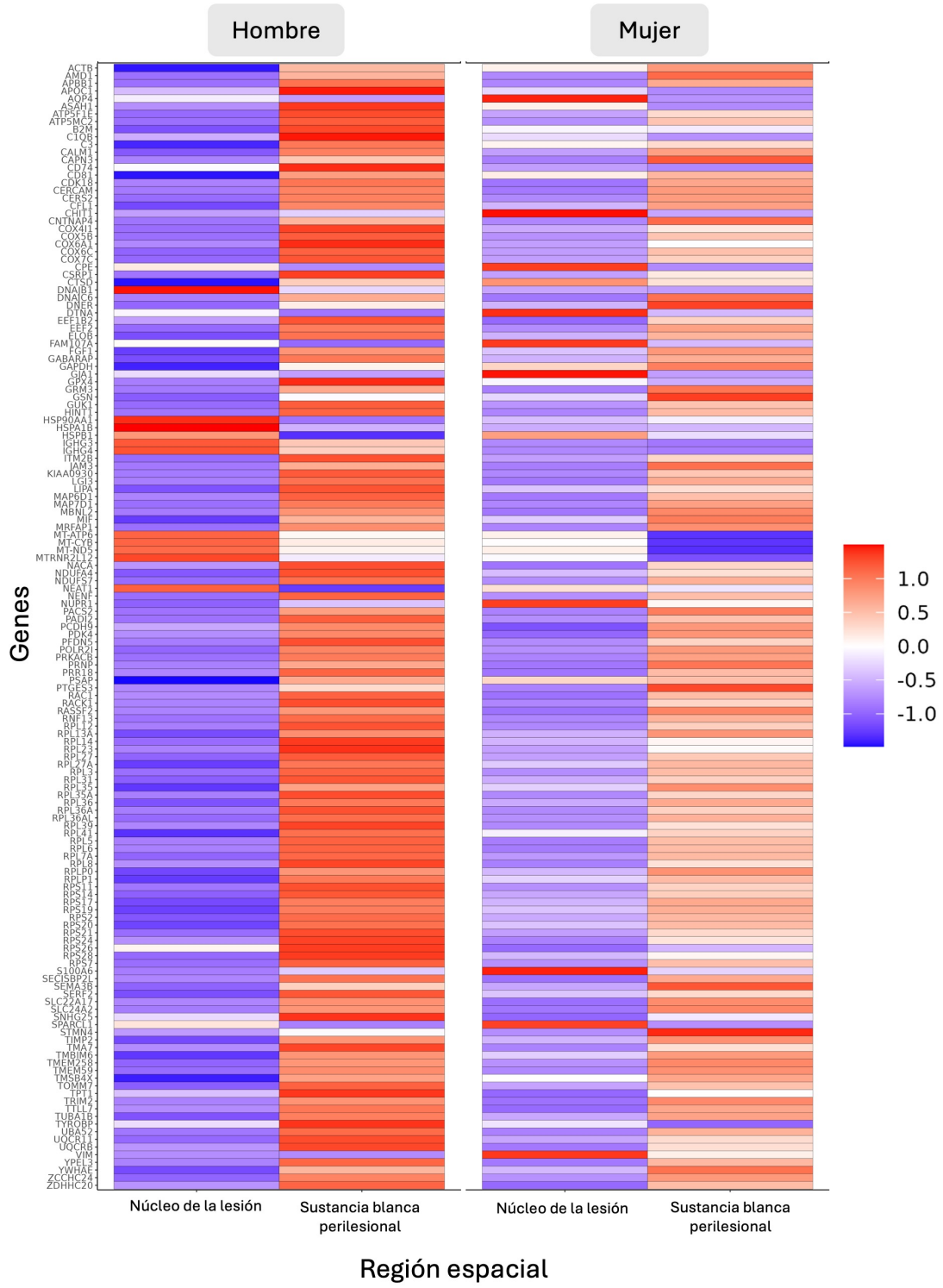

Figura A6. Mapa de calor con la expresión normalizada de los genes diferencialmente expresados entre el núcleo de la lesión de esclerosis múltiple y la sustancia blanca perilesional solo en uno de los dos sexos. Para amplificar visualmente las diferencias, la expresión de cada gen se transformó en un *z-score* a partir de la media y la desviación estándar de la expresión de cada grupo comparado. Los tonos rojos reflejan la sobreexpresión del gen en ese grupo, y los tonos azules indican subexpresión del gen.