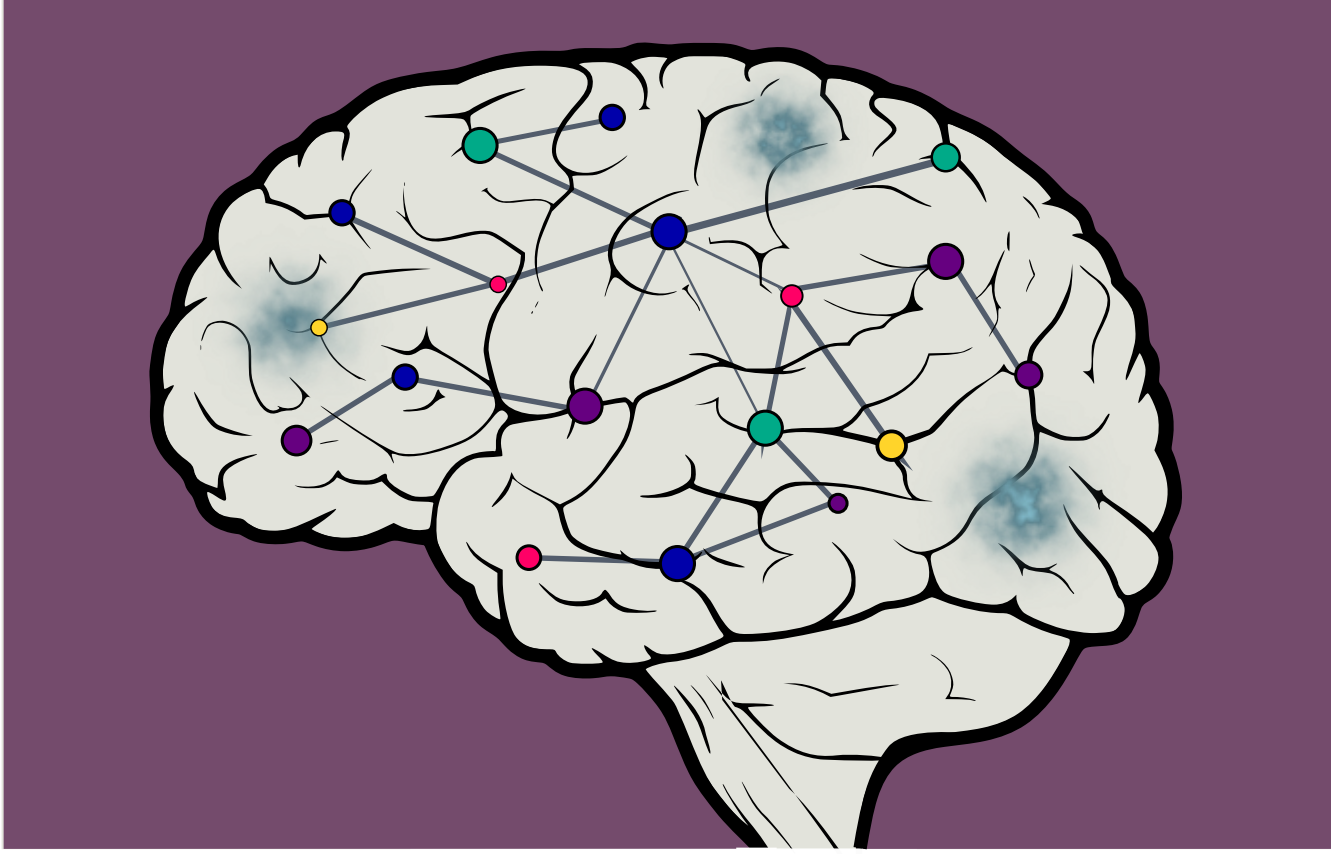


Sex differences and cross-disease molecular mechanisms in multiple sclerosis: insights from transcriptomics and metagenomics analyses

Irene Soler Sáez

Supervisors
Dr. Francisco García García
Dr. Marta R. Hidalgo García
Dr. María De La Iglesia Vayá

Valencia, October 2025

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VALENCIA BIOMEDICAL RESEARCH FOUNDATION
CENTRO DE INVESTIGACIÓN PRÍNCIPE FELIPE

Computational Biomedicine Laboratory

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

PhD in Biotechnology

International PhD mention

Sex differences and cross-disease molecular mechanisms in multiple sclerosis: insights from transcriptomics and metagenomics analyses

Valencia, October 2025

Author: Irene Soler Sáez

Supervisors: Dr. Francisco García García

Dr. Marta R. Hidalgo García

Dr. María de la Iglesia Vayá

Tutor: Máximo Ibo Galindo Orozco

A mi madre

ACKNOWLEDGEMENTS

Echando la vista atrás quiero agradecer a todas las personas que han formado parte de este camino. En primer lugar, quiero dar las gracias a mis directores de tesis Paco, Marta y Mariam. Muchas gracias por vuestra energía positiva y vuestra calidad humana. Gracias por vuestro apoyo constante, por la confianza depositada en mí y por vuestras orientaciones a lo largo de estos años.

A todas las personas que forman o han formado parte de la CBL, gracias por compartir no solo el espacio de trabajo, sino también tantas risas y buenos momentos. Soy muy afortunada de haber compartido estos años con vosotros. Al inicio fuimos compañeros de trabajo, ahora puedo decir que me llevo muchos amigos de esta aventura.

Quiero dar las gracias también a todas esas personas que han aparecido en este camino. En especial, a Sara, por abrirme las puertas de su grupo y hacer posible mi estancia de investigación. A todo su equipo y las personas que conocí en Alemania, gracias por acogerme como a una más y hacerme sentirme como en casa. A los investigadores e investigadoras que he conocido, por compartir su experiencia y enseñarme nuevas formas de mirar la ciencia. A la comunidad de RSG-Spain, es genial formar parte de este grupo.

Finalmente, a mi familia y a mis amigos de dentro y fuera del mundo de la ciencia. Aunque muchos de vosotros no tengáis muy claro qué hago delante del ordenador tantas horas al día, siempre me habéis apoyado como los que más. Gracias por estar siempre a mi lado.

A la persona que está leyendo los agradecimientos ahora: quizá haya pasado tiempo desde que defendí esta tesis, pero si estás leyendo esto probablemente te encuentres en la misma situación. Mucho ánimo y disfruta al máximo de esta experiencia.

This doctoral dissertation was conducted at the Computational Biomedicine Laboratory, led by Dr. Francisco García García and located at the Príncipe Felipe Research Center (CIPF) in Valencia, Spain. The development of this work was made possible thanks to the funding provided by the *Formación de Profesorado Universitario* program (FPU20/03544) of the Spanish *Ministerio de Ciencia, Innovación y Universidades*. Moreover, between September and November 2024, the doctoral candidate undertook a research stay at the Institute of Medical Microbiology and Hygiene in the University Medical Center of Johannes Gutenberg University in Mainz, Germany. This stay was conducted in the Mucosal Microbiology and Immunology laboratory led by Dr. Sara Vieira-Silva, and was supported by the EMBO Scientific Exchange Grant (SEG10845).

ABSTRACT

The central nervous system (CNS) integrates motor, sensory, and autonomic functions. Its dysfunction gives rise to a diverse range of neurological disorders, whose onset and progression are influenced by multiple factors, including intrinsic host features. This thesis aims to contribute to the molecular characterization of CNS-related diseases through two complementary perspectives: identifying sex-differential mechanisms in multiple sclerosis (MS) and examining how the characterization of MS, together with other neurodegenerative diseases, provide insights into the neurobiology of brain tumors.

The thesis is centered on the characterization of sex differences in MS through the analysis of single cell transcriptomics and metagenomics data. MS is a chronic autoimmune and neurodegenerative disease that typically progresses from acute, inflammation-driven episodes to progressive stages dominated by neurodegeneration. Among the biological factors underlying this clinical and pathological heterogeneity we found sex. Females exhibit a two- to three-fold higher risk of developing MS and experience more pronounced inflammatory activity. Meanwhile, males are more prone to suffer rapid and severe neurodegeneration.

In the first study, we investigated sex-based molecular differences in MS by analyzing single-cell transcriptomic datasets from the CNS and the peripheral blood mononuclear cells. All together, they represented the different clinical courses of the disease. Following a systematic literature screening, we processed the selected datasets and performed cell type annotation to generate cell type–specific landscapes. These included differentially expressed gene profiles, functional enrichment analyses, protein–protein interaction networks, differential signaling pathway activity, and cell–cell communication networks for females, males, and their sex-differential profiles. In secondary-progressive MS, female neurons may activate protective responses against neurodegeneration, including enhanced GABAergic signaling activity and increased myelin repair potential, whereas male neurons displayed greater vulnerability presenting higher expression of genes related with excitotoxicity. In relapsing-remitting MS, the inflammatory-predominant form, female immune cells presented higher expression of genes related to inflammation, while males exhibited higher expression of genes associated with mitochondrial impairment. Larger differences were reported in CD8+ T cells from primary progressive MS, with females presenting higher expression of genes that may favor homeostasis. Meanwhile, males exhibited cytolytic profiles that may promote neurodegeneration. Comparing the peripheral blood results from both MS subtypes, we defined a sex-differential gene signature of 67 genes related to stimuli responses such as reactive oxygen species, cytokines, lipids, and leukocyte differentiation.

We extended the characterization of sex differences in MS by exploring the gut microbiota through the integration of 16S rRNA sequencing datasets. Individual analyses were conducted for each study, followed by meta-analyses based on pairwise comparisons that accounted for disease status and sex. Despite the strong heterogeneity across datasets, we identified consistent sex-associated microbial taxa. These results were validated computationally in an independent

cohort. Most of the differences were identified when comparing MS females with MS males: *Eggerthella*, *Eisenbergiella*, and *Flavonifractor* were more abundant in females, and *Prevotella* was more abundant in males. These genera have been previously linked to immune modulation and short-chain fatty acid production, which may contribute to the immune responses involved in the disease. Moreover, we reported some associations of these taxa with disease duration and MS subtype, suggesting sex-differential contributions to MS progression.

The final part of this thesis broadened the study of MS into the field of tumor neurobiology. Specifically, melanoma brain metastases (MBM) is an aggressive clinical condition with high incidence among patients with melanoma, limited treatment options, and poor prognosis. Since melanocytes originate from neural crest cells, previous studies have suggested that MBM may exhibit neuron-like expression patterns. Notably, specific associations have been reported between MBM and neurodegenerative disorders. In this work, we aimed to investigate whether MBM shares molecular mechanisms with the neurodegenerative diseases MS, Alzheimer's disease and Parkinson's disease. For neurodegenerative diseases, we performed differential expression analyses between cases and controls in individual datasets, followed by meta-analyses to derive consensus signatures. Then, we compared these profiles with MBM expression profiles evaluating two complementary scenarios: the brain-specific metastatic signature, obtained by comparing MBM with extracranial melanoma metastases, and the tumoral signature, derived from MBM compared to non-tumoral brain tissue. The first scenario revealed 53 dysregulated genes enriched in 11 functional categories, particularly related to the extracellular matrix. The second identified 195 dysregulated genes, mainly involved in development and cell differentiation, chromatin remodeling and nucleosome organization, and translation. Across both scenarios, two genes (ITGA10 and DNAJC6) emerged as consistent markers, highlighting their potential role at the intersection between neurodegeneration and tumor progression.

This doctoral thesis was conducted entirely through the analysis of publicly available datasets, applying meta-analysis strategies whenever possible to enhance the robustness of the findings. Given the large volume of results, interactive web platforms were developed: https://bioinfo.cipf.es/cbl-atlas-ms/ for the single-cell transcriptomic characterization of sex differences in MS, https://irsoler.shinyapps.io/metaanalisis_16S_MS/ for the integrative analysis of sex differences in MS through metagenomics studies, and https://bioinfo.cipf.es/metafun-mbm/ for the neurodegenerative traits of MBM. Collectively, this thesis contributes to a better characterization of the molecular sex-differential mechanisms in MS and the neurodegenerative features of MBM, which may foster future research that ultimately promotes translational applications.

RESUMEN

El sistema nervioso central (SNC) integra funciones motoras, sensoriales y autónomas. Su disfunción ocasiona diversos desórdenes neurológicos, cuyo inicio y progresión están influidos por múltiples factores, incluidas las características intrínsecas del individuo. Esta tesis tiene como objetivo contribuir a la caracterización molecular de las enfermedades del SNC con dos perspectivas complementarias: la identificación de los mecanismos diferenciales por sexo en la esclerosis múltiple (EM) y cómo la caracterización de la EM, junto con otras enfermedades neurodegenerativas, proporciona información sobre la neurobiología de tumores cerebrales.

Esta tesis se centra en la caracterización de las diferencias de sexo en la EM analizando datos de transcriptómica unicelular y metagenómica. La EM es una enfermedad crónica, autoinmune y neurodegenerativa que típicamente progresa desde episodios agudos, dirigidos por inflamación, hasta etapas progresivas dominadas por neurodegeneración. Entre los factores biológicos que subyacen a su heterogeneidad clínica y patológica está el sexo. Las mujeres presentan un riesgo dos a tres veces mayor de desarrollar EM y sufren una actividad inflamatoria más pronunciada. Mientras tanto, los hombres son más propensos a sufrir una neurodegeneración rápida y severa.

En el primer estudio, investigamos diferencias de sexo en la EM mediante el análisis de datos de transcriptómica unicelular, tanto del SNC como de la sangre periférica. En conjunto, los datos analizados representan los diferentes cursos clínicos de la enfermedad. Tras una revisión sistemática de la literatura, procesamos los conjuntos de datos seleccionados y realizamos la anotación de tipos celulares para generar atlas específicos por tipo celular. Estos incluyeron perfiles de genes diferencialmente expresados, análisis de enriquecimiento funcional, redes de interacción proteína–proteína, actividad diferencial de rutas de señalización e interacciones de comunicación celular para mujeres, hombres y sus perfiles diferenciales por sexo. En la EM secundaria progresiva, las neuronas femeninas podrían activar respuestas protectoras contra la neurodegeneración, incluidas una mayor señalización GABAérgica y un mayor potencial de reparación de mielina, mientras que las neuronas masculinas mostraron una mayor vulnerabilidad presentando mayor expresión de genes relacionados con la excitotoxicidad. En la EM remitente-recurrente, el subtipo predominantemente inflamatorio, las células inmunitarias femeninas presentaron mayor expresión de genes relacionados con la inflamación, mientras que los hombres exhibieron una mayor expresión de genes asociados con el deterioro mitocondrial. Las mayores diferencias se identificaron en las células T CD8+ de la forma primaria progresiva, donde las mujeres presentaron una mayor expresión de genes que podrían favorecer la homeostasis. Por su parte, los hombres mostraron perfiles citolíticos que podrían promover la neurodegeneración. Comparando los resultados de la sangre periférica de ambos subtipos de EM, definimos una firma génica diferencial por sexo de 67 genes relacionados con respuestas a estímulos como especies reactivas de oxígeno, citoquinas, lípidos y diferenciación de leucocitos.

Ampliamos la caracterización de las diferencias de sexo en la EM explorando el microbioma intestinal mediante la integración de conjuntos de datos de secuenciación de ARNr 16S. Se realizaron análisis individuales de cada estudio, seguidos de metaanálisis con comparaciones por

pares definidas considerando la presencia o ausencia de enfermedad y el sexo. A pesar de la fuerte heterogeneidad entre estudios, identificamos taxones microbianos consistentes asociados al sexo. Estos resultados fueron validados computacionalmente en una cohorte independiente. La mayoría de las diferencias se identificaron al comparar mujeres con EM frente a hombres con EM: *Eggerthella*, *Eisenbergiella* y *Flavonifractor* fueron más abundantes en mujeres, y *Prevotella* fue más abundante en hombres. Estos taxones han sido previamente vinculados con la modulación inmune y la producción de ácidos grasos de cadena corta, por lo que podrían influir en las respuestas inmunológicas asociadas a la enfermedad. Asimismo, identificamos algunas asociaciones de estos taxones con la duración de la enfermedad y el subtipo de EM, lo que sugiere contribuciones diferenciales por sexo a la progresión de la EM.

La parte final de esta tesis amplió el estudio de la EM hacia el campo de la neurobiología tumoral. Específicamente, las metástasis cerebrales de melanoma (MCM) constituyen una condición clínica agresiva con alta incidencia entre pacientes con melanoma, opciones terapéuticas limitadas y un pronóstico desfavorable. Dado que los melanocitos se originan de las células de la cresta neural, estudios previos sugieren que las células metastásicas cerebrales de melanoma pueden expresar patrones de tipo neuronal. Específicamente, se han reportado asociaciones específicas entre MCM y trastornos neurodegenerativos. En este trabajo, investigamos si MCM comparte perfiles transcriptómicos con las enfermedades neurodegenerativas EM, enfermedad de Alzheimer y enfermedad de Parkinson. En el caso de las enfermedades neurodegenerativas, realizamos análisis de expresión diferencial entre casos y controles a partir de datos individuales, seguido de un metaanálisis para obtener firmas consenso. Posteriormente, comparamos los resultados obtenidos con los perfiles de expresión de MCM evaluando dos escenarios complementarios: la firma metastásica específica del cerebro, obtenida comparando MCM con metástasis extracraneales de melanoma, y la firma tumoral, derivada de MCM comparado con tejido cerebral no tumoral. El primer escenario reveló 53 genes desregulados enriquecidos en 11 categorías funcionales, particularmente relacionadas con la matriz extracelular. El segundo identificó 195 genes desregulados, principalmente involucrados en desarrollo y diferenciación celular, remodelado de la cromatina y organización del nucleosoma, y traducción. Se identificaron dos genes significativos en ambos escenarios (ITGA10 y DNAJC6), destacando su papel potencial en la intersección entre neurodegeneración y progresión tumoral.

Esta tesis se llevó a cabo íntegramente mediante el análisis de conjuntos de datos disponibles públicamente, aplicando estrategias de metaanálisis siempre que fue posible para mejorar la robustez de los resultados. Dado el gran volumen de resultados, se desarrollaron plataformas web interactivas: https://bioinfo.cipf.es/cbl-atlas-ms/ para la caracterización transcriptómica unicelular de las diferencias de sexo en la EM, https://irsoler.shinyapps.io/metaanalisis_16S_MS/ para el análisis integrativo de diferencias de sexo en la EM mediante estudios metagenómicos, y https://bioinfo.cipf.es/metafun-mbm/ para los patrones neurodegenerativos de MCM. En conjunto, este trabajo contribuye a una mejor caracterización de los mecanismos moleculares diferenciales por sexo en EM y de las características neurodegenerativas de MCM, lo que puede fomentar futuras investigaciones que, en última instancia, promuevan aplicaciones traslacionales.

RESUM

El sistema nerviós central (SNC) integra funcions motores, sensorials i autonòmes. La seva disfunció dona lloc a una àmplia varietat de trastorns neurològics, l'inici i la progressió dels quals estan influenciats per múltiples factors, incloent característiques intrínseques específiques de l'hoste. Aquesta tesi té com a objectiu contribuir a la caracterització molecular de les malalties relacionades amb el SNC des de dues perspectives complementàries: l'identificació dels mecanismes diferencials entre sexes en l'esclerosi múltiple (EM), i com la caracterització de l'EM, juntament amb altres malalties neurodegeneratives, pot aportar noves claus per entendre la neurobiologia dels tumors cerebrals.

Aquesta tesi es centra en la caracterització de les diferències de sexe en l'EM mitjancant l'anàlisi de transcriptòmica de cèl·lules individuals i de metagenòmica. L'EM és una malaltia autoimmune i neurodegenerativa crònica que progressa des d'episodis aguts, dominats per inflamació, fins a etapes progressives on predomina la neurodegeneració. Entre els factors biològics subjacents a aquesta heterogeneïtat patològica es troba el sexe. Les dones presenten un risc dos o tres vegades superior de desenvolupar EM i manifesten una activitat inflamatòria més pronunciada. Mentrestant, els homes són més propensos a sofrir una neurodegeneració ràpida i severa.

En el primer estudi, investigàrem les diferències moleculars de sexe en l'EM mitjançant l'anàlisi de conjunts de dades tant de transcriptòmica unicel·lular del SNC i de la sang perifèrica. En conjunt, les dades analitzades representen els diferents cursos clínics de la malaltia. Després d'una revisió sistemàtica de la literatura, processàrem els conjunts de dades adequats i duguérem a terme l'anotació cel·lular per a generar atles específics per tipus cel·lular. Aquests inclogueren perfils de gens diferencialment expressats, anàlisis d'enriquiment funcional, xarxes d'interacció proteïna–proteïna, inferència d'activitat de rutes de senyalització i interaccions de comunicació cel·lular per a dones, homes i els seus perfils diferencials. En l'EM secundària progressiva, les neurones femenines semblen activar mecanismes protectors contra la neurodegeneració, incloent una major activació de la senyalització GABAèrgica i un major potencial de reparació de la mielina, mentre que les neurones masculines mostraren més vulnerabilitat amb una major expressió de gens relacionats amb l'excitotoxicitat. En l'EM remitidora-recurrent, la forma predominantment inflamatòria, les cèl·lules immunitàries femenines presentaren major expressió de gens relacionats amb la inflamació, mentre que els homes mostraren una major expressió de gens associats al deteriorament mitocondrial. Les diferències més notables es detectaren en els limfòcits T CD8+ de l'EM primària progressiva, on les dones presentaren una major expressió de gens que podrien afavorir l'homeòstasi. Per part seua, els homes van mostrar perfils citolítics que podrien promoure la neurodegeneració. Comparant els resultats de la sang perifèrica dels dos subtipus d'EM, definírem una signatura diferencial de sexe composta per 67 gens relacionats amb respostes a estímuls com espècies reactives d'oxigen, citoquines, lípids i diferenciació leucocitària.

Ampliàrem la caracterització de les diferències de sexe en l'EM amb l'estudi del microbioma intestinal a través de la integració de conjunts de dades de seqüenciació d'ARNr 16S. Es

realitzaren anàlisis de cada estudi, i posteriorment metanàlisi, amb comparacions per parelles estratificades per la presència o absència de malaltia i el sexe. Malgrat l'heterogeneïtat entre estudis, identificàrem tàxons consistentment associats al sexe. Aquests resultats foren validats computacionalment en una cohort independent. Les diferències més significatives es detectaren entre dones amb EM i homes amb EM: *Eggerthella*, *Eisenbergiella* i *Flavonifractor* foren més abundants en dones, mentre que *Prevotella* ho fou en homes. Aquests tàxons han estat prèviament vinculats amb la modulació immune i la producció d'àcids grassos de cadena curta, pel que podrien influir en les respostes immunològiques associades a la malaltia. A més, associàrem alguns d'aquests tàxons amb la durada de la malaltia i amb el subtipus d'EM, suggerint contribucions diferencials per sexe a la progressió de la EM.

La part final de la tesi amplià l'estudi de l'EM cap al camp de la neurobiologia tumoral. En concret, les metàstasis cerebrals de melanoma (MCM) constitueixen una condició clínica agressiva, amb una alta incidència entre els pacients amb melanoma, opcions terapèutiques limitades i un pronòstic desfavorable. Atés que els melanòcits s'originen de les cèl·lules de la cresta neural, estudis previs suggereixen que les cèl·lules metastàtiques cerebrals de melanoma poden expressar patrons de tipus neuronal. Específicament, s'han reportat associacions específiques entre MCM i malalties neurodegeneratives. En aquest treball, investigàrem si MBM comparteix mecanismes moleculars amb les malalties neurodegeneratives EM, la malaltia d'Alzheimer i la malaltia de Parkinson. En el cas de les malalties neurodegeneratives, realitzem anàlisis d'expressió diferencial entre casos i controls a partir de dades individuals, seguit de metanàlisis per a obtindre els patrons consens. Posteriorment, comparàrem els resultats obtinguts amb els perfils d'expressió de MCM avaluant dos escenaris complementaris: la signatura metastàtica específica cerebral, obtinguda comparant MCM amb metàstasis extracranials de melanoma, i la signatura tumoral, derivada de la comparació de MCM amb teixit cerebral no tumoral. El primer escenari va revelar 53 gens desregulats enriquits en 11 categories funcionals, particularment relacionades amb la matriu extracel·lular. El segon va identificar 195 gens desregulats, principalment implicats en desenvolupament i diferenciació cel·lular, remodelatge de la cromatina i organització del nucleosoma, i traducció. En ambdós escenaris, dos gens (ITGA10 i DNAJC6) emergiren com a marcadors consistents, ressaltant el seu potencial paper en la intersecció entre neurodegeneració i progressió tumoral.

Aquesta tesi doctoral es dugué a terme íntegrament mitjançant l'anàlisi de conjunts de dades públicament disponibles, aplicant estratègies de metanàlisi sempre que fou possible per a reforçar la robustesa dels resultats. Atés el gran volum de resultats generats, es desenvoluparen plataformes web interactives: https://bioinfo.cipf.es/cbl-atlas-ms/ per a la caracterització transcriptòmica unicel·lular de les diferències de sexe en EM, https://irsoler.shinyapps.io/metaanalisis_16S_MS/ per a l'anàlisi integratiu del microbioma de les diferències de sexe en EM mitjançant estudis metagenòmics i https://bioinfo.cipf.es/metafun-mbm/ per a les característiques neurodegeneratives de MCM. En conjunt, aquest treball contribueix a una millor caracterització dels mecanismes moleculars diferencials per sexe en l'EM i de les característiques neurodegeneratives de MCM, el que pot fomentar futures investigacions que, en última instància, promoguen aplicacions translacionals.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Aβ	Amyloid-β
AD	Alzheimer's disease
AD-CT	Alzheimer's disease - cortex
AD-HP	Alzheimer's disease - hippocampus
ASV	Amplicon sequence variant
Bact1	Bacteroides 1
Bact2	Bacteroides 2
BBM	Breast brain metastasis
BH	Benjamini-Hochberg
BMI	Body mass index
CAR	Chimeric antigen receptor
Ca2+	Calcium
cDNA	Complementary DNA
CIPF	Centro de Investigación Príncipe Felipe
CNS	Central nervous system
dbRDA	Distance-based redundancy analysis
DL	DerSimonian–Laird
DMM	Dirichlet multinomial mixture
DMT	Disease-modifying therapy
DNA	Deoxyribonucleic acid
DOI	Digital object identifier
EBNA1	Epstein-Barr nuclear antigen 1
EBV	Epstein-Barr virus
EDSS	Expanded disability status scale
EMBL-EBI	European Molecular Biology Laboratory – European Bioinformatics Institute
ENA	European Nucleotide Archive
FAIR	Findability, Accessibility, Interoperability, and Reuse
FDR	False discovery rate
FGF	Fibroblast growth factor
GABA	Gamma-aminobutyric acid
GEO	Gene Expression Omnibus

GO	Gene Ontology
GTDB	Genome Taxonomy Database
HGNC	HUGO Gene Nomenclature Committee
HLA	Human leukocyte antigen
HMP	Human Microbiome Project
HVG	High variable gene
IDF	Impact of disease in females
IDM	Impact of disease in males
INSDC	International Nucleotide Sequence Database Collaboration
ITS	Internal transcribed spacer
KEGG	Kyoto Encyclopedia of Genes and Genomes
K^+	Potassium
lncRNA	Long non coding RNA
logFC	Logarithm of fold change
MAD	Median absolute deviation
MAST	Model-based analysis of single-cell transcriptomics
MBM	Melanoma brain metastases
MetaHIT	Metagenomics of the Human Intestinal Tract
MHC	Major histocompatibility complex
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
Na^+	Sodium
NCBI	National Centre for Biotechnology Information
NFT	Neurofibrillary tangle
NK	Natural killer
OPC	Oligodendrocyte precursor cell
ORA	Over-representation analysis
PBMC	Peripheral blood mononuclear cell
PC	Principal component
PCA	Principal component analysis
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction

PD	Parkinson’s disease
PD-SN	Parkinson’s disease - substantia nigra
PD-ST	Parkinson’s disease - striatum
PDGF	Platelet-derived growth factor
PPI	Protein-protein interaction
PPMS	Primary progressive multiple sclerosis
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Prev	Prevotella
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
RRMS	Relapsing-remitting multiple sclerosis
Rum	Ruminococcus
rRNA	Ribosomal ribonucleic acid
SCFA	Short-chain fatty acid
scRNA-seq	Single-cell RNA sequencing
SDID	Sex differential impact of disease
SE	Standard error
SNN	Shared Nearest Neighbor
snRNA-seq	Single-nucleus RNA sequencing
SPMS	Secondary progressive multiple sclerosis
SRA	Sequence Read Archive
tSNE	T-distributed stochastic neighbor embedding
UCSC	University of California Santa Cruz Cell Browser
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
USA	United States of America
WGS	Whole-genome shotgun

1. General introduction

This general introduction is structured to outline the biological and bioinformatic concepts required to contextualize the topics addressed in this doctoral thesis. It begins with the characterization of multiple sclerosis (MS) disease, then extends to other central nervous system (CNS) disorders, and concludes with an overview of the types of omics data analyzed throughout this work.

1.1. MULTIPLE SCLEROSIS DISEASE

MS is a chronic and autoimmune condition characterized by the neurodegeneration of the CNS. Although its etiology remains incompletely understood, MS is considered a multifactorial disease in which both genetic susceptibility and environmental factors contribute to disease onset and progression[1,2].

Neurodegeneration in MS is primarily immune-mediated by peripheral blood immune cells that infiltrate the CNS crossing the blood-brain barrier. These cells trigger autoreactive responses against the myelin sheath, the lipid-rich neuronal envelope that acts as an electrical insulator to ensure the effective transmission of nerve impulses. The resulting myelin destruction, known as demyelination, leads to the formation of focal plaques that cause structural and functional damage to the affected CNS regions. Clinical manifestations vary widely depending on the location of these plaques, resulting in motor, sensory, visual, and cognitive impairments that compromise the autonomy of the individual, ultimately leading to the death of the patient[1–3].

1.1.1. EPIDEMIOLOGY AND SOCIAL IMPACT

Unlike other neurodegenerative disorders such as Parkinson's or Alzheimer's diseases, which primarily affect older individuals, MS onset ranges between 20 and 40 years. As a result, it affects the young adult population, being the leading cause of non-traumatic neurological disability in this age group[4].

According to the Multiple Sclerosis International Federation, nearly three million individuals were living with MS worldwide in 2020, based on data collected from 115 countries. These results represent a 30% increase compared to the previous global survey conducted in 2013. However, the reasons underlying this prevalence growth remain inconclusive. It is not known whether this increase reflects improved diagnostic capabilities and longer life expectancy, or a substantially increase in MS incidence[5].

Moreover, the global prevalence of MS is not uniform, as it varies significantly across geographical regions. Higher prevalence rates, exceeding 100 cases per 100,000 inhabitants, are observed in Europe, North America, and Australia. Meanwhile, countries in South America, Africa, and Southeast Asia reported lower prevalence rates, with fewer than 25 cases per 100,000 inhabitants (**Figure 1.1**). These geographical disparities are thought to be multifactorial, including differences in public awareness, access to healthcare systems, and genetic and environmental risk factors[5,6].

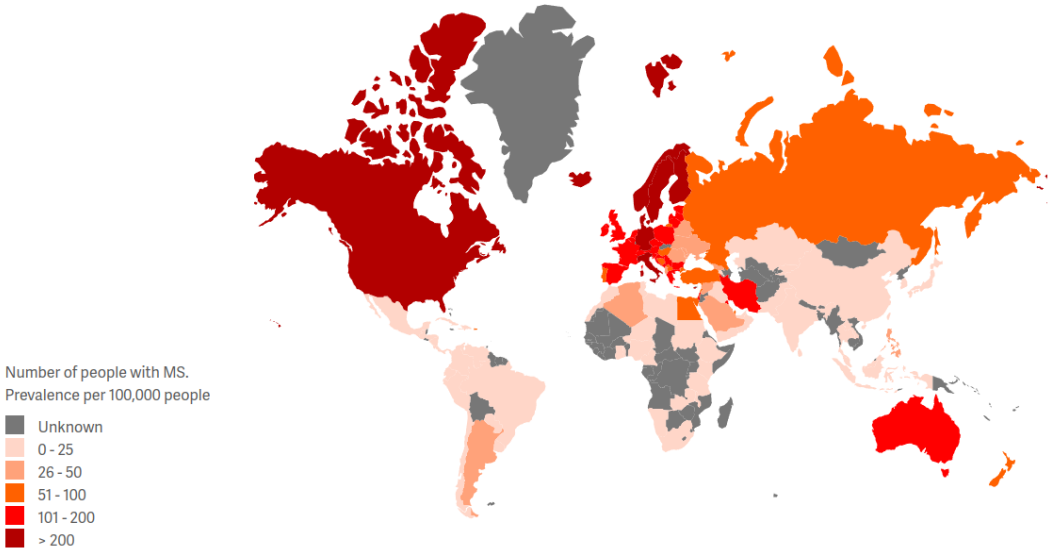


Figure 1.1. Global prevalence of multiple sclerosis in 2020. Color scale indicates the number of individuals diagnosed with multiple sclerosis per 100,000 inhabitants in each country, based on data compiled in the third edition of the *Atlas of MS* published by the Multiple Sclerosis International Federation. Figure from Walton *et al.* 2020[5].

Regardless of the geographical region, MS highly impairs the quality of life of the patients, affecting not only physical health but also mental and emotional well-being. Affected individuals frequently experience fatigue and reduced physical mobility, which limit their ability to perform daily activities[7]. In addition to motor deficits, MS is often associated with the dysfunction of the autonomic nervous system, leading to urinary tract disturbances, sexual dysfunction, and cardiovascular or gastrointestinal dysregulation[8]. Cognitive impairments are also reported, particularly affecting recent memory, attention capacity, verbal fluency, conceptual reasoning, and visuospatial perception.

Furthermore, they also present elevated rates of comorbidities that are both psychological (e.g., depression, anxiety) and somatic (e.g., hypertension, diabetes)[9].

Overall, these clinical manifestations result in increased dependency in daily activities, reduced social participation and loss of employment. Many patients require continuous support from caregivers, which are normally their partners or close relatives. Sociodemographic factors such as socioeconomic status, education, and social support also impact MS outcomes. The economic burden is also substantial, with estimated total annual costs reaching €40,303 per patient in Europe, encompassing direct medical expenses and patient care[10].

1.1.2. PRIMARY AFFECTED TISSUES

MS is a neurodegenerative and autoimmune disorder, primarily impacting the CNS and the immune system. Investigating both systems is essential for a better understanding of the disease, as each exhibits its own particularities yet interrelated features.

The CNS, composed of the brain and the spinal cord, serves as the body's principal control center. It processes inputs from peripheral tissues and organs, integrates incoming information, and coordinates responses to maintain physiological homeostasis. The CNS is responsible for a variety of actions, including processing and responding to sensory and motor stimuli, and maintaining the involuntary autonomic systems[11].

Meanwhile, the primary role of the immune system is to protect the organism from potentially harmful factors, both exogenous and endogenous. Immune cells can be established within tissues as resident immune cells or circulate in the bloodstream as peripheral immune cells. The latter comprises the peripheral blood mononuclear cells (PBMCs), which are the cell types analyzed in this work. In addition, the response mechanisms driven by immune cells are classified into two categories: the innate immune system, which provides rapid, non-specific protection through physical barriers (such as skin and mucous membranes) and phagocytic cells; and the adaptive immune system, which operates with longer response times and exhibits high specificity, primarily mediated by lymphocytes (B and T cells)[12].

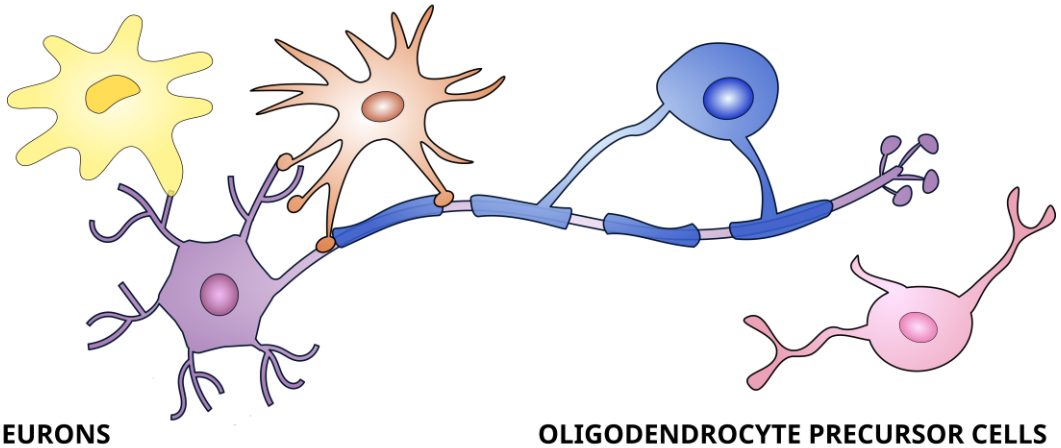
Thus, the CNS and the immune system are composed of multiple cell types, which perform specific and interconnected activities to execute the attributed functions that are summarized in **Figure 1.2.**

A

MICROGLIA
Immune cells of the CNS. Remove cellular debris by phagocytosis*.

ASTROCYTES
Support neurons by providing nutrients and maintaining ionic balance. Maintain the blood brain barrier.

OLIGODENDROCYTES
Constitute and maintain the myelin sheaths.

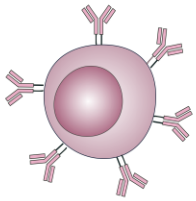


NEURONS
Transmit information through electrical and chemical signals.

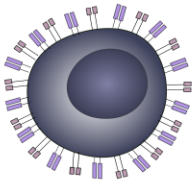
OLIGODENDROCYTE PRECURSOR CELLS
Precursors to mature oligodendrocytes. Respond to demyelination by proliferating and differentiating.

B

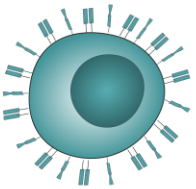
B CELLS
Produce antibodies against specific antigens.



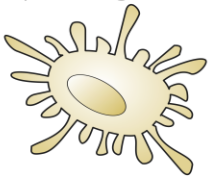
CD8+ T CELLS (cytotoxic T cells)
Recognize and eliminate infected or cancerous cells presenting foreign antigens.



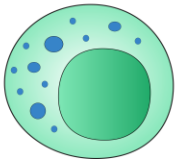
CD4+ T CELLS (helper T cells)
Assist in the activation of B cells and cytotoxic T cells.



DENDRITIC CELLS
Professional antigen-presenting cells.



NATURAL KILLER CELLS
Eliminate infected or cancerous cells laking self-markers.



MONOCYTES
Macrophage precursors. Phacogitosis* of particles** and microorganisms.

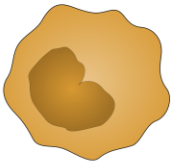


Figure 1.2. Functions of the major cell types from (A) the central nervous system and (B) the peripheral blood immune system. (*Previous page*) (A) Cell types from the central nervous system: astrocytes, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells. (B) Cell types from peripheral blood mononuclear cells: B cells, CD8+ T cells, CD4+ T cells (from the adaptive immune system) and dendritic cells, natural killer cells and monocytes (from the innate immune system). *Phagocytosis: endocytosis of elements via plasma membrane extensions. **The term *particles* refers to metabolites and debris.

1.1.3. PATHOPHYSIOLOGY

In the CNS of healthy individuals, neuronal axons are covered by the lipid-rich plasma membranes of the oligodendrocytes. This structure, called myelin, serves as an electrical insulator that ensures the effective transmission of nerve impulses. It enables the rapid and efficient propagation of action potentials and the maturation and maintenance of neural circuits. In MS, this myelin sheath is destroyed by immune-mediated mechanisms, leading to a pathological process known as demyelination[13].

In MS pathogenesis, a subset of peripheral T lymphocytes becomes autoreactive against membrane proteins located on the surface of the myelin sheaths. The specific mechanisms responsible for this loss of immune tolerance remain under investigation. The most widely accepted hypothesis involves the molecular mimicry between antigens from the Epstein-Barr virus (EBV) and the proteins from the myelin sheath. Specifically, structural similarity has been described between the EBV nuclear antigen 1 (EBNA1) and the myelin-associated protein GlialCAM, potentially triggering cross-reactive immune responses[14]. In fact, the research published by Bjornevik *et al.* 2022[15], conducted on a cohort of over 10 million military personnel, revealed that the risk of developing MS increases more than 32-fold following EBV infection. Meanwhile, individuals who were EBV-negative exhibit almost no risk for developing the disease. Given that over 90% of the global population is infected with EBV, current research is focused on describing differential immunological responses to explain why only a subset of individuals develop MS[16,17].

The mechanism that results in neuronal damage is illustrated in **Figure 1.3.** The autoreactive T cells interact with B cells, leading to the production of antibodies against myelin-associated proteins (**Figure 1.3-2a**). These interactions may occur in both the peripheral blood system and the CNS[18].

The infiltration of the autoreactive lymphocytes into the CNS is promoted by the increased permeability of the blood-brain barrier, the structure responsible for regulating

the selective traffic flow of molecules between the bloodstream and the CNS. Once inside the CNS, the lymphocytes trigger autoimmune responses characterized by the production of pro-inflammatory cytokines and antibodies (**Figure 1.3-3**), leading to the demyelination of the neuronal axons (**Figure 1.3-4**). The infiltrating T cells can also activate the microglial cells (**Figure 1.3-2b**), promoting demyelination, inflammation and the phagocytosis of myelin debris[19].

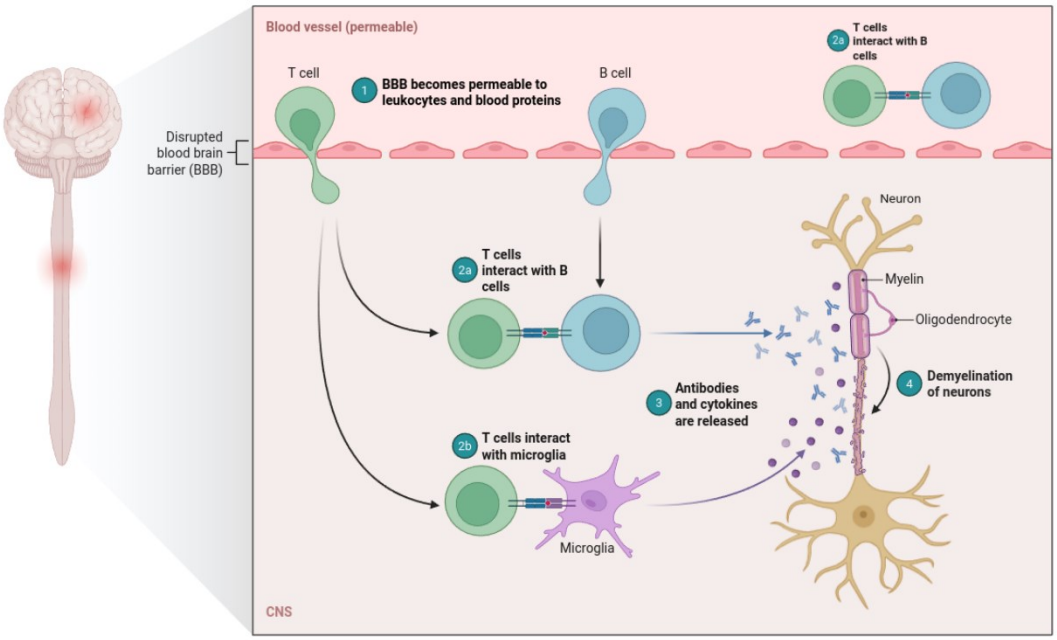


Figure 1.3. Mechanism underlying neuronal damage in multiple sclerosis. Autoreactive lymphocytes initiate autoimmune responses in the CNS, leading to the production of pro-inflammatory cytokines and antibodies. These processes, together with microglial activation, result in inflammation and axonal demyelination. *BBB: blood brain barrier; CNS: central nervous system.* Figure adapted from the *Biorender* template created by Akiko Iwasaki and titled *Pathogenesis of Multiple Sclerosis*.

The CNS regions affected by the demyelination process are identified as focal and delimited areas of damage, known as lesions or plaques. In the early stages, lesions are classified as acutely active. They are characterized by active demyelination and oligodendrocyte damage, activated microglia, and high infiltration of phagocytic peripheral immune cells. When the myelin is completely destroyed and debris has been

cleared from the center of the lesion, it progresses into a chronic active stage. At this phase, an inflammatory rim is created around the lesion, conformed by activated immune cells and microglia. Meanwhile, astrocytes initiate the process of generating a glial scar in the center of the lesion. Active lesions can transition into a chronic inactive state, where the inflammatory activity of the rim decreases and the astrocytes complete the formation of the compact glial scar[20,21].

The number and activity status of MS lesions evolve over the course of the disease. Magnetic resonance imaging (MRI) is the standard tool for monitoring lesion evolution and CNS atrophy progression. A contrast agent (typically gadolinium) is administered. Lesions that retain the contrast molecules indicate active inflammation, high immune activity and increased vascular permeability. In contrast, chronic inactive lesions do not show contrast enhancement, and are frequently reported together with progressive CNS atrophy. Longitudinal tracking of lesion activity and brain volume loss with MRI, together with the evolution of the patient's disability, allows the classification of MS patients into clinical subtypes as detailed in the following section[22,23].

1.1.4. MULTIPLE SCLEROSIS SPECTRUM

MS exhibits high heterogeneity in disability progression and MRI activity, both across patients and within individuals over time. To facilitate communication among researchers and clinicians, the first clinical classification based on patient phenotypes was established in 1996[24]. This system was updated in 2013 to incorporate advances in biomedical imaging, biomarker characterization, and a deeper understanding of the pathological hallmarks[25]. Under this description, three clinical subtypes are recognized. The most common form is called relapsing-remitting MS (RRMS), which fluctuates between periods of relapses, driven by exacerbated inflammatory status and followed by subsequent remissions. When RRMS transitions into a progressive phase of disability, it gives rise to secondary progressive MS (SPMS). Lastly, primary progressive MS (PPMS) is characterized by a continuous neurological decline from the onset, predominantly driven by neurodegeneration and not preceded by peaks of inflammatory relapses. Despite this general description, the manifestations of each subtype can vary greatly over time. A representative illustration of the time course for each subtype is exemplified in **Figure 1.4**[26].

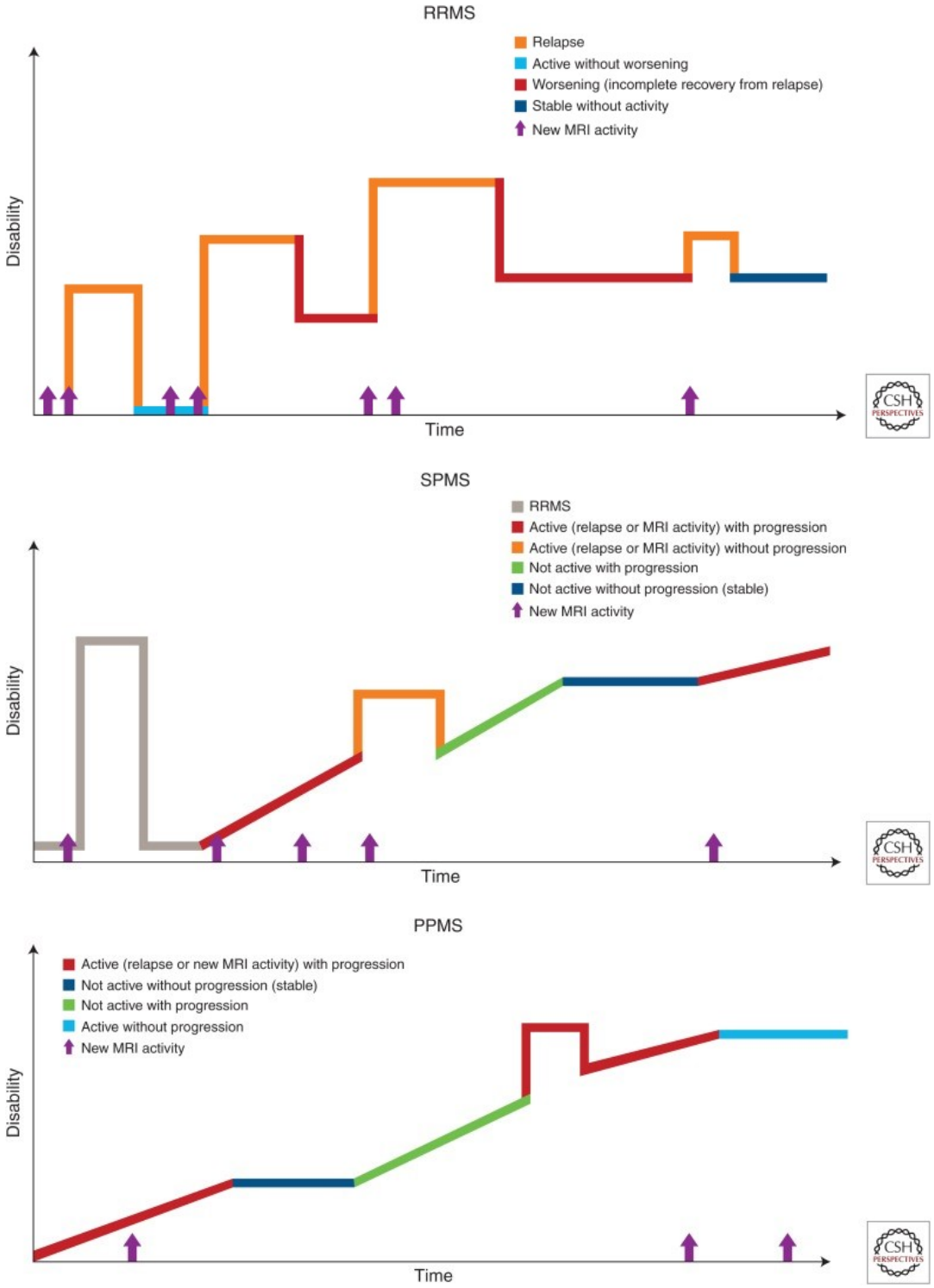

RRMS
Relapse
Active without worsening
Worsening (incomplete recovery from relapse)
Stable without activity
New MRI activity
Disability
Time
SPMS
RRMS
Active (relapse or MRI activity) with progression
Active (relapse or MRI activity) without progression
Not active with progression
Not active without progression (stable)
New MRI activity
Disability
Time
PPMS
Active (relapse or new MRI activity) with progression
Not active without progression (stable)
Not active with progression
Active without progression
New MRI activity
Disability
Time

Figure 1.4. Illustration of disability progression across the three main subtypes of multiple sclerosis. (*Previous page*) (A) Relapsing-remitting MS (RRMS), (B) Secondary progressive MS (SPMS), and (C) Primary progressive MS (PPMS). Line colors indicate different subclinical phases of MS. The terms *active* and *inactive* refer to the presence or absence of new activity detected by magnetic resonance imaging (MRI), respectively. *MS: multiple sclerosis.* Figure from *Klineova et al.* 2018[26].

The pathophysiological mechanisms underlying MS are also highly heterogeneous. As a result, two patients may exhibit the same degree of disability, although the molecular processes may differ in type and their relative contribution, as illustrated in **Figure 1.5**[27]. These processes include mechanisms such as neuronal degeneration and impaired remyelination subsequent to myelin loss.

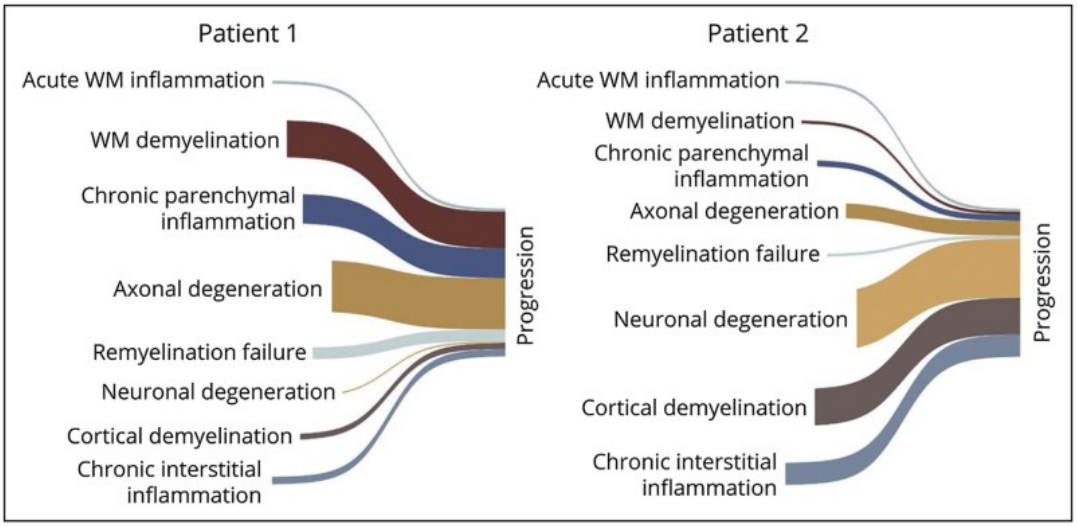


Figure 1.5. Illustrative comparison of the underlying pathological mechanisms in two multiple sclerosis patients with the same degree of disability. The width of each pathway indicates the relative contribution of each process to overall disease progression. *WM: white matter*. Figure from Krieger *et al.* 2024[27].

1.1.5. DIAGNOSIS

The diagnostic approach to MS consists of conducting clinical tests aimed at excluding alternative conditions with similar clinical manifestations, while simultaneously identifying hallmark features of MS. Frequent diagnostic methods include[28]:

- Neurological examinations to assess the status of the CNS functionality.
- Blood analysis to exclude infectious, metabolic, or autoimmune disorders that may mimic MS symptomatology.
- Cerebrospinal fluid analysis to detect oligoclonal bands that may reflect the presence of antibodies in the CNS capable of recognizing autoantigens.
- Neuroimaging through MRI techniques for the identification of CNS demyelinating lesions, differentiating MS from other neurological disorders.

If alternative diseases are discarded, the 2017 revised McDonald criteria are applied to confirm the diagnosis of MS and to determine the clinical subtype[29]. In the case of the RRMS, multiple combinations of clinical, radiological, and laboratory findings can establish the diagnosis (**Table 1.1**). In contrast, the diagnosis of progressive subtypes requires other outcomes: the individual must exhibit progressive neurological symptoms without remission for at least 12 months, along with at least two of the following three attributes[4,13,29]:

- The presence of at least one demyelinating lesion in specific brain regions.
- The presence of at least two demyelinating lesions in specific spinal cord regions.
- Detection of autoantigen-specific antibodies in the cerebrospinal fluid.

Table 1.1. Revised 2017 McDonald criteria for the diagnosis of relapsing-remitting multiple sclerosis. Possible diagnostic combinations based on the number of relapses (column 1), the number of CNS lesions detected by magnetic resonance imaging (column 2) and the required additional evidence to confirm multiple sclerosis diagnosis (column 3). *Multifocal lesion: a group of lesions with a common origin affecting different anatomic areas of the brain. *CNS: central nervous system.* Table adapted from J. A. Thompson *et al.* 2017[29].

Number of relapses	Number of lesions	Additional evidence required
≥ 2	≥ 2	None
≥ 2	1	Medical history indicating a previous relapse involving a different region of the CNS from the reported lesion
≥ 2	1	Multifocal lesion*
1	≥ 2	Presence of autoantigens-recognizing antibodies in the cerebrospinal fluid
1	≥ 2	Medical history indicating a previous relapse involving a different region of the CNS from the reported lesion
1	1	Multifocal lesion* and presence of autoantigens-recognizing antibodies in the cerebrospinal fluid
1	1	Multifocal lesion* and medical history indicating a previous relapse involving a different region of the CNS from the reported lesion

Although the implementation of the revised 2017 McDonald criteria represented a major breakthrough in MS diagnosis by incorporating MRI-based imaging, in recent years additional biomarkers have been proposed. These include the detection of kappa free light chains in cerebrospinal fluid, the identification of demyelinating lesions with the central vein sign via MRI or paramagnetic rim lesions, and the use of emerging

automated imaging techniques based on machine learning algorithms, all of which have shown promising diagnostic value[30]. Thanks to these advances, an update known as the revised 2024 McDonald criteria was proposed. Preliminary drafts have already been presented at scientific meetings, including the 2025 Annual Meeting of the American Academy of Neurology (https://www.aan.com/msa/Public/Events/Details/18130, last accessed July 18, 2025) and the European Committee for Treatment and Research in Multiple Sclerosis (https://ectrims.eu/webinar-highlights-2024-mcdonald-diagnostic-criteria-for-multiple-sclerosis/, last accessed July 18, 2025). Their official publication is expected soon.

1.1.6. TREATMENTS

To date, MS is not yet curable. Available treatments are primarily prescribed to alleviate symptoms, reduce the frequency of relapses, prevent the formation of new lesions, and delay disability progression. The most widely used medications are disease-modifying therapies (DMTs), which modulate or suppress immune responses. The choice of an appropriate DMT is individualized, taking into account the patient's disability, disease activity, comorbidities, and the response to previous therapies[31,32].

Considering the balance between therapeutic efficacy and the risk of adverse effects, the prescription of DMTs commonly follows an escalation approach. First-line therapies are the first to be administered, offering moderate efficacy but a favorable safety profile. These DMTs include injectable compounds such as interferon beta and glatiramer acetate, and oral agents like dimethyl fumarate and teriflunomide. Collectively they are referred to as classic immunomodulators, as they act by modulating the peripheral immune response, shifting the balance between pro-inflammatory and anti-inflammatory cytokines. If the patient presents insufficient response, clinicians may substitute the corresponding DMT to an alternative first-line therapy, or consider the combination of different compounds[31,32].

In patients whose disease activity persists, evidenced by reporting new relapses or MRI-detected lesions, the treatment is escalated to second-line therapies. These treatments present higher efficacy but are associated with an increased risk of adverse effects. One example is natalizumab, which blocks lymphocyte migration into the CNS without markedly altering their peripheral blood levels[33]. Another example is fingolimod, which retains lymphocytes in secondary lymphoid organs, thereby reducing their presence in the peripheral blood and the CNS[34]. If these therapies fail to improve the patient's disability, clinicians can also consider third-line treatments. These include cytotoxic immunosuppressants such as cladribine, which deplete both T and B cells from the

peripheral blood[35]. Third-line treatments offer high efficacy, at the expense of causing immunosuppression which can lead to serious illness derived from opportunistic infections[31,32].

An alternative to the escalation strategy is the induction approach. It consists of the early administration of second-line DMTs to achieve rapid control of the symptoms. This strategy is often used with patients with highly active MS, characterized by frequent relapses or rapid accumulation of disability. Regardless of whether an escalation or induction approach is employed, the treatment may need to be temporarily or permanently interrupted if significant adverse effects are reported[36].

It is important to note that DMTs are significantly less effective in progressive forms of MS (SPMS and PPMS) than in RRMS. Only a limited number of DMTs have been approved for use in progressive MS. Additionally, their prescription is restricted to subpopulations of patients who exhibit signs of active inflammation on MRI. Specifically, ocrelizumab was approved for the treatment of PPMS[37] and siponimod for SPMS[38].

Considering all above, there is a need for treatments with fewer adverse effects to reduce disability and potentially cure the disease. These treatments should also be applicable to patients with progressive subtypes, where inflammation is not pronounced. Different clinical trials are currently ongoing, all together covering the three MS subtypes. Among the most promising approaches are Bruton's tyrosine kinase inhibitors, small molecules that can cross the blood–brain barrier and modulate CNS-resident inflammatory cells.[39] Other strategies include CD19 chimeric antigen receptor (CAR)-T cell therapies[40], antiviral therapies and vaccines[41,42], combination of DMTs with gut microbiota modulators[43] and sex hormone-based treatments[44].

In addition to DMTs, patients with RRMS may also receive corticosteroids during relapses, given their anti-inflammatory properties. Physiotherapy and psychological support also play an important role mitigating the clinical manifestations of the disease and improving patients' overall quality of life. Furthermore, it is important to manage patient-specific comorbidities[45].

1.1.7. RISK FACTORS

MS is a multifactorial disease resulting from the synergistic interaction of genetic and environmental factors. It is suggested that, given a certain genetic predisposition, environmental factors and lifestyle-related factors could act as triggers for the autoimmune response. Although MS is not directly inherited, genetic susceptibility has

been reported. Specifically, the probability of developing MS increases proportionally with the closer degree of kinship to an affected family member[46].

Some of the main genetic risk factors are associated with specific alleles of the human leukocyte antigen (HLA) complex. HLA family genes play a central role in antigen presentation, being implicated in the immunological process of discriminating between self and foreign proteins. The HLA-DRB1*15:01 haplotype is particularly notable, as it can triple the risk of MS in individuals carrying at least one copy of this allele. Consequently, it corresponds to the most common variant found in MS patients. Conversely, other alleles such as HLA-A*02 appear to have a protective effect, reducing the risk of developing the disease[47,48]. Additional genetic variants contributing to MS susceptibility have been identified, including polymorphisms in genes related with the immune system, the mitochondrial activity and the CNS repair capacity[49].

In addition to the genetic susceptibility, different environmental and lifestyle-related factors have been identified to contribute to the risk of developing MS. Among the most widely accepted are vitamin D deficiency, residence in high-latitude regions with limited sunlight exposure, EBV infection, active smoking, and female sex. More recently, other aspects have become increasingly relevant, including obesity during adolescence, chronic exposure to environmental pollutants, and circadian rhythm disruption associated with night-shift work[47,50,51].

1.1.8. SEX DIFFERENCES

Among the factors described previously, being female has been associated with an increased risk of developing MS. In a broader context, sex is a relevant variable that can modulate the onset, progression and outcome of numerous diseases, including both autoimmune[52] and neurodegenerative[53] conditions.

Before presenting the content of this section, it is noteworthy to define the terms *sex* and *gender*. They are not synonymous, although both influence health and disease and are interconnected (**Figure 1.6**). The term *sex* refers to the biological characteristics that determine whether an individual is male or female, such as chromosomal endowment, reproductive organs and hormonal composition. In contrast, gender refers to the sociocultural dimension, referring to how individuals perceive themselves and behave within society according to socially constructed norms of femininity and masculinity. Although these concepts are usually considered in binary terms (sex: female–male; gender: woman–man), it is important to acknowledge that individuals may exhibit intermediate or non-binary characteristics in both biological and social contexts[54].

In this doctoral thesis, we explore MS from a sex-based perspective, identifying differences between female and male individuals.

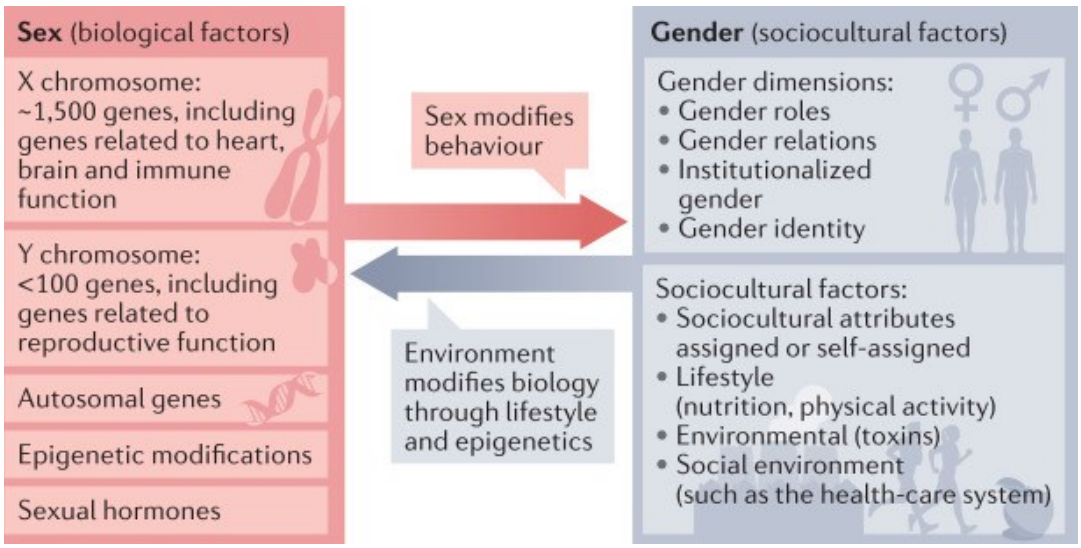


Figure 1.6. Associated terms with the definition of sex and gender. Figure from Regitz-Zagrosek *et al.* 2023[55].

1.1.8.1. Epidemiology

Females tend to develop MS at an earlier age compared to males. They also exhibit a higher prevalence, with a female-to-male ratio ranging from 2:1 to 3:1. This ratio has been increasing in recent years, primarily due to a rising incidence among females rather than a decrease in incidence among males. Females also present more severe inflammatory patterns, with increased production of pro-inflammatory cytokines[56,57]. In contrast, males experience faster CNS deterioration, characterized by more significant atrophy and greater neuronal loss[58]. Furthermore, the study performed by Tolaymat *et al.* 2020[59] reported that males may exhibit a higher proportion of lesions with detectable rims on MRI. Thus, females appear to display a greater capacity to maintain the neurofunctional activity despite the inflammatory burden[60].

These differences are also reflected in the clinical categorization of the disease. Females are more prone to suffer RRMS, while the progressive forms tend to show a more

balanced ratio between sexes[61]. In all clinical subtypes (RRMS and MS progressive forms), females exhibited higher relapse activity compared to males[61,62]. However, male sex has been associated with a shorter time and an earlier age to transition from RRMS to SPMS[63].

Regarding treatments, the potential influence of sex on therapeutic response is not always explicitly assessed. In clinical trials that stratified data by sex, clear sex-based differences are not often reported[64]. However, some studies have documented disparities. For instance, it has been suggested that teriflunomide and fingolimod have a more pronounced effect in reducing relapse rates in female patients. Conversely, in the PROMISE PPMS study, where interferon-beta was evaluated as a therapy for PPMS, males appeared to benefit more from this treatment than females[65].

Studies conducted in humans and animal models have investigated these sex-differential disparities in MS, primarily focusing on the influence of sex chromosomes, the role of sex hormones, and the contribution of specific genes and molecular mechanisms[66].

1.1.8.2. Sexual chromosomes

To study the effects of sex chromosomes independently of the hormonal influence, researchers commonly use the *Four Core Genotypes* mouse model[67,68]. This system generates four experimental groups based on the insertion or deletion of the *Sry* gene, which determines the development of male gonads. The typical XX females possess ovaries, while XY males develop testes. Then, XY mice lacking the *Sry* gene retain male sex chromosomes but develop ovaries, whereas XX females which express *Sry* maintain female sex chromosomes but develop testes. It was observed that, independently of the hormonal endowment, model organisms for MS with XX chromosomes exhibited a higher rate of CNS remyelination compared to those with XY chromosomes[69]. Alternatively, XY organisms suppressed the expression of genes that promote hyperactive T-cell responses[70]. Potential explanations for these XX versus XY differences include the presence or absence of Y-linked genes, differences in X chromosome gene dosage, and variations in parental imprinting of X-linked genes[71].

1.1.8.3. Sexual hormones

Sex hormones also play a crucial role in the pathogenic mechanisms of MS. They influence the CNS and the immune system responses, primarily based on neuroprotective and immunomodulatory properties of hormones such as testosterone, progesterone and estrogens[72]. The clinical relevance of sex hormones in MS has been particularly studied in the context of pregnancy. One of the first studies performed with a large prospective cohort was elaborated by Confavreux *et al* in 1998[73]. In this work, the authors observed a 70% reduction in relapse rates during the third trimester of pregnancy, followed by a 70% increase in the first three months postpartum.

Testosterone and progesterone have been associated with neuroprotective effects, the enhancement of anti-inflammatory responses in the peripheral blood, and the improvement of CNS functionality in MS[74]. Notably, progesterone has also shown remyelination-promoting effects[74]. Among estrogens, estriol is the most widely studied. Elevated estriol levels have been correlated with reduced T cell infiltration into the CNS, decreased inflammation, and enhanced remyelination through the promotion of oligodendrocyte precursor cell (OPC) proliferation[74,75]. Interestingly, *in vitro* experiments with EBV infected B cells have reported that another estrogen named estradiol can induce transcription of viral genes such as EBNA2[76]. Thus, suggesting a potential interaction between sex hormones and viral mechanisms involved in MS.

Considering the benefits of sex hormones, different clinical trials evaluated testosterone and combinations of progesterone with estriol as potential treatments for males and females, respectively[77]. However, the results of the clinical trials are inconsistent, possibly due to the heterogeneity in experimental design, MS subtype and time of intervention among other factors. While some studies reported no benefit from the administration of sex hormones, others revealed a gradual improvement in the progression of MS disability[77]. It should also be considered that the effectiveness of the currently administered DMTs may be affected by the female hormonal status, as it varies across the different phases of their life[78].

1.1.8.4. Focusing on molecular mechanisms

An alternative approach to explore sex differences is to focus on the molecular and cellular mechanisms that are altered, regardless of whether these differences originate from sex chromosomes, sex hormones, or autosomal and/or mitochondrial genes. One example of this approach is the work of Català-Senent *et al.* 2023[79], who conducted a

meta-analysis of bulk RNA-seq and microarray datasets that were available in the scientific literature at the time. By integrating a total of nine transcriptomic studies from the CNS and peripheral blood, the authors identified sex-specific gene expression patterns associated with MS for each tissue. The functional analysis of the CNS results revealed sex-dependent differences in immune-related functions. Furthermore, females with MS exhibited alterations in purine and glutamate metabolism compared to males, while males with MS exhibited significant functions related to metal ion stress responses, among others.

Sex differences can also be evaluated through the characterization of specific molecular mechanisms. The study elaborated by Noriko Itoh *et al.* 2023[80] is based on the characterization of the synaptic mitochondrial function. The authors identified that male mice with MS exhibited significant morphological abnormalities compared to females. Moreover, the mitochondria located at the axon terminals of male neurons consumed less oxygen compared to those of females, pointing to a sex-dependent difference in mitochondrial bioenergetics.

Additional mechanisms underlying sexual dimorphism in MS pathophysiology are explored throughout this doctoral thesis. These are further detailed in the sections titled *Contextualization, motivation and objectives* of the chapters *Study I* and *Study II* of this manuscript.

1.2. BEYOND MULTIPLE SCLEROSIS DISEASE: EXPLORING CENTRAL NERVOUS SYSTEM DISORDERS

As mentioned in the previous section, the chapters corresponding to *Study I and Study II* of this doctoral thesis focus on characterizing molecular mechanisms in the context of a specific disease (MS) and a specific risk factor (sex). However, the underlying processes involved, such as neuronal dysfunction, oxidative stress, and inflammation, are not exclusive to MS. These mechanisms may represent hallmarks that are also implicated in other CNS disorders.

The CNS is a highly complex system responsible for integrating sensory information, regulating motor functions, maintaining systemic homeostasis, and executing cognitive processes such as memory, language, and decision-making. To perform these functions, the CNS relies on a unique structural and cellular organization, characterized by high metabolic demands, intricate cellular interactions, and a limited regenerative capacity[81].

Given the complexity and wide range of functions of the CNS, it is susceptible to a broad spectrum of disorders. Some of these diseases involve both the brain and spinal cord, whereas others predominantly affect only one of these structures. Among them we can find:

- Neurodegenerative disorders, including Alzheimer's disease and Parkinson's disease[82].
- Injury-related conditions, such as traumatic brain injury and spinal cord injury[83].
- Infections with viral, bacterial, or fungal origin[84].
- Cancer, which ranges from primary malignant brain tumors to secondary tumors arising from metastases of extracranial origin[85,86].

Collectively, these disorders represent one of the leading contributors to the global health burden. Their impact is highly significant due to their poor clinical outcomes. Moreover, many of these conditions are chronic and progressive, leading to long-term physical, cognitive, and emotional disability[87]. The economic impact is also substantial. Considering brain disorders alone, the global economic burden was estimated at $1.14 trillion in 2019[88].

Although CNS disorders are distinct disease entities, characterized by specific clinical features and progression trajectories, they are not completely independent. These conditions often share common molecular mechanisms, which may act in the same direction across disorders, or can be altered in opposite directions (i.e., being enhanced in one disorder while reduced in another).

Neuronal damage is a central process altered in CNS disorders. It implies axonal degeneration, dendritic spine loss, synaptic dysfunction, and ultimately neuronal death, leading to impaired neuronal connectivity[89,90]. In neurodegenerative diseases, this process constitutes a major hallmark, as neuronal injury and the consequent neuronal death occur continuously over time[91]. Meanwhile, in injury-related conditions acute damage arises in response to traumatic events[92]. Infections affecting the CNS, such as viral encephalitis or bacterial meningitis, can also produce extensive neuronal injury through direct effects[93]. Furthermore, in primary tumors and metastases, neurons may be compromised either by direct mechanical compression from tumor cells, or indirectly through tumor-induced mechanisms such as excitotoxicity and interactions with the CNS cell types[94].

Another example of shared mechanism is neuroinflammation. It involves the activation of the resident immune cells, together with the recruitment of peripheral immune cells. As a result, they release pro-inflammatory cytokines and chemokines within the CNS[95]. In neurodegenerative diseases, chronic neuroinflammation accelerates neuronal loss and disease progression[96]. Acute CNS injury triggers a robust inflammatory cascade that contributes to secondary injury and long-term deficits[97]. Similarly, CNS infections trigger severe inflammatory responses that compromise neuronal survival and may cause long-term neurological damage[98]. In the oncological context, neuroinflammation contributes to tumor progression, favoring inflammatory pathways to promote angiogenesis, immune evasion, and metabolic interactions of tumoral cells with neurons and glial cells[99,100].

Neuronal damage and neuroinflammation are two mechanisms that exemplify the interconnected nature of pathological processes across different CNS disorders. Consequently, studying one condition can provide valuable insights into others. In this scenario, this doctoral thesis also explores MS within the broader context of neurodegenerative diseases, with a particular focus on Alzheimer's disease (AD) and Parkinson's disease (PD). The ultimate goal of *Study III* is to explore the potential links between these three neurodegenerative diseases and melanoma brain metastasis (MBM) for a better understanding of its tumor biology.

Beyond the previously introduced MS pathology, the following subsections describe the three additional diseases under study.

1.2.1. ALZHEIMER'S DISEASE

AD is the most common form of dementia and represents one of the leading causes of disability and mortality worldwide of this century. Clinically, AD manifests as a progressive and incurable neurodegenerative disease characterized by cognitive decline, memory impairment, and brain deterioration that severely compromise the quality of life and the daily functioning of the patients, ultimately leading to their death[101]. The precise molecular mechanisms underlying its onset and progression remain incompletely understood. However, two neuropathological features define AD: the extracellular accumulation of amyloid-β (Aβ) plaques and the intracellular aggregation of hyperphosphorylated tau protein into neurofibrillary tangles (NFTs); both associated with synaptic dysfunction and neuronal loss[102]. Additional processes involved in disease progression are chronic neuroinflammation, oxidative stress, and mitochondrial dysfunction[103].

1.2.2. PARKINSON'S DISEASE

PD is the second most common neurodegenerative disorder after AD and the most prevalent movement disorder worldwide. Clinically, PD is characterized by bradykinesia (i.e., the slowness and difficulty in initiating voluntary movements) that is used as a diagnostic criterion. Typically, it is accompanied by resting tremors, rigidity, and postural instability. Beyond motor symptoms, patients frequently experience a broad spectrum of non-motor manifestations[104]. PD is defined by the progressive loss of dopaminergic neurons in the substantia nigra pars compacta —a critical region for regulating voluntary motor control— and by the intracellular accumulation of misfolded α-synuclein aggregates named Lewy bodies[105]. The underlying mechanisms remain incompletely understood; however, they are known to involve mitochondrial dysfunction, impaired protein clearance, and chronic neuroinflammation[106].

1.2.3. MELANOMA

Melanoma is one of the most common types of skin cancer due to an uncontrolled growth of melanocytes. Melanocytes are cell types specialized in the production of the pigment called melanin, which gives color to the skin and protects us from ultraviolet radiation[107]. Cutaneous melanoma is often detected by changes in pre-existing moles or the appearance of new pigmented lesions, typically characterized by asymmetry, irregular borders, color heterogeneity, and a morphology that changes over the time[108]. At the molecular level, melanoma is characterized by high mutational burden and dysregulation of key oncogenic pathways, which lead to uncontrolled proliferation, survival, and immune evasion[109]. Unlike other solid tumors, melanoma exhibits a high metastatic potential, with the ability to disseminate to organs including the lung, liver, bone, and brain[109]. Importantly, it presents high propensity to metastasize to the brain, which occur in up to 40–60% of patients with advanced disease[110].

1.3. OMICS TECHNOLOGIES

High-throughput technologies, commonly referred to as omics approaches, represent a broad set of methodologies designed to enable the large-scale acquisition of data related to specific molecular types of interest, such as deoxyribonucleic acid (DNA), ribonucleic

acid (RNA), proteins or metabolites. These approaches constitute a paradigm shift in scientific research. Unlike reductionist methods which typically focus on a limited number of variables, omics technologies generate unbiased and high-dimensional datasets, that enable the exploration of the molecular basis that underlie the conditions of interest[111].

This section provides an overview of the omics methodologies covered in this doctoral thesis. Specifically, we describe transcriptomic approaches based on microarray data, bulk RNA sequencing (RNA-seq), and single-cell RNA sequencing (scRNA-seq), as well as metagenomic analyses derived from 16S ribosomal RNA (rRNA) gene sequencing.

1.3.1. TRANSCRIPTOMICS

Transcriptomics is the discipline that allows the identification and quantification of RNA molecules within a biological sample, allowing researchers to determine the complete set of transcripts present at a specific time and under defined conditions. This collection of RNA molecules constitutes the transcriptome of the biological sample. Its significance relies on identifying and quantifying which portions of the genome are actively expressed under particular conditions, being crucial for characterizing molecular biological changes, such as disease-associated molecular alterations[112].

Starting with the extraction of total RNA molecules from the biological sample of interest, two main transcriptomic technologies can be employed: microarrays and bulk RNA-seq[111]. These technologies capture the average gene expression across all cells within a sample, providing information about the global transcriptional activity[113]. However, this approach inherently masks cellular heterogeneity, as it does not distinguish between different cell types or states within the sample. To overcome this limitation, scRNA-seq was developed to profile gene expression at the resolution of individual cells[114]. This advancement allowed for the identification of cell populations and the characterization of the corresponding cell-specific transcriptional programs, boosting the understanding of tissue complexity, which is of particular interest in diseases such as MS.

In the following paragraphs are described the details of each strategy, from the processing of the biological sample to the generation of *count matrices*. In these matrices, rows correspond to genes, columns to samples, and each row-column position represents the quantified abundance of the given gene in the corresponding sample. Once this data structure is obtained, researchers can perform different analyses depending on

their scientific objective. For instance, when analyzing samples subjected to different conditions, one common approach is to perform a differential gene expression analysis to identify changes in expression levels across the different experimental groups[113].

1.3.1.1. Microarrays

Microarrays represent one of the earliest high-throughput technologies developed for transcriptomic profiling (**Figure 1.7, left**). They consist of a solid surface, typically a glass slide, where thousands of single-stranded DNA probes are immobilized and distributed in an ordered pattern. Each probe corresponds to the sequence of a specific gene from the organism under study. Meanwhile, RNA molecules extracted from the sample are reverse transcribed into complementary DNA (cDNA) and labeled with fluorescence. When the labelled cDNA is added to the microarray, it hybridizes to its complementary probe, and the resulting fluorescence signal is detected and quantified using image analysis. The fluorescence intensity at each spot reflects the relative abundance of the corresponding transcript in the sample. Microarrays remain a valuable and cost-effective tool for transcriptome-wide analysis, particularly for organisms with well-annotated genomes. However, their utility is limited to probe design, as only transcripts represented on the microarray can be detected. Additionally, they present a reduced sensitivity for low-abundance transcripts or novel isoforms[115]. These limitations were overcome with bulk RNA-seq technology.

1.3.1.2. Bulk RNA-seq

Bulk RNA-seq is based on the massive sequencing of transcripts present in a sample (**Figure 1.7, right**). First, the isolated RNA from the sample is fragmented and converted into cDNA through reverse transcription. Then, a series of experimental steps is performed —ligation, amplification by polymerase chain reaction (PCR), etc.— to generate libraries, which are collections of cDNA fragments specifically prepared for sequencing. The libraries are loaded into a high-throughput sequencing platform called sequencer, which determines the ordered nucleotide sequence of several molecules simultaneously. The nucleotide sequence obtained from each molecule is referred to as a read. This automated process allows the large-scale acquisition of transcriptomic data from the entire sample[113,116].

The next steps in the analysis are performed at the bioinformatic level. The first step is to check the quality of the sequencing. Alignment algorithms are then implemented to

determine which gene (or transcript) each read corresponds to. To do this, the genome (set of genes) or transcriptome (set of transcripts) of the corresponding organism is used as a reference, in which the features are already annotated. Each read is then positioned on the reference sequence based on its similarity and assigned to the corresponding gene or transcript. If no references are available for the organism under study, the reads are joined in an orderly manner according to the sequence overlaps found (strategy called *de novo* assembly). Finally, the number of reads for each gene in each sample is quantified, obtaining as a result the raw count matrix[113,116].

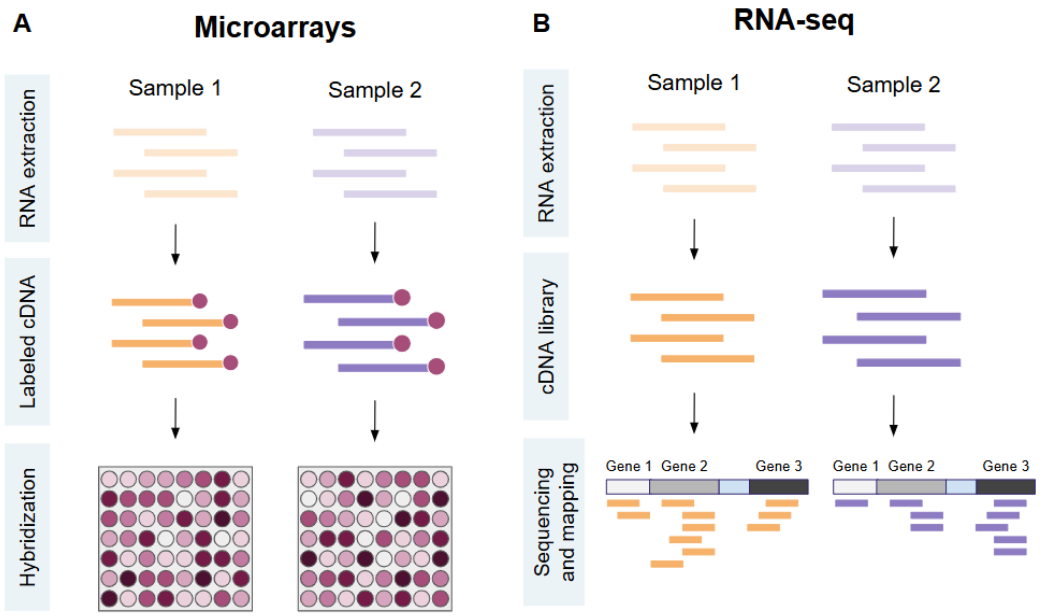


Figure 1.7. Illustration of microarray (left) and RNA-seq (right) technologies for transcriptome profiling. (Left) Microarray illustration: RNA is reverse transcribed to cDNA. The cDNA is labelled with fluorescence and hybridized to complementary probes immobilized on a solid surface, where fluorescence intensity reflects the relative abundance of each transcript. (Right) RNA-seq illustration: RNA is reverse transcribed to cDNA. High-throughput sequencing of the cDNA library generates short reads that are mapped to a reference genome; determining which gene each read belongs to. *cDNA: complementary DNA.*

1.3.1.3. Single-cell and single nucleus RNA-seq

scRNA-seq is based on the same principles as bulk RNA-seq, except that sequencing is performed at single-cell resolution. After sample collection, tissues must be processed to first dissociate and then isolate the cells. The feasibility of isolating intact single cells depends on the structural characteristics of the tissue. For easily dissociable tissues such as blood, viable individual cells can be isolated, enabling scRNA-seq. However, in complex tissues like the CNS, where cells are highly interconnected, cell dissociation often leads to plasmatic membrane damage. In this scenario nuclei are often isolated instead, and single-nucleus RNA sequencing (snRNA-seq) is performed, without sequencing cytoplasmic and mitochondrial RNAs. The overall bioinformatic workflow is equivalent in both cases[117]. Thus, when no distinction is needed throughout this doctoral thesis, we will use the terms *cells* and *scRNA-seq* to refer indistinctly to both scRNA-seq and snRNA-seq approaches.

Once isolated, individual cells are lysed to release their RNA, which is then captured and amplified for sequencing. Among the most commonly used platforms we found the *10x Genomics Chromium* system, which was the one applied in the datasets analyzed in this thesis (**Figure 1.8**). This technology employs a droplet-based microfluidic system to encapsulate individual cells, each in its own droplet. Within the droplet, transcripts are labeled with two types of barcodes: a cell barcode (to identify the cell of origin) and a unique molecular identifier (UMI), to distinguish biological transcripts from the technically amplified transcripts. These labels are added during the reverse transcription, generating the cDNA. The resulting cDNA molecules undergo amplification and library preparation, similar to bulk RNA-seq, and finally they are sequenced. After this process, reads are mapped to a reference genome obtaining the raw count matrix. Each position in the resulting matrix represents the number of distinct transcripts (identified by different UMIs) for each gene in each sequenced cell. This approach generates expression profiles for thousands to millions of individual cells (or nuclei) in a single experiment, enabling the identification of distinct cell types, transient cellular states, and rare subpopulations that would otherwise be masked in bulk analyses[118].

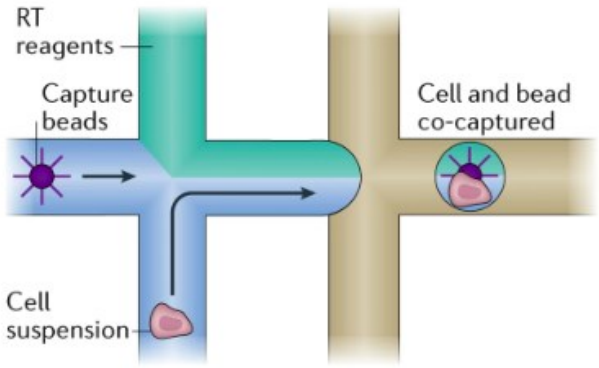


Figure 1.8. Illustration of cell isolation using a microfluidic droplet system. The device integrates microchannels that merge reagents and cell suspensions into an aqueous phase, which are then segmented into droplets by the controlled flow of an immiscible oil phase. Each droplet serves as an independent physical space for cell lysis, reverse transcription, and barcoding. Figure from Prakadan *et al.* 2017[118].

1.3.2. METAGENOMICS

Microorganisms are microscopic living entities including bacteria, archaea, viruses, fungi and protists. The majority of microbial species cannot be cultured in laboratories, leaving a substantial proportion of the microbial diversity uncharacterized. To address this limitation, sequencing strategies were developed, giving rise to the field of metagenomics. Metagenomics is the omics discipline that enables the identification and quantification of genetic material from microbial communities without prior cultivation[119].

Metagenomics was initially developed through marker gene sequencing approaches. These methods target genes that are conserved across specific groups of organisms, yet comprising enough variability to discriminate among different taxa. Commonly, the 16S rRNA gene is used for bacteria and archaea; the internal transcribed spacer (ITS) regions for fungi; and the 18S rRNA gene for eukaryotic microorganisms such as protozoa. More recently, whole-genome shotgun (WGS) sequencing technologies were developed, enabling the identification and quantification of all genes present in a sample rather than focusing on single marker genes. This approach involves high-throughput sequencing of entire microbial communities through random fragmentation of the extracted DNA, followed by sequencing and computational assembly to generate a complete microbial representation of the samples[120].

In this doctoral thesis, we analyze data derived from 16S rRNA sequencing of the human gut microbiota. The 16S rRNA gene is a highly conserved component of the small subunit of prokaryotic ribosomes, being used as a gene marker for identification and quantification of bacterial and archaeal taxa. This gene comprises approximately 1,500 base pairs that are subdivided in conserved and hypervariable regions (**Figure 1.9**). The conserved regions are sequences that do rarely change across diverse taxa, enabling the design of primers that can amplify segments of the 16S rRNA gene from a wide range of microorganisms. Meanwhile, hypervariable regions exhibit sequence variability to discriminate taxa. Typically, nine hypervariable regions (V1 to V9) are recognized. Different sequencing approaches target specific regions depending on the desired taxonomic resolution and the selected sequencing technology[121].

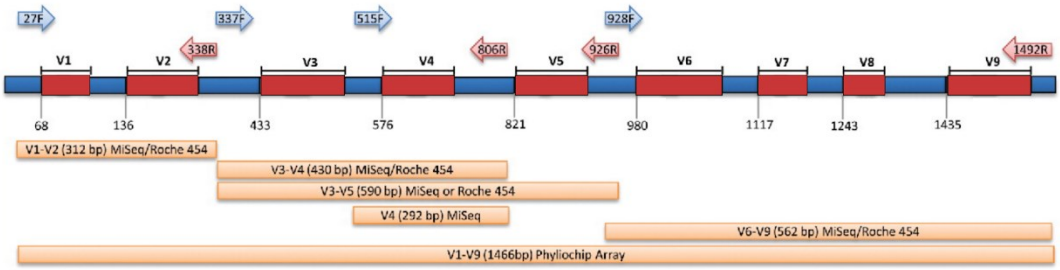


Figure 1.9. Schematic illustration of the 16S ribosomal RNA gene. Conserved and hypervariable regions are delimited with blue and red colors, respectively. Numbers indicate the starting positions of each hypervariable region. Blue arrows represent forward primers, and red arrows represent reverse primers used for paired-end sequencing. Regions highlighted in orange correspond to the amplified regions, indicating their approximate lengths, and the commercial sequencing platforms that provide these services. *F: forward; R: reverse.* Figure adapted from Shailesh K. Shahi *et. al*. 2017[121].

16S rRNA-based metagenomic studies typically begin with sample collection, followed by the extraction of microbial DNA using physical, chemical, or combined lysis methods to recover the genetic material from the microbial communities. The 16S rRNA gene is then amplified using the previously defined primers (**Figure 1.9**). Similar to RNA-seq protocols, several DNA processing steps are performed prior to sequencing. After sequencing, the resulting reads represent short nucleotide sequences corresponding to 16S rRNA gene fragments from the microorganisms present in the sample, which are then processed to generate the count matrix[120].

1.3.3. REUSING OMICS DATA TO ANSWER NEW BIOLOGICAL QUESTIONS

Omics approaches support *in silico* analyses, enabling the reuse of existing datasets to address new biological questions. In recent years, part of the scientific community has advocated for data sharing, as it enhances the reproducibility of published findings, facilitates independent validation of results, and accelerates the dissemination of scientific knowledge[122]. In 2016, a group of researchers proposed the FAIR principles (from *Findability, Accessibility*, *Interoperability*, and *Reuse*) as a practical framework to promote and guide data sharing. These principles not only encourage open data practices but also provide guidelines to ensure that shared data can be easily accessed, understood, and reused by other researchers[123]. Moreover, international organizations actively promote data sharing as part of broader open science strategies. The European Union established open science policies since Horizon 2020, and entities such as UNESCO provide recommendations to foster open and collaborative research practices[124].

Springer Nature, in collaboration with Digital Science and Figshare, regularly publishes *The State of Open Data* reports, which survey researchers worldwide across diverse disciplines. The most recent editions, released in 2023 and 2024, highlight an increasing trend in the percentage of researchers sharing their data. Their main motivations include compliance with the funder requirements to share the data, but also the promotion of transparency and reproducibility in research, and the potential to increase the visibility and impact of their work. Conversely, the main barriers to data sharing remain the lack of adequate training, concerns about privacy and data protection, and the limited incentives to support open science practices[125,126].

A major advantage of data sharing is that omics datasets generated to address a specific research question can be reanalyzed to explore new hypotheses beyond the original scope of the study. Among the benefits we find that it eliminates the need to repeat costly and invasive experiments. It also facilitates the integration of multiple datasets, increasing the statistical power, and ultimately leading to more robust and reproducible findings. However, data sharing also presents multiple challenges. In studies involving human subjects, the ability to share data may be restricted by national privacy protection laws, even when data are anonymized. Moreover, there is a lack of standardization in both data and metadata formats. As a result, we can identify inconsistencies in the processing state of shared data (e.g., whether the count matrices are raw or normalized), incomplete documentation and missing variables, unclear definition of variable units, and unreported technical or biological biases known only to the original authors[127–130].

This doctoral thesis was conducted entirely through the reuse of publicly available data. The following sections detail the resources employed to identify relevant datasets, along with the principles applied to ensure their proper identification and selection.

1.3.3.1. Scientific databases

Scientific databases are organized digital repositories designed to store and provide access to large volumes of structured data. In the omics field, these databases serve as infrastructures for storing high-throughput datasets and their associated metadata, which may describe experimental design, sample characteristics, and processing methods. Depending on the type of omics and the type of data structure, we can find different specialized databases.

For transcriptomic data, the most widely used repositories are Gene Expression Omnibus (GEO) and ArrayExpress, currently integrated in BioStudies. The GEO database is created and maintained by the United States National Centre for Biotechnology Information (NCBI). It provides a comprehensive archive for raw and processed data, as well as the corresponding metadata files[131]. Meanwhile, ArrayExpress was developed by the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI). Initially, it was designed as a repository for microarray datasets although it has been expanded to store sequencing-based data. Like GEO, researchers can retrieve raw data, processed data, and metadata files[132].

On the other hand, there are databases specifically designed to store raw sequencing data, which represents the most common format for metagenomic datasets. The two most widely used repositories are the Sequence Read Archive (SRA) from the United States and the European Nucleotide Archive (ENA) in Europe. The SRA was developed by the NCBI as part of the International Nucleotide Sequence Database Collaboration (INSDC). Using the unique identifier of a study of interest, researchers can retrieve datasets and metadata through tools such as the SRA Run Selector[133]. The ENA serves as the European counterpart for sequence storage and is maintained by EMBL-EBI, also within the INSDC framework. Since 2016, ENA has published annual reports in *Nucleic Acids Research* detailing its status, impact, updates, and ongoing developments[134].

In addition to these scientific databases, omics data and metadata records are often shared through alternative strategies. These include:

- Supplementary materials directly linked in the original publications.

- Institutional or laboratory websites, where research groups present the data associated with their projects.
- Interactive web tools designed to facilitate data exploration and visualization of the results from a specific publication.
- General data-sharing platforms such as Zenodo (https://zenodo.org, last accessed 28 July, 2025) or Figshare (https://figshare.com, last accessed 28 July, 2025). They allow datasets to be stored with an associated *Digital Object Identifier* (DOI), ensuring proper citation and long-term accessibility.
- Specialized databases, like the University of California Santa Cruz Cell Browser (UCSC Cell Browser) for scRNA-seq data.

Finally, in some publications it is specified that data and metadata files are *available upon request*, requiring contact with the corresponding authors to obtain access.

1.3.3.2. Systematic review

A systematic review is conducted to identify the datasets that would be analyzed to address the scientific question of interest. Systematic reviews involve an in-depth structured screening of the literature to collect the complete scientific evidence available on a specific topic. In this doctoral thesis, the systematic review process was implemented following the PRISMA (from *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) guidelines[135,136].

PRISMA guidelines establish a framework to promote standardization and reproducibility in the identification of studies for new analyses. The process begins with a concise formulation of the research question. Once defined, we determined the inclusion criteria, which delineated the mandatory characteristics that the eligible datasets must meet. Next, an exhaustive search is conducted in multiple databases and public resources, filtering according to the previously established requirements (identification phase). The identified studies are individually evaluated to ensure that there is no reason for exclusion (screening and eligibility phases). Finally, the data and metadata from the selected studies are downloaded, subject to confirmation of their accessibility, the quality and the status of the provided data (inclusion phase)[137,138]. The recommendation to visualize the results is through a four-phase PRISMA flow diagram, as illustrated in **Figure 1.10**.

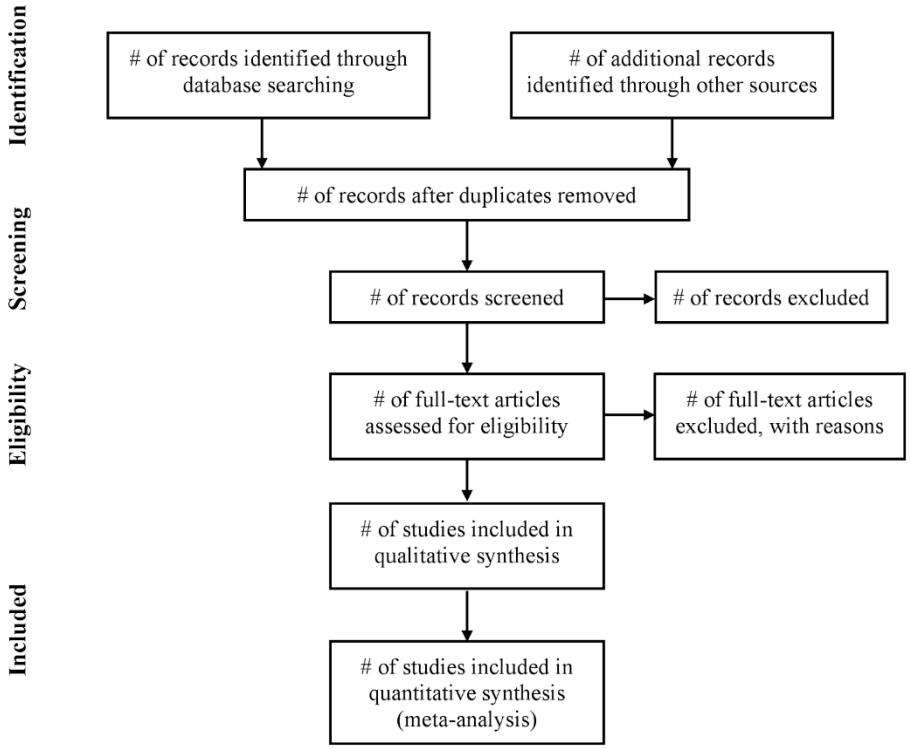


Figure 1.10. Flow diagram to recapitulate the systematic review results in accordance with the PRISMA statement. Specification of the number of studies (#) retained at the identification, review, eligibility and inclusion phases, as well as the corresponding reasons for exclusion. Figure from Liberati *et al.* 2009[135].

PRISMA guidelines were followed in all the projects presented in this doctoral thesis. Whenever sufficient datasets were available, a meta-analysis approach was applied. However, in instances where data were limited, individual analyses were conducted instead.

1.3.3.3. Meta-analysis

Once the systematic review has been completed, the sets of data that meet the predefined criteria for analysis are retrieved. After individually processing and analyzing each dataset, the results may differ across studies even when the experimental groups are comparable. These discrepancies can be attributed to a variety of sources, such as

varying sample sizes, differences in experimental protocols, laboratory instrumentation, sequencing platforms and the computational and statistical methodologies. Meta-analysis strategies can be employed to synthesize and integrate the findings from multiple independent studies. Thus, meta-analysis is a statistical approach that enables researchers to combine results from several studies addressing the same research question, thereby providing consensus profiles that account for variability among individual studies[139,140].

In the context of this doctoral thesis, we performed meta-analyses based on the combination of effect sizes from individual studies. The effect sizes measure both the extent and the direction of change of a variable in a given comparison. For instance, a standard effect size measure in differential expression analyses is the logarithm of fold change (logFC) value, where the absolute value reflects the extent of the change and the sign indicates the group in which the corresponding transcript is more abundant.

The outcome of the meta-analysis includes, among other metrics, the combined effect size along with its confidence interval and its statistical significance. This combined effect size represents a consensus estimate that integrates evidence across individual studies, providing a more robust and comprehensive understanding of the underlying biological signals[141,142].

2. Motivation and objectives

Diseases affecting the CNS are complex conditions influenced by multiple factors, including host-related characteristics. Sex is a major determinant, as it modulates the disease onset, clinical manifestations, and progression of CNS disorders. This thesis is motivated by the need to better characterize the molecular mechanisms underlying sex-related differences in MS, where sex plays a critical role as outlined in the *General Introduction* chapter.

Although each CNS disorder has its own clinical and pathological identity, the processes involved may be shared across conditions. These mechanisms may converge within a single disease group, and also extend to other types of pathology. Within this framework, part of this thesis was inspired by emerging evidence that brain melanoma metastatic cells exhibit characteristics linked to neurodegenerative disorders[143]. This observation, together with previous clinical associations, led us to investigate the extent to which gene expression patterns may be shared between these two disease types.

Considering all the above, the major aim of this doctoral thesis is to advance the understanding of CNS-associated diseases by reusing publicly available omics data, exploring novel scenarios beyond those originally addressed in their associated publications. By doing so, this work aims to provide a better understanding of the molecular features and mechanisms potentially involved in the pathophysiology of CNS disorders. The two research axes are approached through specific objectives developed across three independent studies:

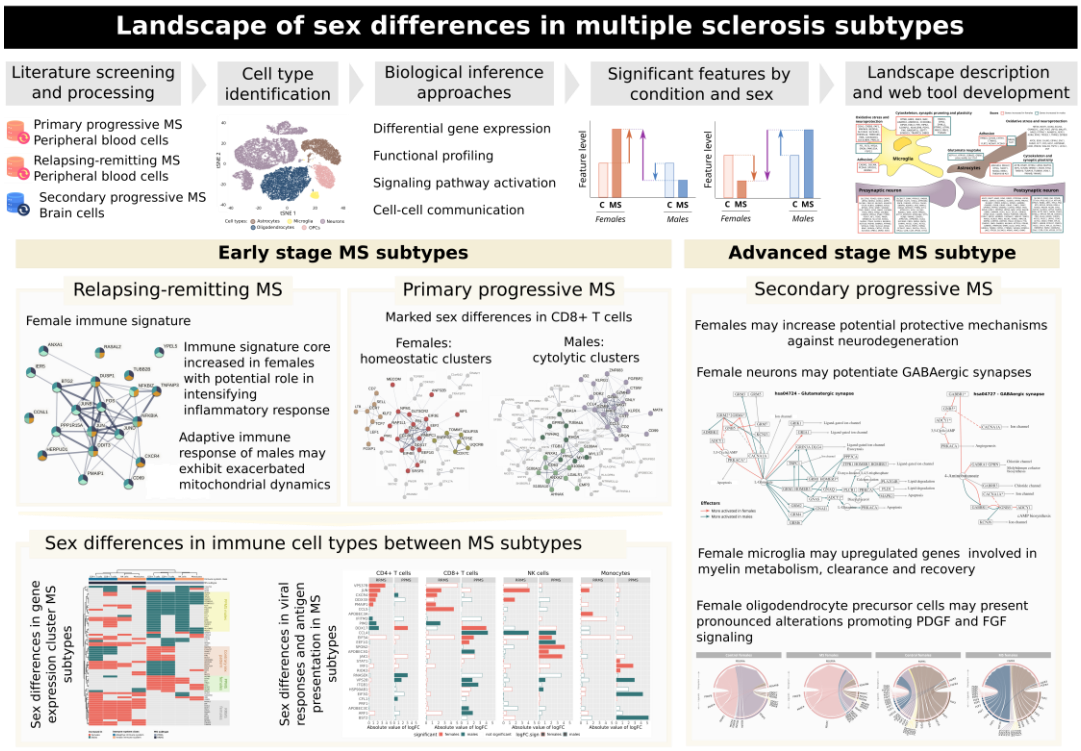
1. Study I: to identify and characterize sex-related differences across MS subtypes by generating cell type specific landscapes of gene expression, functional profiles, signaling pathways, and intercellular signaling communication through the analysis of scRNA-seq data.

2. Study II: to detect and explore sex-related differences in the gut microbial composition of MS patients analyzing 16S metagenomic data.

3. Study III: to uncover shared transcriptomic alterations and associated biological functions between melanoma brain metastases and neurodegenerative diseases.

The three studies are organized into separate chapters. Each chapter begins with the background section and the formulation of the study objectives, followed by the description of how we retrieved the data and the methodological approaches employed. Results are subsequently presented and finally discussed. This manuscript concludes with a final section that integrates the findings from all chapters, the conclusions of this work, and the scientific contributions generated throughout this doctoral period.

3. STUDY I: Single cell landscape of sex differences in the different courses of multiple sclerosis



This chapter was adapted from the preprint *Single cell landscape of sex differences in the progression of multiple sclerosis* by Irene Soler-Sáez, Borja Gómez-Cabañes, Rubén Grillo Risco, Cristina Galiana Roselló, Lucas Barea Moya, Héctor Carceller, María de la Iglesia-Vayá, Sara Gil Perotín, Vanja Tepavčević, Marta R. Hidalgo and Francisco García-García.

https://www.biorxiv.org/content/10.1101/2024.06.15.599139v1

3.1. INTRODUCTION

3.1.1. MOLECULAR MECHANISMS UNDERLYING MULTIPLE SCLEROSIS IN THE CENTRAL NERVOUS SYSTEM

The main pathological hallmark of MS is the damage that occurs to myelin sheaths that surround neuronal axons. For additional context, refer to the *General Introduction* of this thesis. Rather than being a specific alteration, myelin damage involves complex multi-faceted processes. In this work, we focus on three key aspects: excitotoxicity, cellular stress responses, and mechanisms for myelin clearance and repair.

The demyelination process disrupts the synaptic transmission leading to synaptic dysfunction[144]. To understand the synaptic alterations in MS, we first need to describe the general structure and functionality of synapses (**Figure 3.1**). Synapses are specialized junctions through which neurons communicate with each other in the presence of glial cells (astrocytes and microglia). The neuron transmitting the signal is referred to as the presynaptic neuron, while the one receiving the signal is the postsynaptic neuron. The narrow gap between them is known as the synaptic cleft. At the synapse, the presynaptic neuron releases chemical molecules called neurotransmitters, which are recognized by the postsynaptic neurons via membrane receptors. After this interaction, the chemical signal is transformed into an electrical signal, known as the action potential. This signal propagates along the axon through sequential depolarization and repolarization waves, mediated by the opening and closing of voltage-gated sodium (Na^+) and potassium (K^+) channels. Once the action potential reaches the axon terminal, depolarization induces the opening of calcium (Ca^{2+}) channels. The influx of Ca^{2+} triggers the fusion of synaptic vesicles (small sacs containing neurotransmitters) with the plasmatic membrane. Neurotransmitters are consequently released into the synaptic cleft by exocytosis to transmit the signal to the next neuron. Based on the neurotransmitter released, the signal transmission on the postsynaptic neuron is modulated. The effect on the postsynaptic neurons can be excitatory (e.g., glutamatergic synapses), inhibitory (e.g., GABAergic synapses) or modulatory (e.g. dopaminergic synapses)[145].

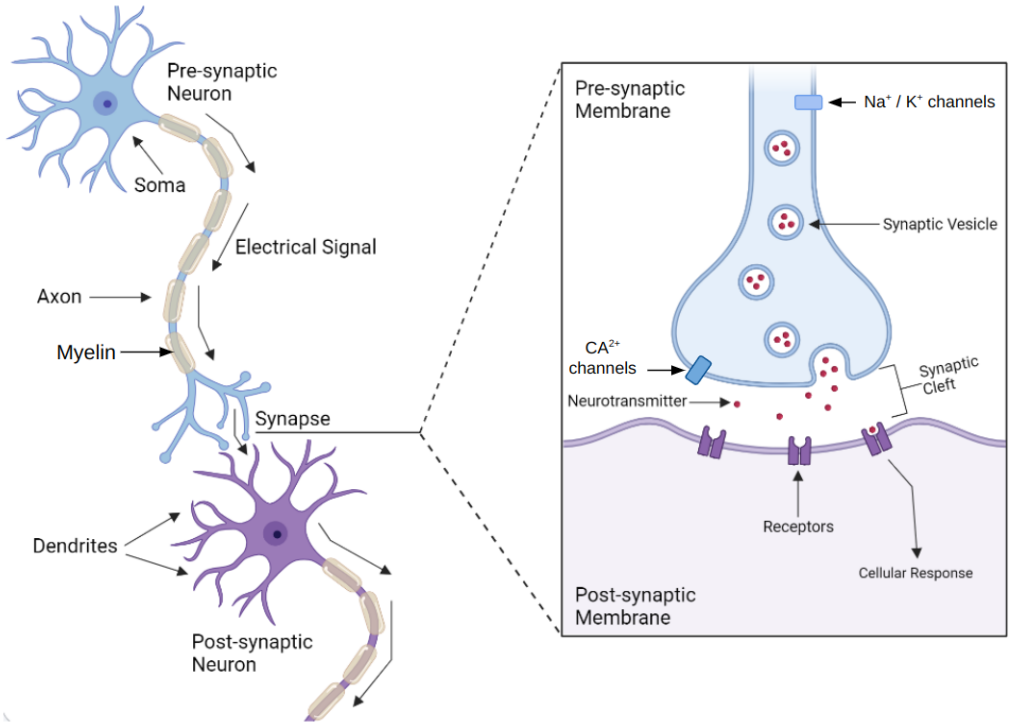


Figure 3.1. Illustration of a generic neuronal synaptic transmission. (Left) Architecture of synaptic communication between a presynaptic (blue) and postsynaptic (purple) neuron. (Right) Detail of the synapse specifying components of the presynaptic and postsynaptic neurons. Figure adapted from the *Biorender* template created by Genesis Cedeno and titled *Neuron Synapse*.

The axonal degeneration and synaptic loss characteristic of MS[146] are associated with a persistent state of neuronal hyperexcitability. Among the contributing mechanisms, we found imbalanced ion homeostasis. Different studies in model organisms pointed to alterations in K^+, Na^+ and Ca^{2+} channels[147–149], that favor Ca^{2+} accumulation at the presynaptic terminals. This situation is accentuated by an excess of glutamate release into the synaptic cleft[150]. Glutamate is the main excitatory neurotransmitter in the CNS, essential for physiological synaptic transmission, plasticity, and cognition. However, when its extracellular concentration is highly disproportionate it becomes toxic, generating a phenomenon known as excitotoxicity. In MS, this event has been associated to increased intracellular accumulation of Ca^{2+}, dysregulation of glutamate receptors[151], mitochondrial dysfunction, and oxidative stress, which collectively contribute to synaptic degeneration and neuronal death[152]. Moreover, elevated brain glutamate levels have been correlated with accelerated brain atrophy and faster disability progression[153]. Despite being less well characterized, impaired GABAergic synapses —the main

inhibitory neurotransmitter in the CNS— have also been observed in model organisms for the disease.

Excitotoxicity not only causes damage in the affected synapses, as it is expanded to neighboring neurons and glial cells. These glial cells exhibit a dual role in the disease, balancing neuroprotective and neurotoxic responses in an environment characterized by oxidative stress, hypoxia, and neuroinflammation[157–159]. At the affected synaptic clefts, astrocytes can secrete compounds that exacerbate synaptic dysfunction by accelerating the neuronal accumulation of Ca^{2+} and the loss of glutamate uptake capacity, contributing to excitotoxicity. Moreover, astrocytes undergo a process known as reactive astrogliosis, characterized by morphological and functional changes under permanent stress signals. This leads to the formation of a glial scar in the MS lesion area, which acts as protective mechanism although it hinders remyelination and axonal regeneration[160,161]. Astrocytes also support myelin repair by recruiting microglia in the lesion areas[162]. Although microglia contributes actively to inflammation through the release of reactive oxygen species, it also presents beneficial phagocytic capacity to clear myelin debris from damaged axons. Microglia can also support the myelin repair process by producing trophic factors, promoting the survival and maturation of OPCs[163,164].

As previously described, oligodendrocytes are specialized cells responsible for the formation and maintenance of myelin sheaths in the CNS. The loss of the myelin sheath not only impairs the propagation of the action potentials along the axon, but also disrupts the metabolic support that oligodendrocytes provide to neurons, exacerbating axonal damage and neurodegeneration[165]. OPCs are intended to repair this damage. This population maintains a proliferative and migratory capacity that enables them to differentiate into mature oligodendrocytes. However, the process is highly complex and tends to fail in MS patients. Firstly, OPCs must be recruited in the lesion where the injury is located. However, neuroinflammation, oxidative stress and the glial scar may cause insufficient or delayed recruitment of OPCs. Furthermore, they do not always receive the necessary signals that promote differentiation, which delay the repair process and the area becomes too damaged to be repaired[166]. As with other cell types, oligodendrocytes also have a dual role where they can promote and mitigate neuroinflammation[167].

3.1.2. ROLE OF THE PERIPHERAL IMMUNE SYSTEM IN MULTIPLE SCLEROSIS

Although the etiology of MS remains unknown, it is well established that CNS damage is mediated by autoreactive subsets of cells from the peripheral immune system. The scientific community has not reached a definitive consensus on when the immune self-response is initiated. However, the most supported hypothesis points to the loss of immune tolerance in the periphery, not in the CNS. As a result, the autoreactive cells would cross the blood-brain barrier and directly contribute to neuroinflammation, demyelination, and axonal injury in the CNS[18].

The presence of autoreactive cells in MS implies that the immune system is aberrantly activated upon recognizing self-antigens as foreign compounds to be eliminated. Such process is initiated through the antigen presentation event (**Figure 3.2**). Antigen presenting cells fragment proteins into smaller peptides, called antigens, that are presented to T cells. Then, the T cells assess whether these antigens originate from self-proteins and should be not targeted, or whether they derive from foreign sources and require the activation of the immune response[168].

Cells present antigens through the major histocompatibility complex (MHC), a set of proteins encoded by HLA gene family. HLA genes (and the corresponding MHC) can be classified into class I and class II. Specifically, MHC-I class is detected on the surface of all nucleated cells, including the PBMCs. They present internal peptides to CD8+ T cells, allowing the detection of internal molecules from viral infections or oncogenic transformation. On the contrary, MHC-II class is primarily located on professional antigen-presenting cells (dendritic cells, monocytes and B cells). These cell types present to the CD4+ T cells exogenous peptides derived from phagocytosed elements (e.g., bacterial infection). Genome-wide association studies have identified more than 100 single nucleotide polymorphisms associated with MS[169], where HLA genes from both class I and class II presented several associations. Notably, some of them were linked with an elevated risk of suffering MS, while others appeared to confer a protective effect[170].

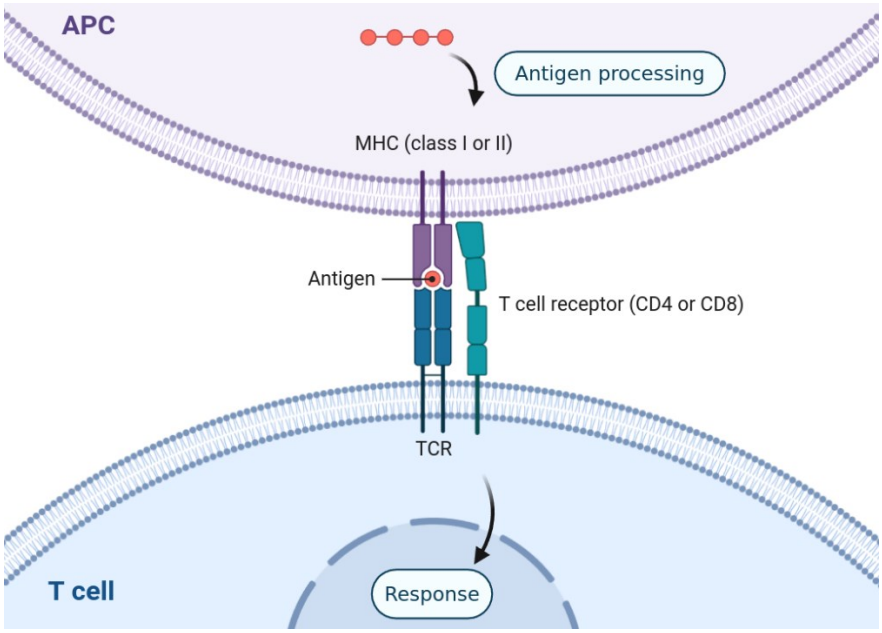


Figure 3.2. Illustration of a generic antigen presentation event. (Purple cell) Antigen-presenting cell processing proteins into peptides (antigens), that are presented on the cell surface via the major histocompatibility complex molecules of class I or II. (Blue cell) T cell that recognizes the antigen via the T cell receptor complex, with the assistance of co-receptors, triggering downstream signaling that can lead to an activation of the immune response. Interactions: CD4+ T cells for MHC-II, CD8+ T cells for MHC-I. *APC: antigen-presenting cell; MHC: major histocompatibility complex; TCR: T cell receptor complex.* Figure adapted from the *Biorender* template created by Catherine C and titled *Autocrine Signal Loop*.

Another related mechanism associated with antigen presentation is the viral EBV infection. Specifically, individuals who developed mononucleosis, a symptomatic manifestation of EBV, were at higher risk of developing MS[171], while those who were never infected had a lower risk[172,173]. The most accepted hypothesis is that EBV-reactive B and T cells cross the blood-brain barrier and, due to molecular mimicry, they recognize and attack self-antigens from the myelin of the CNS[174,175].

The involvement of PBMCs in the course of MS is not limited to the failure of self-antigen recognition. These cells exhibit a complex cytokine profile that contributes to neuroinflammation. The inflammatory state not only induces CNS damage but also increases the blood-brain barrier permeability, facilitating the continuous infiltration of PBMCs into the CNS[176]. Additionally, immune cells from MS patients display mitochondrial impairment that leads to increased production of reactive oxygen species, exacerbating the pathogenesis[177]. Different studies examined the transcriptional profile

their PBMCs to identify molecular signatures associated with disease status, severity, and treatment efficacy. These profiles not only allowed the stratification of MS patients based on their molecular signatures and clinical features, but also differ from those observed in healthy individuals[178–180].

Current MS treatments focus on modulating or suppressing the immune response. For instance, interferon beta modulates cytokine production to promote a more anti-inflammatory state, while ocrelizumab is a monoclonal antibody that depletes B-cells from the peripheral blood[181]. Although these therapies do not cure the disease, they significantly reduce relapse rates and mitigate disease severity, particularly in the RRMS subtype. In PPMS, where relapse-remission cycles are not reported and neurodegeneration progresses more steadily, these treatments showed reduced efficacy[182]. However, the immune system also plays a significant role in progressive forms. Autoreactive immune cells were identified within CNS lesions, with CD8+ T cell abundance positively correlating with axonal damage[183]. Moreover, PBMCs transcriptomic profiles from PPMS also differ from those of healthy individuals[184,185].

3.1.3. INHERENT CHARACTERISTICS OF SINGLE CELL TRANSCRIPTOMIC DATA

The scRNA-seq technology enables the characterization of gene expression at single-cell resolution. Precisely, this ability to obtain the transcriptomic profile of individual cells exhibits inherent characteristics that distinguish scRNA-seq from other transcriptomic approaches, such as bulk RNA-seq and microarrays[114].

One of these characteristics is commonly designated as the *curse of dimensionality*. Each cell is represented by the expression of thousands of genes, resulting in large and complex matrices with thousands of cells and thousands of features. This high dimensionality substantially increases computational demands compared to bulk approaches, requiring considerable processing time and memory to perform tasks such as clustering and dimensionality reduction. Consequently, large-scale analyses frequently rely on high-performance computing infrastructure[186].

A second distinctive property of scRNA-seq data is its sparsity, as individual cells contain intrinsically small amounts of RNA. Moreover, due to technical limitations in capturing and sequencing, some transcripts fail to be detected (known as *dropout* effect). This large number of zeros complicates the distinction between true biological absence of expression and technical noise within each sequenced cell[187].

To address these challenges, specific bioinformatic strategies have been developed. Common approaches include filtering out non-informative genes to retain those that capture most of the biological variability, applying dimensionality reduction techniques to facilitate the visualization of transcriptional heterogeneity, and employing statistical frameworks specifically designed to handle zero-inflated distributions, among others[186,188].

3.1.4. BIOINFORMATIC STRATEGIES FOR ANALYZING SINGLE CELL RNA-SEQ DATA

For each identified cell population, we applied different bioinformatics strategies. One of the most widely used approaches in transcriptomic studies is differential gene expression analysis. Its aim is to compare the gene expression levels between two or more conditions; to identify genes whose expression significantly increases or decreases among the groups of interest. Unveiling changes in gene expression enables the discovery of potential biomarkers, and improves the characterization of the molecular mechanisms that may operate in the pathogenesis of diseases[189]. When we implement this strategy at cell type level, we can characterize how different populations contribute to the development of the different conditions, revealing both cell type–specific and shared transcriptional alterations[190].

Once differentially expressed genes are identified, we can explore the biological context in which these changes occur by implementing functional characterization approaches. Thus, the goal is to determine in which biological functions, processes or pathways are the dysregulated genes involved. To perform these analyses, we required two elements: i) curated databases that catalogue the associations between functions and genes, and ii) the statistical framework according to how we wish to evaluate the results from the differential gene expression analysis.

In this work we used the *Biological Processes* database of the Gene Ontology (GO) as the curated resource for functional annotation[191,192]. Functions are called GO terms, each corresponding to a set of annotated genes, with no limitation on the number of genes included. The relationships of GO terms are organized hierarchically, with higher-level terms describing broad concepts (parent terms), and more specific terms appearing lower in the hierarchy (child terms) (**Figure 3.3**). At lower levels of the hierarchy, the specificity increases, allowing for a more detailed description of biological processes. This structure is defined as a directed acyclic graph where each term is a node, and the edges represent their relationships. The most common relationships between terms are *part_of* and *is_a*. Importantly, genes annotated to a child term with these relationships

are automatically inferred to participate in all ancestor terms. This concept is called *Propagation of Gene Ontology Annotation*[193].

Regarding the statistical framework, we used the *over-representation analysis* (ORA)[194] on gene subsets of interest. The method evaluates whether specific functions are represented more frequently among the selected genes than would be expected by chance. For the functions that appear significant, we state they are overrepresented or enriched in the corresponding condition.

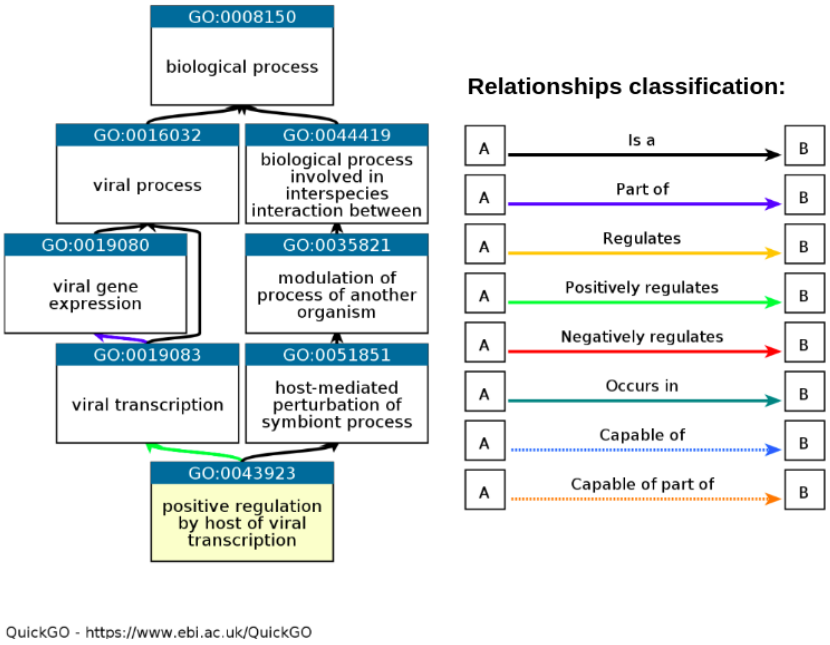


Figure 3.3. Gene Ontology graph for the term positive regulation by host of viral transcription (identifier GO:0043923). Nodes represent biological processes and edges the relationships between them. The child term is highlighted in yellow, being at the bottom of the hierarchical structure, while the parental terms are displayed at the top. Legend: relationship types that can exist between the generic term A (child) and the generic term B (parent), which establish the edge color in the graph. *GO: Gene Ontology.* Figure adapted from https://www.ebi.ac.uk/QuickGO/term/GO:0043923 and accessed May 5, 2025.

We also performed protein-protein interaction (PPI) analyses. In this approach, we construct a network in which nodes represent proteins (encoded by the selected genes) and edges denote known or predicted interactions between them. The database we used to determine the protein-protein associations is STRING[195,196], which includes functional (e.g., co-expression, co-regulation), physical (e.g., direct interaction) or inferred (e.g.,

literature mining) associations. These associations determine whether the proteins that comprise the network are connected. The statistical analysis assesses whether the observed interactions exceed those expected by chance, pointing to non-random biological interactions.

An additional bioinformatic approach consists of the evaluation of differential signaling pathways activation. Signaling pathways are molecular cascades primarily composed of proteins and/or few small molecules and metabolites. These cascades mediate the signal transmission from a stimulus that activates a receptor to a final protein effector that executes the biological function. Each step in the cascade sequentially activates or inhibits the next component of the pathway, often through post-translational modifications such as phosphorylation or glycosylation[197]. The database used in this work to report the signaling pathways is the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG represents signaling pathways through maps, where nodes represent the proteins and edges how the signal is transmitted along the pathway[198]. The effector protein points to the function associated with it. As an example, the apoptosis pathway is represented in **Figure 3.4**. Unlike ORA and PPI analyses, which inputs gene lists of interest, the computation of signaling pathway activation is performed directly from gene expression levels. Using a quantitative mathematical approach, we can assess the pathway activity in each sample. For comparisons across two or more conditions, differential activation analysis determines the condition with significantly increased (or decreased) pathway activity[199].

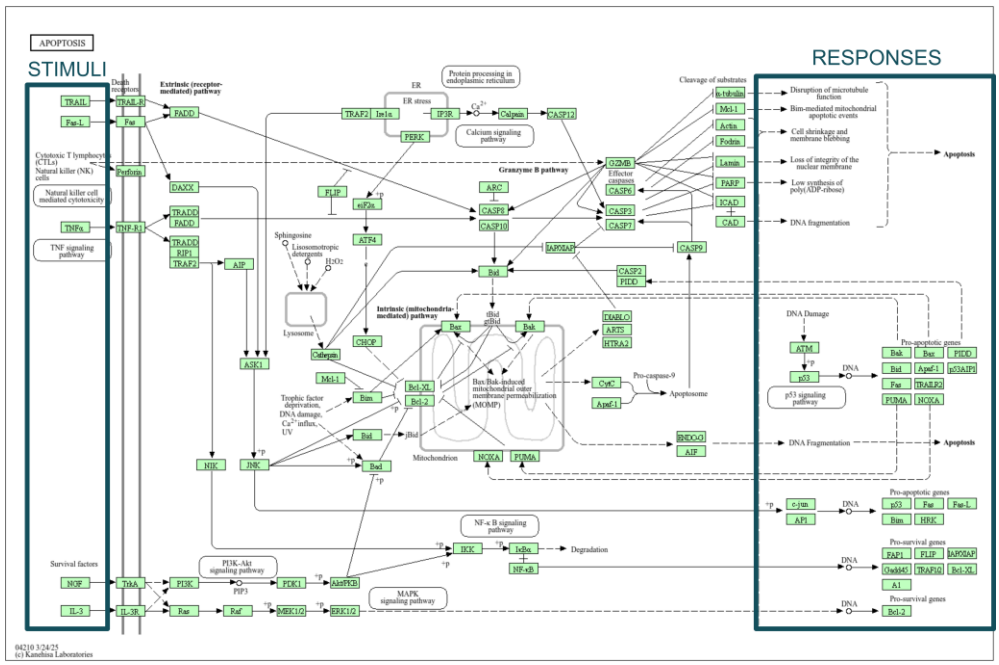


Figure 3.4. KEGG pathway map for the term Apoptosis (identifier: 04210). *(Previous page)* Nodes represent proteins, and edges indicate the interactions among the proteins. For clarity, the highlighted regions defined the area of the pathway where are located most of the stimuli (*stimulus* region), and most of the effectors with their associated functions (*responses* region). Figure adapted from https://www.kegg.jp/pathway/hsa04210 and accessed May 7, 2025.

The behavior of cells is constantly shaped by the signals they receive from the surrounding multicellular environment. These signals modulate intracellular processes such as gene expression and pathway activation in response to both physiological and pathological stimuli. Using bioinformatic pipelines, it is possible to infer cell–cell communication interactions based on the expression patterns of genes encoding ligands and receptors, both qualitatively (i.e., identification of the interaction) and quantitatively (i.e., the strength of the interaction). In these analyses, the sending cell is defined as the cell type that expresses the ligand and emits the signal, while the target cell expresses the receptor and executes the response. This approach enables not only the identification of potential signaling interactions, but also the comparison of interaction strengths across different conditions, characterizing changes in their cellular crosstalk[200,201].

3.2. CONTEXTUALIZATION, MOTIVATION AND OBJECTIVES

Sex contributes as an important variable to epidemiological and clinical aspects of MS. As discussed in the *General Introduction,* females tend to develop MS earlier, exhibit a higher prevalence, and suffer more severe inflammatory patterns[56]. In contrast, males experience more rapid CNS deterioration, with more significant atrophy[58]. These differences are reflected in the clinical subtypes of the disease. Females are more prone to suffer RRMS (with more relapses), while the progressive forms appear at a more balanced ratio between sexes[61]. Several studies seek to explain these differences by investigating the molecular mechanisms underlying MS pathology, primarily focusing on sex chromosomes and hormones. Of the characteristics explored in this doctoral thesis, the following paragraphs summarize the evidence on sex chromosomes and sex hormones, and, where known, their relationship with MS.

Excitotoxicity. Neurons with an XY chromosomal endowment, compared to XX, appeared to be more susceptible to injury under oxidative stress and excitotoxicity[203];

conditions that could resemble the MS environment. A recent review by Kniffin and Brian *et. al.* 2024[204] recapitulates how sex hormones modulate glutamatergic synapses. The removal of gonadal hormones in mice led to decreased synaptic activity in both sexes. When administering testosterone in males and estrogens in females, the activity of signal transmission is recovered[205,206]. Although fewer studies have focused on progesterone, current evidence suggests it may promote GABAergic transmission[204]. In model organisms of MS, male neurons were more vulnerable to the neuroinflammatory environment than female neurons, as male neurons exhibited greater axon damage and neuron loss[202]. Interestingly, our previous work from José Català-Senent *et al*. 2023[79] analyzing CNS bulk RNA-seq data identified the function *Glutamate metabolic process* was underrepresented in MS females, but not in males who seem to be more susceptible to excitotoxicity[79].

Glial cell response. Sex differences are known to influence neuroinflammatory responses within the CNS[207–209]. However, their impact on MS glial reactivity remains not fully understood. In MS model organisms, there is no consensus regarding whether astrocytes exhibit greater reactivity in males or females[210,211]. Nevertheless, some evidence suggests that male rodents may be more prone to develop reactive gliosis in response to MS-like injury and neuroinflammation[66]. This complexity is also reported in microglia responses to MS. Krementsov Lab identified that p38α signaling pathway confers neuroprotective effects in male microglia, but not in females[212,213]. Moreover, males tend to accumulate more microglia cells at the lesion borders[214,215]. These findings can be supported by studies in different brain areas of healthy mice, where higher microglial density is reported in males compared to females[216].

Myelin recovery. Some rodent studies have shown that, under healthy conditions, males exhibit a greater number of oligodendrocytes with an increased expression of myelin-associated proteins[217,218]. However, in response to demyelination injury, OPCs with XX chromosomes tend to exhibit higher proliferative capacity[69,219]. Sex hormones promote myelin recovery in both sexes[220]. In MS female model organisms, the administration of progesterone reduces disease severity; partly promoting myelin repair and oligodendrocyte differentiation[221]. Estradiol is also associated with beneficial effects, as correlates with decrease relapses preventing the loss of oligodendrocytes in lesion areas[222]. With a similar outcome, testosterone appears to prevent demyelination and promote myelin repair in MS male lesions[223].

Adaptive immune system. Sex also influences immune responses in MS, with both sex chromosomes and hormones modulating cytokine expression that determines the balance between proinflammatory and anti-inflammatory status[224]. Pro-inflammatory responses of CD4+ T cells may be exacerbated in MS females, while MS males tend to

exhibit a more immunomodulatory phenotype[225]. Moreover, a higher number of $CD4^+$ T cells were detected in the cerebrospinal fluid of females during relapses compared to males[226]. Regarding $CD8^+$ T cells, the sex-specific differences in MS are less well characterized. It is suggested that $CD8^+$ T cells are also influenced by hormonal regulation, where females tend to exhibit more coordinated responses and males appear to display markers of exhaustion[227]. The findings also remain inconclusive for B cells, as some studies reported a higher proportion of B cells in CNS lesions of female MS patients[228], while others have not found significant differences[229].

Innate immune system. Natural killer (NK) cells have been studied in the context of pregnancy, where an increase in NK cell activity is associated with reduced pro-inflammatory responses during MS, pointing to hormonal influence[230]. Meanwhile, sex differences in monocyte subpopulations were identified in naïve untreated patients[210]. Differentiated monocytes –called macrophages– located in the lesion areas reported increased iron accumulation in males, which could potentiate excitotoxic mechanisms that may contribute to the more rapid neurodegeneration[59].

Clinical courses and treatment outcomes. Tessa Dhaeze and colleagues conducted a study using XX and XY wild type mice that lacked their own immune system[231]. These animals were administered with immune cells derived from either female (XX) or male (XY) mouse models of MS. Notably, immune cells derived from female donors produced a phenotype resembling RRMS, while those from male donors were more similar to PPMS, regardless of the sex of the recipient animal[231]. Most clinical trials did not exhaustively explore sex-specific outcomes. Among those that consider the results separately by sex, some therapies show no significant sex-based differences. Others, like glatiramer acetate, reported greater efficacy in male patients[65].

Considering the current state of the art, sex plays an important role influencing different aspects of MS pathophysiology. However, its characterization is challenging due to the multifactorial nature of the processes involved. Most studies tend to focus either on the contribution of sex chromosomes, the sex hormones or in isolated genes and proteins. To our knowledge, there is a lack of characterization of sex differences in MS from a holistic perspective, considering simultaneously the different cell populations involved in the disease. Thus, the main objective of this chapter is to contribute to a better understanding of the molecular mechanisms underlying sex differences in MS. To achieve this, we develop an *in silico* strategy analyzing scRNA-seq datasets, defined by the following specific objectives:

1. To identify scRNA-seq and snRNA-seq datasets that explore MS, selecting them based on predefined inclusion and exclusion criteria, and retrieving eligible datasets from the corresponding repositories.

2. To process each selected dataset individually through bioinformatic workflows to obtain processed and cell-type annotated data suitable for downstream analyses.

3. For each MS subtype and cell type, to perform computational biological inference strategies addressing the following questions:

    a. Which genes are differentially expressed between sexes? (Differential gene expression analysis).

    b. Which biological functions are dysregulated in a sex-dependent manner? (Functional profiling analyses).

    c. Which signaling pathways show sex-differential activation patterns? (Signaling pathway analysis).

    d. Do patterns of intercellular communication differ between sexes? (Cell-cell communication analysis).

4. To interpret and contextualize the results from objective 3, identifying molecular mechanisms that may underlie the sex-differential pathophysiology.

5. To develop an open-access, user-friendly web resource for dissemination of the complete results to the scientific community.

3.3. MATERIALS AND METHODS

3.3.1. WORKFLOW DESCRIPTION

This project was conducted entirely *in silico*, analyzing publicly available scRNA-seq datasets that were composed of the raw count matrices and the associated metadata. The procedures prior to the count matrix generation can be consulted in the *General Introduction* chapter.

A comprehensive literature screening was performed to identify studies eligible for analysis (**Figure 3.5, yellow box**). Using these raw count matrices as input, we conducted individual bioinformatic analyses for each selected dataset (**Figure 3.5, blue boxes**). To ensure consistency across these studies, we applied a standardized analysis pipeline, while also accounting for dataset-specific characteristics when necessary[188,232,233]. Data were analyzed based on the following steps: i) preprocessing, ii) quality control filtering, iii) normalization, iv) highly variable genes selection, v) dimensionality reduction, vi) description of sources of variability, vii) cell identity clustering, and viii) cell type annotation with gene marker evaluation. Once cell types were annotated, we applied downstream biological inference approaches to address the biological research questions (**Figure 3.5, green boxes**). In detail, we explored statistical changes in gene expression patterns, biological processes, PPI networks, signaling pathways, and cell-cell communication interactions. Results were obtained for each cell type in females, males, and the sex-differential comparison.

The coming sections comprised how datasets were identified and a detailed description of each computational step previously mentioned. The bioinformatics code was developed using R (version 4.1.2) and can be found freely available at https://github.com/IrSoler/cbl-atlas-ms. **Supplementary Table 3.S1** provides a record of the versions of corresponding R packages.

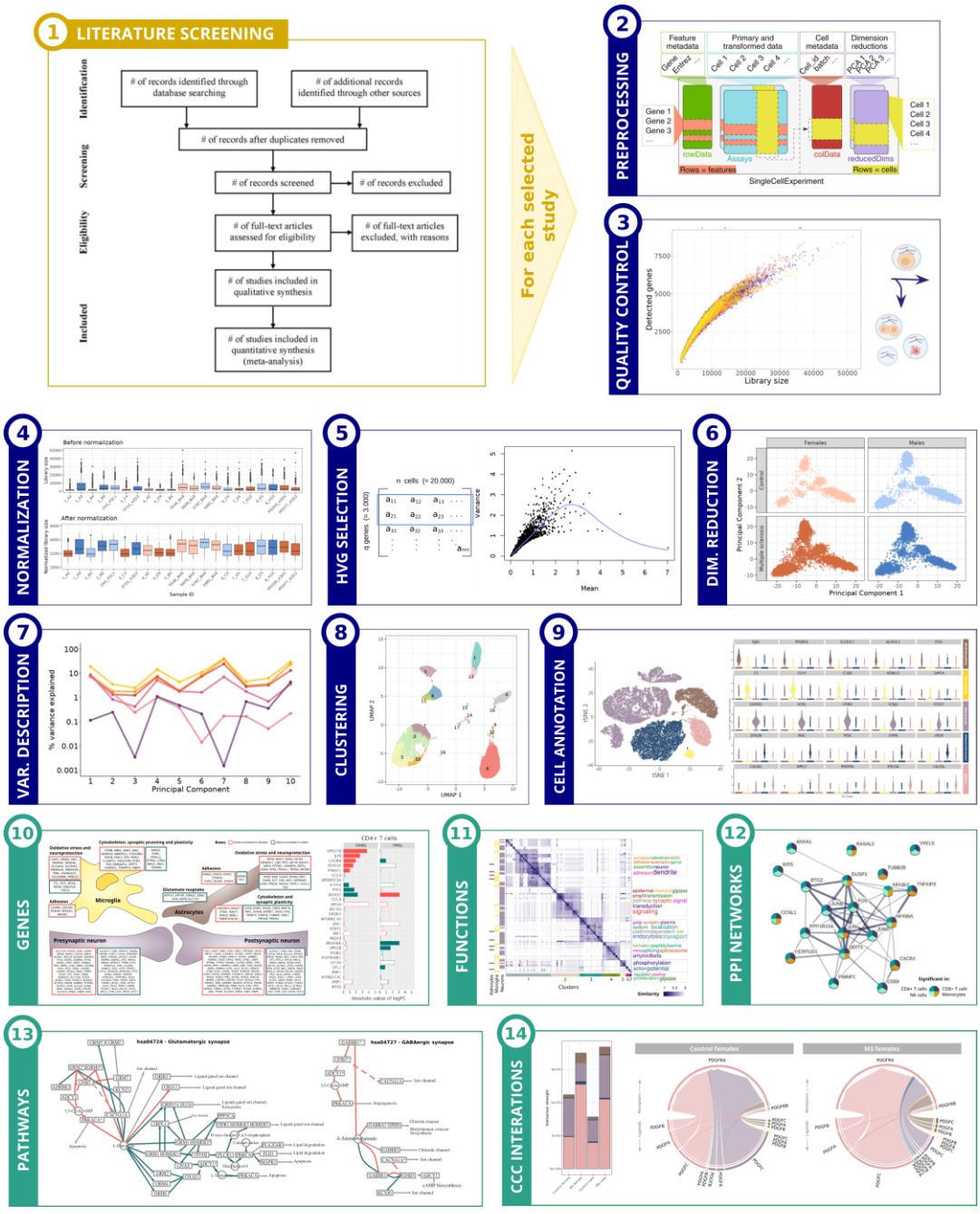


Figure 3.5. Workflow of this research. Steps followed in the scRNA-seq data analysis: literature screening in yellow box, computational processing of cells in blue boxes and biological inference approaches in green boxes. *CCC: cell-cell communication; DIM: dimensionality; HVG: high variable gene; ID: identifier; PPI: protein-protein interactions; VAR: variable.*

3.3.2. LITERATURE SCREENING

Scientific literature screening considering PRISMA guidelines was conducted until April 2022, in the public repositories GEO[234], ArrayExpress[235] and UCSC Cell Browser[236]. PubMed was also used to screen peer-reviewed publications with associated datasets.

The inclusion criteria were adapted to the layout of each repository:

- For GEO and ArrayExpress, an advanced search was performed using the keywords *multiple sclerosis* and [*single cell* or *single nuclei* or *single nucleus*], and *Homo sapiens* was selected as the organism of interest. In the case of GEO, the study type was indicated as *Expression profiling by high throughput sequencing*, while in ArrayExpress *RNA assay* and *sequencing assay* were selected in the technology section.
- For the UCSC Cell Browser platform, the list of available datasets was reviewed by selecting those addressing MS in humans.
- Finally, the keywords *multiple sclerosis* and *scRNA-seq* or *snRNA-seq* were introduced in the PubMed browser. The resulting literature was reviewed to identify studies with associated data available.

After this step, each of the selected studies was manually reviewed, discarding those that met any of the following exclusion criteria:

- Methodology: studies that did not generate scRNA-seq or snRNA-seq omics data.
- Disease examined: studies not based on MS.
- Experimental design: studies that included only control individuals or only MS patients.
- Sex information not reported.
- Sample size: datasets generated with fewer than three different individuals per condition and sex (control females, MS females, control males, MS males).
- Unavailability of the gene expression matrix or metadata files.

The raw count matrices and metadata files from the final selected studies were downloaded and stored in the Centro de Investigación Príncipe Felipe (CIPF) computational infrastructure. A detailed description of the included datasets is provided

in the *Results* section. However, to facilitate understanding of the subsequent bioinformatic analyses, we briefly outline that three datasets were analyzed, labelled in this work as: SPMS-CNS (scRNA-seq dataset with CNS samples from SPMS), RRMS-PBMCs (snRNA-seq dataset with PBMCs samples from RRMS), and PPMS-PBMCs (snRNA-seq dataset with PBMCs samples from PPMS).

3.3.3. STANDARDIZATION OF GENE AND GROUP NOMENCLATURE

After downloading the raw count matrices from each study, we verified that all of them used the official HGNC (HUGO Gene Nomenclature Committee) gene symbols[237] as primary identifiers. Experimental group labels were also standardized combining both condition and sex variables, resulting in four distinct groups: control females, MS females, control males, and MS males.

3.3.4. QUALITY CONTROL FILTERING

Not all profiles obtained after sequencing correspond to single, viable, and intact cells, as technical artifacts may occur[238,239]. In this work, we analyzed datasets generated with the droplet-based *10x Genomics Chromium* platform, where we may find:

- **Empty droplets:** droplets that do not contain cells, only capturing environmental RNA.
- **Failure to maintain cellular integrity (damaged cells):** during tissue dissociation and cell isolation, mechanical or enzymatic stress may break the plasma membrane of the cell, causing RNA leakage that results in partial RNA loss.
- **Inefficient cell/gene barcoding, amplification and/or sequencing (partially processed cells):** droplets in which not all genes were detected and quantified as they are present in the cell. This may result in uneven gene detection, underrepresenting the true transcript abundance.
- **Multiplets:** droplets that may encapsulate two cells instead of one (doublets), three (triplets), etc. These can be:
    - Homotypic multiplets (cells of the same type), which artificially increase the counts of a single population.

- Heterotypic multiplets (cells of distinct types), creating hybrid expression profiles.

The primary objective of cell-based quality control is to retain only those columns of the count matrix that represent single viable cells. To identify whether cells should be discarded for any of the previously described reasons, we calculated the following metrics per cell: library size (total counts), number of expressed genes, and the mitochondrial gene ratio. Empty droplets, damaged and inadequately processed cells were expected to present lower values for the library size and number of expressed genes, as they had lost part (or all) of the RNA present in the cell. On the contrary, we expected higher values on these metrics for multiplets, since they combine the transcriptomic profiles of two or more cells[238,239].

To refine multiplet detection, we used the scDblFinder R function incorporated in the homonymous package[240]. In brief, the algorithm classifies cells into clusters based on their expression profiles. Then, it computationally generates artificial doublets and combines the expression profiles of cell pairs from the same cluster (homotypic artificial doublets) and across clusters (heterotypic artificial doublets). Finally, the transcriptomic profiles of the artificial doublets are compared with the original data. Cells with expression patterns similar to the artificial doublets are labeled as potential multiplets.

Another indicator of data quality is obtained by plotting the library size (X-axis) and the number of expressed genes (Y-axis). As a result, we expect a positive correlation, where the larger the library size, the more genes are detected. If the library size is large but few genes are detected, this could be indicative of environmental contamination or amplification biased towards particular RNA molecules. If we find the opposite situation, with a low library size but many detected genes, it could also be indicative of technical issues, e.g. sequencing process not completed.

To complement the library size and the number of expressed genes, elevated values of mitochondrial gene ratio suggest membrane damage, as the compromised cells often exhibit cytosolic RNA leakage due to increased permeability. Mitochondrial genes were identified by their names starting with “MT-”. As scRNA-seq preserves the RNA present in the nucleus, the cytoplasm and the mitochondria, we would expect higher fractions in viable cells compared to snRNA-seq, which only maintain the integrity of the nucleus.

The establishment of appropriate cutoffs for the quality control metrics is dependent on both technical factors (e.g., the implement protocol, the sequencing depth) and the biological context[241]. Thus, the quality control evaluation was carried out individually for each of the studies, taking into account their particularities. We visualized the

distribution of each metric using violin plots per sample. We compared pre- and post-filtering distributions to explore if some cells should be discarded.

For SPMS-CNS dataset we removed nuclei that met any of the following criteria:

- Less than 1000 counts for library size.
- Less than 500 detected genes.
- Anomalous values (higher outliers) for mitochondrial gene ratio identified with *isOutlier* function from scuttle R package[242]. Specifically, this function calculates the Median Absolute Deviation (MAD – average of all differences between the arithmetic mean and each value in absolute terms). Nuclei were discarded if the value exceeded 3 MADs.
- Multiplets detected with scDblFinder R package[240].

For the RRMS-PBMCs and PPMS-PBMCs datasets, prefiltering plots revealed that the authors had already filtered the data[243], confirming no need for additional filtering.

The distributions of the quality control metrics were also visualized to infer whether to discard all cells from a particular sample. Their examination revealed no anomalous patterns, so we decided not to discard any sample from any of the three evaluated studies.

Gene-level quality control was also conducted. We determined the number of cells in which its expression was detected. If this value is extremely low, the gene is discarded as it is considered to not be present in the analyzed samples. For all datasets, genes were discarded genes if detected in less than 3 cells. Non-coding RNA genes were also removed following the regex patterns in parentheses: pseudogenes (P[0-9]+$), tRNAs (^TR.-), small nuclear RNAs (^RNU) and small nucleolar RNAs (^SNORD – for small nucleolar RNA, C/D box genes; ^SNORA – for small nucleolar RNA, H/ACA box genes; and ^SCARNA – for small Cajal body-specific RNA genes). HGNC guidelines were used to define these regex expressions[237].

3.3.5. NORMALIZATION

Normalization is performed to minimize technical variability while maintaining true biological differences, enabling the comparison between samples. In this study, we used a scaling factor normalization, in which each cell is assigned a specific factor that reflects technical-related biases. Gene counts were divided by this factor, ensuring proportional adjustment across genes within each cell.

To determine the approach for calculating the size factor, we accounted for the distribution of scRNA-seq data. On the one hand, the limited RNA content and *dropout* effect result in count matrices with a high proportion of zeros. On the other hand, the expression levels of each cell differ both qualitatively and quantitatively. Therefore, it cannot be assumed that the library size is the same across all sequenced cells.

To consider both aspects, the scaling factor was calculated by the deconvolution method (**Figure 3.6**)[244]. First, we clustered cells using the *quickCluster* function of the scran R package[245] to mitigate sparsity. Then, scaling factors were calculated with the *computeSumFactors* function from the same R package[245]. This procedure sums the counts by gene from all cells constituting each cluster generating *pseudocells*. Clusters are not generated once, as the process is performed multiple times generating different *pseudocells* formed by combinations of different cells. Next, the scaling factor for each *pseudocell* is calculated as the ratio of the *pseudocell* library size to the mean value. From this, the cell-specific scaling factors are obtained by a system of linear equations.

Following size factor calculation, we normalized the raw count matrices by dividing the counts from each cell by its respective size factor, and subsequently applied a log2 transformation using the *logNormCounts* function from the scuttle R package[245].

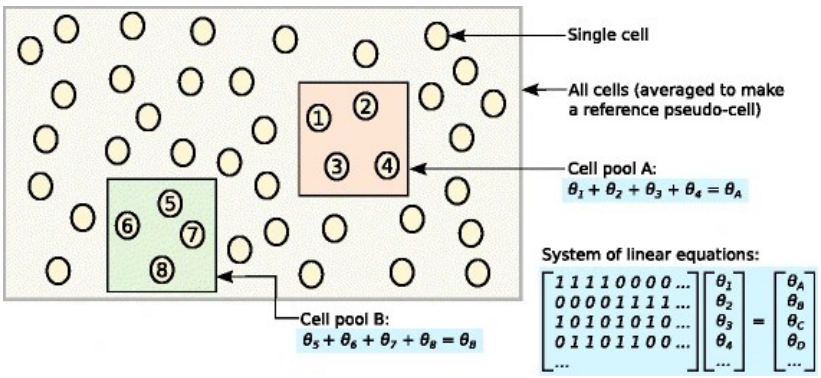


Figure 3.6. Graphical representation of the scaling factors calculation by the deconvolution method. Exemplification of the creation of two *pseudocells* (green and orange boxes), each of them formed by the sum of the expression levels of four different cells (yellow dots). The scaling factor for each *pseudocell* is calculated from its library with respect to the reference *pseudocell*; "deconvolving" the value for each real cell present in the dataset through a system of linear equations. Figure obtained from Lun *et al*. 2016[244].

3.3.6. HIGHLY VARIABLE GENE SELECTION

The following steps are performed prior to cell type annotation. To address the *curse of high dimensionality,* it is considered that some genes may exhibit minimal biological variation or the variability represents technical noise, which can obscure the meaningful biological signals. For that reason, highly variable genes (HVGs) are selected. HVGs constitute a subset of genes with biologically informative signals and non-random variability across cells, such as cell-type specific cell markers or genes driving differential functional responses.

To distinguish technical noise from biological variability, we used functions from the scran R package[245]. In brief, the gene-specific mean and variance values were calculated with the *modelGeneVar* function, obtaining the mean-dependent trend of the variance. Next, the algorithm of *fitTrendVar* function fitted the mean-variance relationship in the log-normalized data, assuming that the majority of genes exhibited baseline technical variation. The resulting trend represents the expected technical variance across genes as a function of their mean expression values. It allowed the decomposition of the total variance into technical and biological fractions. Subsequently, HVGs were selected as the 20% of genes exhibiting the largest biological variance using *getTopHVGs* function. Batch information was considered when possible.

3.3.7. DIMENSIONALITY REDUCTION

Following HVGs selection, transcriptional profiles were further condensed through dimensionality reduction strategies. Without dimensionality reduction, we would consider each gene as an independent dimension. These strategies project the high-dimensional data into a lower-dimensional space, enabling data visualization. In this work, we conducted *Principal Component Analysis* (PCA), *t-distributed Stochastic Neighbor Embedding* (tSNE) and *Uniform Manifold Approximation and Projection* (UMAP); which are described below. As they presented stochastic components, the seed 1234 was established to ensure reproducibility.

PCA was performed on the HVGs normalized expression matrix to capture the major components of transcriptional variation. PCA reduces the number of features to be evaluated, each of which is constituted by a linear combination of the HVGs expression levels. Concisely, the process involves the following steps i) centering the data to the mean expression of each gene, ii) computing the covariance matrix to identify correlated expression patterns, and iii) obtaining the principal components (PCs) through value decomposition. These PCs represent the reduced set of features that glomerate the

variability of the analyzed data. Each successive PC explains the maximum remaining variance: PC1 captures the highest proportion of variability, followed by PC2, and so forth. In this work, the first 50 PCs were obtained with the function *runPCA* from scater R package[242].

To determine the number of PCs of interest, the elbow method was applied using the *findElbowPoint* function from the PCAtools R package[246]. This approach identified the *elbow* in the scree plot (a graphical representation of the percentage of variance explained by each PC). The elbow distinguishes relevant PCs with steep slopes relative to their neighboring PCs, from the non-relevant PCs, which exhibit shallow slopes.

To project the data in a two-dimensional space, we employed the t-SNE and UMAP algorithms. The major objective of these methods is to preserve the intrinsic neighborhood relationships between cells, so that their distribution in two dimensions is as similar as possible to their distribution in the multidimensional space.

First, the tSNE algorithm computes pairwise cell similarities using as input the PCs selected with the elbow point. Specifically, similarity is calculated with a t-distribution-based probability metric, which defines the probability that two cells would be nearest neighbors. Cells are then randomly located in the two-dimensional space and iteratively relocated until the same distribution of close neighbors from the PCs space is found. Therefore, it can be inferred that cells close in the two-dimensional space are also neighbors in the multidimensional space. However, this assumption cannot be extended to cells that are far apart from each other. This process was executed with the *runTSNE* function of the scater R package[242], using as input the PCs previously selected by the *elbow* method. Indirectly, the *perplexity* parameter determines the number of neighbors to be evaluated. After testing values from 5 to 300, the parameter was set to 40.

UMAP is an alternative approach to tSNE for visualizing the distribution of cells in two dimensions. Its methodological basis focuses on calculating the distance that separates cells in multidimensional space[247]. In this case, the distance can be calculated by the method of the researcher's choice. In our case, the default Euclidean distance was used. Cells whose distance is less than a previously established threshold are connected, generating topological structures called *simplex*. Finally, *simplex* are connected based on the relative distance that separates them, so that the cells can be arranged in the dimensions of interest. As *simplex* are connected, we can compare the distribution of cells that are not close to each other. UMAP was performed with the *runUMAP* function of the scater R package[242]. As described in the tSNE methodology, the algorithm uses the previously selected PCs to optimize the process. The distances are calculated

between the nearest neighbors, whose number is selected by the researcher by the parameter *k*. In this work, it was set to 20.

3.3.8. EXPLORATORY ANALYSIS OF SOURCES OF VARIABILITY

In addition to the variability expected from the different cell types, other technical and biological sources of variation can influence gene expression patterns. Therefore, an exploratory analysis of the potential explanatory variables was performed prior to clustering and inferential biological approaches. As we analyzed previously published datasets, the covariates to be studied were depended on the metadata available:

- SPMS-CNS[248] included: sample identifier, age, capture batch, and sequencing batch.
- RRMS-PBMCs[243] included: sample identifier, age, previous treatments, MS severity and batch effect.
- PPMS-PBMCs[243] included: sample identifier, age and batch effect.

In the three datasets we also evaluated the variable of interest, combining the condition (MS or control) and sex (female or male) categories. We also considered that gene expression in individual cells is not constant over time, as it may fluctuate depending on the cell cycle stage. To account for this biological source of variability, we inferred the cell cycle phase using the *cyclone* function from the scran R package[245]. This function incorporates gene pairs for the G1, S, G2-M phases. For each gene pair, the first gene is known to have higher expression in the designated phase compared to the others. The algorithm computes the ratio of gene pairs that satisfy the assigned patterns. These proportions are then transformed into scores that predict the most likely cell cycle stage.

Considering all of these variables, the function *getExplanatoryPCs* from scatter R package[242] was used to calculate the percentage of variance explained by each variable in each PC, and the results were visualized with the function *plotExplanatoryPCs* from the same package[242].

3.3.9. CLUSTERING

The main objective of this step is to categorize cells based on the similarity of their gene expression patterns, obtaining groups that represent distinct cellular identities. Regardless of the specific strategy implemented, the aim of clustering is to ensure that the transcriptional variability within each group is smaller than the differences between groups. In this study, we employed a graph-based methodology due to its scalability across diverse study sizes. The process consists of two phases:

- **Graph construction:** cells (represented by nodes) are interconnected via edges weighted by their transcriptional similarity.
- **Community detection**: highly interconnected cell communities are identified, corresponding to cell populations with similar expression profiles.

The graph was constructed with the *Shared Nearest Neighbor* (SNN) algorithm via the *buildSNNGraph* function from the scran R package[245]. The inputs were the PCs selected with the elbow method. From them:

1. Nodes represented individual cells positioned according to their PC coordinates.
2. Edges connected cells that share mutual nearest neighbors.
3. The weight of the edge quantified the transcriptional similarity, directly proportional to the number of shared neighbors: stronger weights indicate higher similarity in their transcriptomic profiles.

A critical parameter is the number of neighbors evaluated (k). It defines the resolution of the clusters and their composition by modulating the graph's connectivity. Smaller k values confer higher resolution by detecting higher number of populations (finer subpopulations), while larger k values produce broader classifications with reduced number of clusters. Through the evaluation of multiple k sizes, we determined $k = 20$ to be optimal for our analysis.

Following graph-construction, cellular communities were identified using the *walktrap* algorithm implemented in *cluster_walktrap* function from igraph R package[249]. This approach defines clusters as densely interconnected regions. Community connections exhibit both higher frequency and greater edge weights compared to what is expected by chance and to inter-community links. As a result, random walks on the graph tend to become “trapped” within these regions due to their high connectivity.

Once clusters were defined, the next step was to identify which genes are overexpressed in each cluster with respect to the rest, obtaining marker-specific genes per cluster.

Marker genes were identified with the *findMarkers* function from the scran R package[245]. Specifically, differential expression analysis of HVGs was conducted between cluster pairs using the Wilcoxon test. To designate a gene as a marker of a cell identity (cluster), the gene must be significantly overexpressed in all pairwise comparisons between the corresponding cluster and the rest. Adjusted p-values were calculated by the Benjamini-Hochberg (BH) method[250], considering significance when false discovery rate (FDR) < 0.05.

We also explored how distant the clusters were with the *purity* metric. The purity of each clustering was calculated with the *neighbourPurity* function of the bluster R package[252]. We determined to which group the neighbors of each cell belonged. Then, the purity per cell was calculated as the ratio of neighbors from the same group to the total number of neighbors[233].

3.3.10. CELL TYPE ANNOTATION

We next performed the cell type assignment using a reference-based approach, where we compared the cell profiles of each dataset to publicly available profiles of annotated cells. This approach compares normalized expression patterns of HVGs against reference datasets with known populations. As we analyzed samples from CNS and PBMCs, we used different pipelines for each tissue type.

For cell type annotation of the SPMS-CNS dataset, the *brainCells* function from the BRETIGEA R package was employed[251]. The algorithm is specifically designed for the classification of the six CNS major cell types: neurons, astrocytes, endothelial cells, microglia, oligodendrocytes, and OPCs. Precisely, BRETIGEA provides a curated list of 1,000 marker genes per cell type, derived from consensus expression profiles from different CNS datasets. In this study, after testing different parameters, we evaluated the top 60 marker genes. The algorithm calculates cell-type-specific scores for each cell based on the expression of the marker genes, assigning the cell type annotation with the highest score.

Meanwhile, RRMS-PBMCs and PPMS-PBMCs datasets were analyzed with the *SingleR* function from the SingleR R package[252]. The *MonacoImmuneData* dataset from the celldex package was employed as reference[252]. This dataset constitutes a standard reference for human immune cell type annotation in PBMC samples, containing expression profiles of B cells, T cells (both CD4+ and CD8+), NK cells, dendritic cells, and monocytes. It also includes neutrophils, basophils, and progenitor cells, allowing the identification of any contaminating populations that may not have been completely

removed during PBMC isolation. *SingleR* calculated the Spearman correlation between the expression levels of shared genes in the reference cells and in the cells under evaluation, assigning a specific score for each potential cell type. Each cell is then annotated according to the highest scoring cell type. In this study, annotation was performed at the single-cell level, without prior aggregation by clusters, to capture the full heterogeneity of the PBMC populations.

Following cell type annotation using reference datasets, we examined marker genes previously described in the literature to confirm the assigned annotations.

3.3.11. BIOLOGICAL INFERENCE APPROACHES

Once the cells were annotated, we next characterized sex-associated molecular differences in MS pathogenesis within each dataset, which represented a particular MS subtype and tissue. The specificities of each analysis are described in the following sections. In brief, for each cell type, we identified those genes with different expression patterns (differential gene expression analysis), in which functions are these genes involved (functional profiling analysis), and the differential activation of signaling pathway effectors (signaling pathways analysis). Each of these biological inference approaches were evaluated in three scenarios: impact of disease in females, impact of disease in males, and sex-differential impact of disease. We also performed cell-cell communication analyses to quantitatively characterize interaction networks for each group (control females, MS females, control males, MS males).

3.3.11.1. Comparisons

We evaluated three scenarios regardless of the approach or the selected cell type:

- **Impact of disease in females (IDF):** differences among MS patients and healthy individuals being female by the comparison:

  MS females - control females

- **Impact of disease in males (IDM)**: differences among MS patients and healthy individuals being male by the comparison:

  MS males - control males

- **Sex differential impact of disease (SDID)**: sex differences among MS patients without considering the inherent sex variability in healthy individuals, that is, finding differences between IDF and IDM by the comparison:

(MS females - control females) - (MS males - control males)

IDF and IDM evaluate the impact of suffering MS separately for each sex. Meanwhile, SDID comparison identifies the features that change due to sex in MS. Irrespective of any of the former comparisons or the statistical test applied, as a result of the statistical analysis we obtain: i) the test statistic, ii) the p-value associated with the statistic, iii) the adjusted p-value for correcting multiple testing and iv) the logFC representing the direction (logFC sign) and the magnitude (logFC absolute value) of change. For the ease of the logFC interpretation, now we describe a qualitative representation for each of the comparisons (**Figure 3.7**).

A positive logFC (logFC > 0) in the IDF comparison indicates a higher level of the analyzed feature in MS females compared to control females (i.e., a lower level of the analyzed feature in control females compared to MS females). Conversely, a negative logFC (logFC < 0) indicates a higher level of the analyzed feature in control females compared to MS females (i.e., a lower level of the analyzed feature in MS females compared to control females) (**Figure 3.7-A**). The IDM comparison is interpreted in the same way; however, it applies to males instead of females (**Figure 3.7-B**).

The results from SDID comparison indicate if the features are increased in females (logFC > 0) or in males (logFC < 0) in MS. We use this terminology to illustrate results for simplicity's sake. Nonetheless, we can delve deeper into the exact patterns of change by looking at IDF and IDM comparisons. Patterns that could lead to an increased feature in females are: positive logFCs in both IDF and IDM but larger in IDF, positive logFCs in IDF and negative logFCs in IDM, negative logFCs in both IDF and IDM but larger in IDM, positive logFCs in IDF without change in IDM, and negative logFCs in IDM without change in IDF (**Figure 3.7-C, top figures**). Reversely, patterns that could lead to an increased feature in males are: positive logFCs in both IDF and IDM but larger in IDM, negative logFCs in IDF and positive logFCs in IDM, negative logFCs in both IDF and IDM but larger in IDF, positive logFCs in IDM without change in IDF, and negative logFCs in IDF without change in IDM (**Figure 3.7-C, bottom figures**).

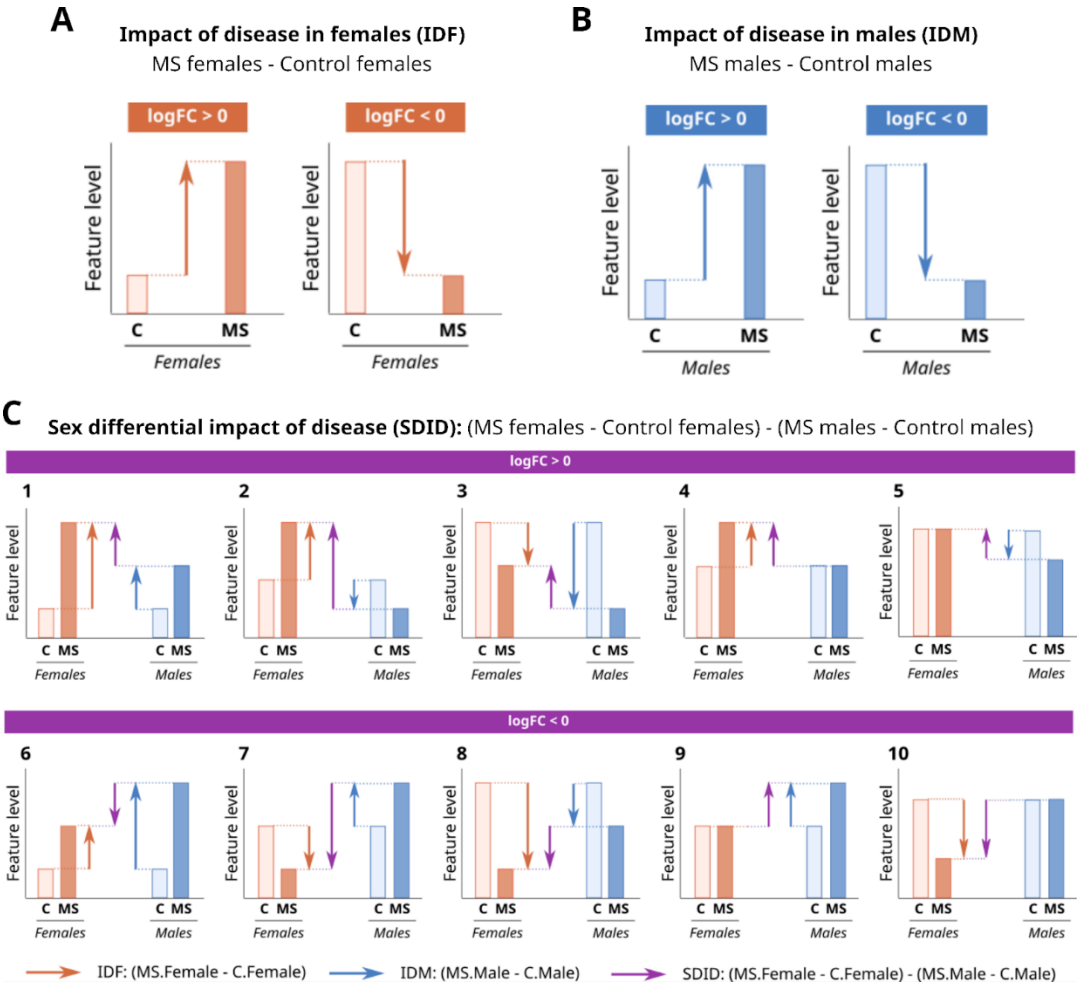


Figure 3.7. Schematic representation of the possible patterns that may arise when evaluating (A) the impact of disease in females, (B) the impact of disease in males and (C) the sex differential impact of disease. For each scenario, the X-axis represents the groups considered in the comparison, and the Y-axis the level of the tested feature (e.g. genes or signaling pathways). Arrows indicate the direction of change, as defined by the sign of the logFC: a positive logFC denotes higher levels in the first group of the comparison (or lower in the second), while a negative logFC indicates the opposite. Levels in control females and control males are shown as equal for simplicity, although this does not necessarily occur in real data. *C: controls; IDF: impact of disease in females; IDM: impact of disease in males; MS: multiple sclerosis; SDID: sex differential impact of disease.*

3.3.11.2. Differential gene expression analysis

The main objective of this analysis is to identify, for each cell type, the number of genes that exhibit a significantly sex-dependent expression pattern in MS testing the three contrasts defined in the *Comparisons* section. This analysis was conducted separately for each dataset and annotated cell type.

We employed the *Model-based Analysis of Single-cell Transcriptomics* (MAST) R package framework[253] using as input the normalized matrix by transcripts per million. Transcript length annotation was downloaded from Biomart Ensembl Genes 106 (GRCh38.p13 version) on 5th December 2022.

MAST approach is specifically designed to account for the characteristics of single cell data (e.g., high sparsity and Poisson-like distribution). For each cell type and comparison, we fitted with the *zlm* function a two-part generalized linear model, called hurdle model, that was composed of:

- Logistic regression: it determines whether a gene is expressed or not (discrete distribution).
- Gaussian linear model: it fits the distribution of counts conditioned on the expression of the genes (continuous distribution).

The model was adjusted using an empirical Bayes method. The model also incorporated the cellular detection rate (fraction of expressed genes per cell) as a covariate to mitigate technical confounding, while adjusting for the biological and technical variables reported by the original authors through fixed effects.

After fitting the model, statistical tests for the three comparisons were performed using the *lrTest* function, and the corresponding logFC values were calculated with *logFC* function from the same package[253]. The p-values were adjusted using the BH method[250]. Genes were considered significant when FDR < 0.05 and absolute logFC > 0.5.

3.3.11.3. Over-representation analysis

ORA is one of the foundational methods in functional enrichment analysis. The process begins with a list of genes of interest, such as those significantly differentially up- or down-regulated in each cell type and comparison. ORA compares the frequency of genes associated with specific biological functions in our list to their frequency in a reference set (that is, the function of interest). The statistical test is then applied to determine

whether the proportion of genes from our list linked to a particular function is significantly greater than what would be expected by chance.

To address the redundancy derived from gene propagation in GO terms, we used the weight01 approach. This method combines the algorithm *elim*, which removes annotated genes from parent terms when the child term is significant; and the algorithm *weight,* that assigns gene scores according to GO term relationships, giving higher weight to genes that specifically contribute to enriched terms rather than broadly annotated ones[254]. The *biological processes* GO terms were obtained from the org.Hs.eg.db R package[255]. ORA was performed with the *runTest* function of the R package topGO, selecting *weight01* algorithm and *Fisher exact test* for the statistical evaluation[256]. Adjusted p-values were calculated by the BH method[250], considering significance when FDR < 0.05.

For result interpretation and visualization, we performed semantic clustering of significant terms using the simplifyEnrichment and simplifyGO R packages[257]. First, we computed the semantic similarity matrices between enriched GO terms. Next, the Louvain community detection algorithm was applied to identify clusters of biologically related terms. As a result, we obtained the representative word clouds for each cluster, with term frequency determining the word size.

3.3.11.4. Protein-protein interaction analysis

PPI analyses were performed using the STRINGdb R package and the STRING web resource[195] to identify relevant molecular networks from the differentially expressed genes in each cell type and comparison. In the resulting networks, proteins were represented as nodes, while the interactions between them constituted the edges. The networks were constructed based on the interactions available in the STRING database, which includes physical and functional associations, and inference relationships based on data mining. Edges were weighted based on confidence of the interaction: the more significant the thickness, the greater the confidence.

Following network construction, an enrichment p-value was calculated to assess whether the network contains more connections than would be expected by chance, which would suggest meaningful relationships among the proteins.

Computationally, the Gene Symbol identifiers were mapped to obtain STRING identifiers with *string_db* function. Interaction networks were obtained with *string_db$plot_network* function by keeping default parameters and examining the

database's total number of physical and functional interactions. Significant networks were considered when PPI p-value < 0.05.

Finally, ORA was conducted on the STRING-derived networks to identify biological processes significantly enriched within each network.

3.3.11.5. Signaling pathway activation analysis

To infer the signaling pathway activation we applied the pipeline from the HiPathia R package[258], based on the annotation from KEGG database[198]. Given that the repertoire of active pathways can vary depending on tissue type, some KEGG pathways were analyzed just in the CNS dataset, others in the PBMCs datasets and the remaining pathways in both tissues (**Supplementary Table 3.S2**).

HiPathia includes a collection of 146 human KEGG signaling pathways, which are decomposed into functional subpathways (**Figure 3.8-A**). Each subpathway corresponds with a fraction of the pathway from one or more initial receptors to a specific effector protein, independently of the upstream stimuli that may trigger it. In total, the 146 pathways were divided into 1,876 subpathways, each representing a signaling unit analyzed in this work.

This analysis allows to distinguish between differential gene expression patterns and the activation of genes within the context of signaling pathways. As illustrated in the schematic representation (**Figure 3.8-B**), we defined a simplified example involving the terminal portion of a signaling subpathway, where feature A activates feature B, the effector that conducts function X. It is possible for feature B to be more expressed in condition 1 than in condition 2, while its activation is greater in condition 2. This situation may occur if, despite its abundance, few feature A molecules are activated in condition 1, transmitting a weaker signal to downstream components such as feature B.

To compute the transmission of the signal from one node to another are considered: 1) the proportion of molecules at the upstream node capable of transmitting the signal (that is, the proportion of activated molecules) and 2) the abundance of the downstream node.

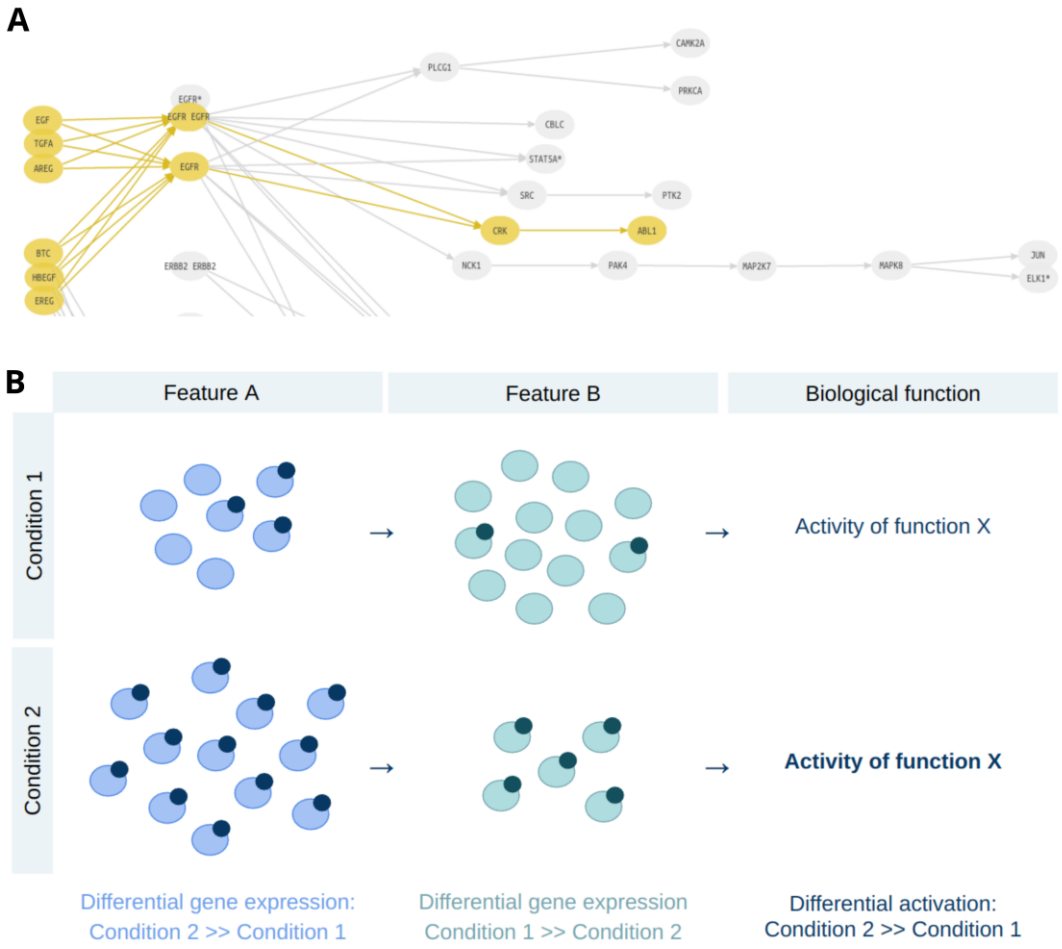


Figure 3.8. Illustration of signaling pathway analysis. (A) Example of a signaling pathway fraction, where nodes represent proteins and edges denote interactions along the signaling cascade. The subpathway is highlighted in yellow. Figure from HiPathia manual (last accessed 22 July, 2025) (B) Conceptual exemplification between differential expression analysis and differential signaling pathway activation. Each circle represents a protein within a two-node signaling pathway, where feature A activates feature B. Feature B acts as the effector that develops function X in two different conditions. The small dark circle within a node indicates that the protein is active, while its absence indicates inactivity.

To estimate the activation level of signaling subpathways, HiPathia algorithm implements a mathematical model that simulates the propagation of the signal. First, gene identifiers are standardized by converting Gene Symbols to ENTREZ IDs using the *translate_data* function. Next, the *normalize_data* function transforms normalized

gene expression into values ranging from 0 to 1, representing the relative transcriptomic abundance of each gene. These scaled values serve as the input for simulating signal transduction. The signal propagation through the subpathway is computed assuming that the stimulus presents a signal of 1. Then, for each node of the pathway, the activation of the downstream node is calculated as:

$$S_n = v_n \cdot \left(1 - \prod_{S_i \in A_n} (1 - S_i)\right) \cdot \prod_{S_j \in I_n} (1 - S_j)$$

Where S_n denotes the signal value at gene n, v_n the normalized expression value of gene n, A_n the activation signals, and I_n the inhibition signals.

Once the signaling activity was computed, we performed differential signaling pathway activation analysis for each cell type testing the three previously defined comparisons (IDF, IDM and SDID). Since the resulting activation matrix shares the same structure as the single cell gene expression matrix (high sparsity and high dimensionality), we applied MAST algorithm[253] previously defined in the *Differential gene expression analysis* section. Adjusted p-values were calculated by BH method[250] and significant results were considered when FDR < 0.05.

3.3.11.6. Cell-cell communication analysis

To complete the characterization of sex differences in MS, we conducted another inference analysis to identify ligand-receptor interactions with the CellChat R package[259,260]. CellChat incorporates their own curated database, called CellchatDB. In detail, the CellChatDB human database comprises 1,939 interactions involving 546 ligands and 507 receptors and including paracrine/autocrine signaling interactions, extracellular matrix-receptor interactions and cell-cell contact interactions.

We computed the communication probabilities for each interaction between pairs of cell types in each group of interest (control females, MS females, control males, and MS males). The first step involved the identification of overexpressed ligands and receptors within each cell type. Adjusted p-values were calculated with the BH method[250], and significance was established at FDR < 0.05. Significant interactions were determined if either the ligand or the receptor were overexpressed. Next, the communication probability between cell types was computed for each significant interaction with the *computeCommunProb* function. These probabilities are modeled according to the law of

mass action using a Hill-type equation. Specifically, the communication probability (Pi,j) from cell type i to cell type j, for a given ligand-receptor pair k, is modeled by:

$$P_{i,j}^{k} = \frac{L_i R_j}{K_h + L_i R_j} \times \left(1 + \frac{AG_i}{K_h + AG_i}\right) \cdot \left(1 + \frac{AG_j}{K_h + AG_j}\right) \times \frac{K_h}{K_h + AN_i} \cdot \frac{K_h}{K_h + AN_j} \times \frac{n_i n_j}{n^2},$$

$$L_i = \sqrt[m1]{L_{i,1} \cdots L_{i,m1}},\ R_j = \sqrt[m2]{R_{j,1} \cdots R_{j,m2}} \cdot \frac{1 + RA_j}{1 + RI_j}.$$

Where i denotes the cell type expressing the ligand (sender), j the cell type expressing the receptor (receiver), L the ligand expression level, R the receptor expression level, and K_h the cooperativity constant, which is typically set to 0.5 by default since the expression values are normalized between 0 and 1. For protein complexes, the expression values were corrected by considering all required subunits. Specifically, the expression values of ligand subunits (L_i) are multiplied together, and then the root is taken based on the number of subunits. In the case of receptors (R_j), the resulting value is modulated depending on whether the subunits have activating (RA_j) or inhibitory (RI_j) activity.

Interactions present in fewer than 10 cells in each cell type were filtered out. Once the probabilities for individual ligand-receptor interactions were computed, communication probability for a signaling pathway was calculated by summing the probabilities of all its associated significant ligand-receptor interactions. The probability of non-significant interactions was set as zero. If the resulting sum was greater than zero, the pathway was considered significantly involved in the communication between the corresponding sender and receiver cell types. For both specific ligand-receptor and pathway levels, the probability of interaction is used as a proxy of the interaction strength for the corresponding interaction.

The differential number of interactions were computed with *netVisual_diffInteraction* function for the comparisons: 1) MS females *vs* control females and 2) MS males *vs* control males. Additionally, chord plots representing significant interactions from specific signaling pathways and cell types were generated with the *netVisual_chord_gene* function.

3.3.12. WEB TOOL

The web tool (https://bioinfo.cipf.es/cbl-atlas-ms/) was developed under the structure of the R shiny package[261], and is hosted in the computational cluster of the CIPF. This resource was created to provide free access to the complete results in an easy-to-use manner. The web is divided into seven sections: 1) *Home* section, to provide a summary of the results available on the website; 2) *Gene expression* section, to explore changes in the expression of the genes of interest; 3) *Functional profiling* section, to delve into the functional profiling results derived from ORA analyses; 4) *Signaling pathways* section, to identify the differential activation of protein effectors in signaling pathways of interest; 5) *Cell-cell communication networks* section, to inspect the ligand-receptor interactions strengths; 6) *Study overview* section, to review the outline of dataset selection and the bioinformatic methods; and 7) *Help* section, to assist in the interpretation of the comparisons and the displayed results.

3.4. RESULTS

This section includes the results obtained throughout the different phases of our study, which are individually defined in the *Materials and Methods* section of this chapter. It begins with the description of the transcriptomic datasets identified through literature screening. Next, the results of the processing steps are exposed, leading to the annotation of the major cell types in each tissue. We finally explored the inferred biological results for each disease subtype (SPMS, RRMS and PPMS) in the correspondind tissue (CNS or PBMCs).

3.4.1. IDENTIFICATION OF SUITABLE DATASETS THROUGH LITERATURE SCREENING

The search was performed until April 2022, where we identified 61 distinct studies (**Figure 3.9**). After evaluating them, 59 studies were discarded due to the following exclusion criteria: samples not derived from human tissues, data type, disease examined, absence of sex information, insufficient sample count, or unavailability of data. As a result, two studies were included in the analysis: UCSC-MS[248] (from the *Multiple sclerosis* identifier in the UCSC Cell Browser database) and GSE144744[243] (from the GEO database).

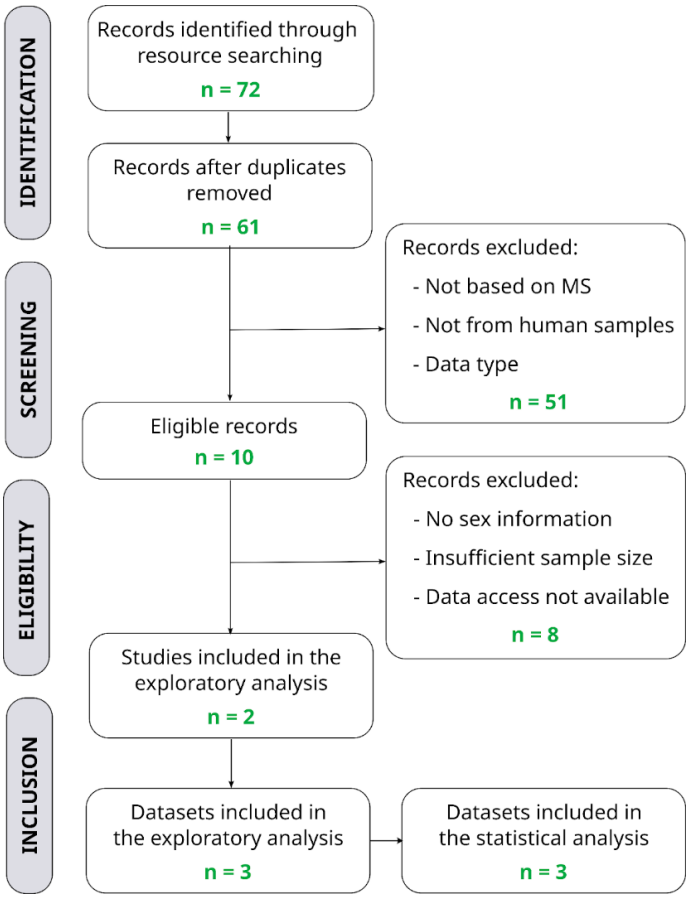


Figure 3.9. Systematic review results following PRISMA guidelines. Specification of remaining study number (n) through the identification, screening, eligibility, and inclusion phases, along with the corresponding exclusion justifications. *PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses*.

The UCSC-MS study[248] stores an snRNA-seq dataset from post-mortem brain tissues to explore SPMS subtype of disease. It contains nuclei from cortical grey matter regions with the adjacent subcortical white matter. SPMS samples were taken from lesion areas, while the control tissue was obtained from unaffected individuals. In our study, this cohort is identified as SPMS-CNS.

Meanwhile, the GSE144744 study contains three independent cohorts of scRNA-seq data from PBMCs. Individually, cohort 1 (labelled in this thesis as RRMS-PBMCs) represents the RRMS subtype. It stores control samples, and paired-sample data acquired

in two time points for MS patients: before and after Natalizumab treatment – a monoclonal antibody that inhibits the PBMCs to migrate towards the CNS[262]. This cohort was partially incorporated into our analysis, as we excluded the paired drug-treated samples from MS patients. Meanwhile, cohort 2 only included female samples, being consequently discarded. Lastly, cohort 3 (labelled in this thesis as PPMS-PBMCs) was composed of control and PPMS untreated samples. It was completely incorporated as it fulfilled the established criteria. The sample distribution for each cohort can be found at **Supplementary Table 3.S3**.

Overall, three different cohorts were individually analyzed in this work. The landscape of sex differences was therefore generated for CNS in the SPMS subtype, and for PBMCs in RRMS and PPMS subtypes of the disease. The technical and biological characteristics of each dataset can be consulted in **Table 3.1**.

Table 3.1. Datasets description. *Control females : MS females : control males : MS males. *GEO: Gene Expression Omnibus; MS: multiple sclerosis; PBMCs: peripheral blood mononuclear cells; PMID: PubMed identifier; PPMS: primary progressive MS; RRMS: relapsing-remitting MS; scRNA-seq: single cell RNA-seq; snRNA-seq: single nucleus RNA-seq; SPMS: secondary progressive MS; UCSC: University of California Santa Cruz.*

	SPMS-CNS	**RRMS-PBMCs**	**PPMS-PBMCs**
Database	UCSC Cell Browser	GEO	GEO
Identifier	UCSC-MS	GSE144744 cohort 1	GSE144744 cohort 3
Sequencing type	snRNA-seq	scRNA-seq	scRNA-seq
Sequencing method	10x Genomics	10x Genomics	10x Genomics
Sequencing platform	Illumina HiSeq 2500	Illumina Nextseq 500	Illumina Nextseq 500
Sample type	*Post-mortem* brain tissue	PBMCs	PBMCs
MS subtype	SPMS	RRMS	PPMS
N samples by group*	4:8:5:4	5:4:5:5	3:3:6:6
Status of the accessible data	unfiltered raw counts	filtered raw counts	filtered raw counts
N cells	48,919	71,592	265,342
N genes	65,217	15,354	15,354
PMID	31316211	33748804	33748804

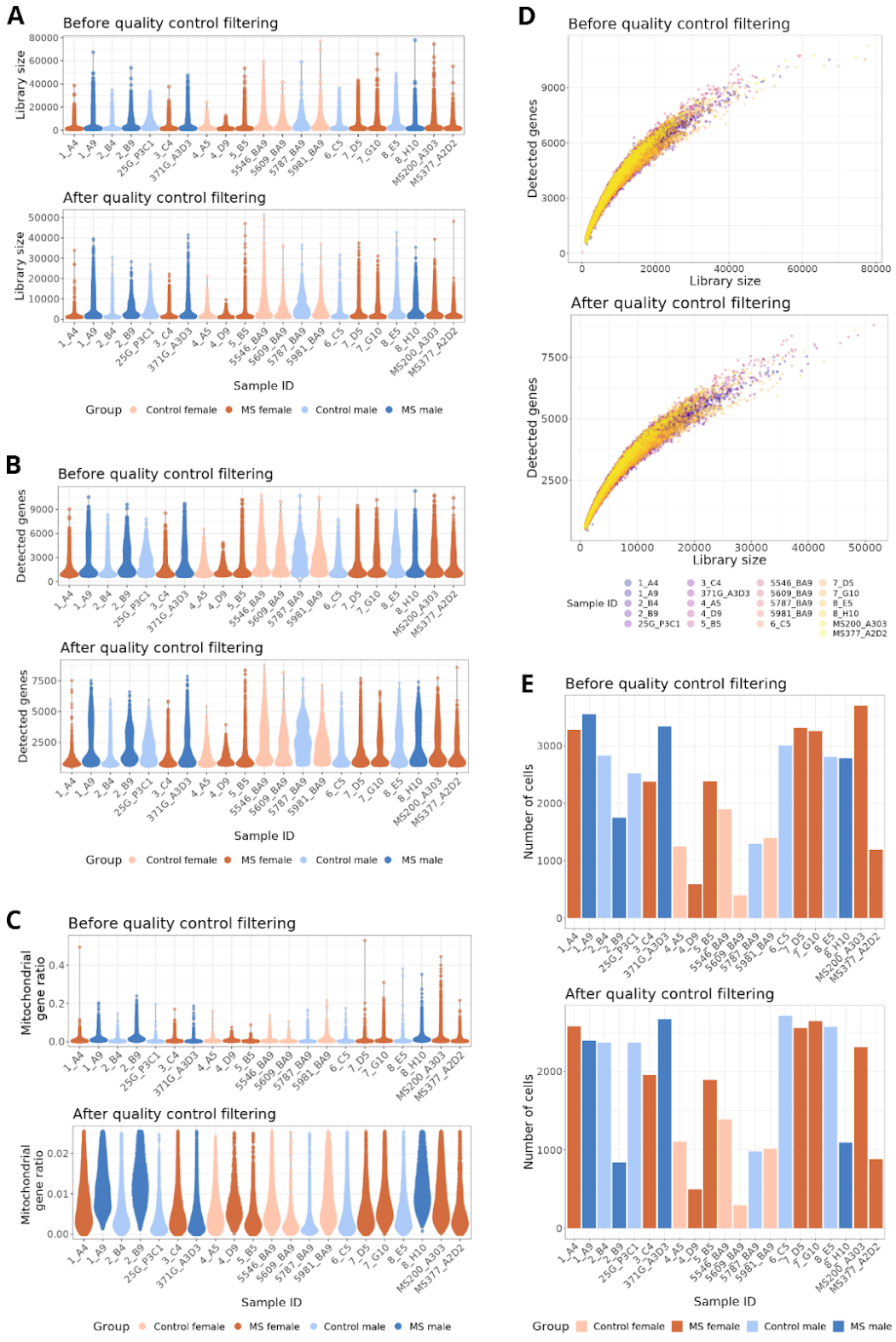
3.4.2. COMPUTATIONAL DATA PROCESSING: FROM QUALITY CONTROL TO CELL TYPE ANNOTATION

To obtain the processed count matrices with the cell types assigned in each cohort, we followed the steps described in the *Materials and Methods* section. Throughout the pipeline, we generated intermediate results that guided dataset-specific decisions at each step of the processing workflow.

Precisely, we first assessed the initial status of each set of data representing the quality control metrics before applying any filtering strategy. In the case of SPMS-CNS, we retrieved the raw and unfiltered gene expression matrix. To define proper cutoffs, we evaluated the distribution of the quality control metrics for each cell, including library size, the number of detected genes, and the proportion of mitochondrial transcripts (**Figure 3.10-A-C, before QC**). We also evaluated the cell distribution by the library size and number of detected genes, where we observed the expected trend: the larger the library size, the greater the number of genes detected (**Figure 3.10-D, before QC**). Based on these metrics, we defined the filtering thresholds (discarded nuclei if: < 1000 counts, < 500 detected genes, and higher outliers for mitochondrial gene ratio). The same plots were represented after filtering to confirm the cutoffs had been applied appropriately (**Figure 3.10-A-D after QC**). We also confirmed that at least 200 cells per sample remained for downstream analysis (**Figure 3.10-E, after QC**).

On the contrary, the data included the study GSE144744 (RRMS-PBMCs and PPMS-PBMCs) consisted of the raw and filtered gene expression matrices (**Supplementary Figures 3.S1A-D and 3.S2A-D**). The graphical representations confirmed that the downloaded data was already filtered by the original authors. Following their description[243], we did not detect cells with fewer than 500 or more than 15,000 library size, or cells with fewer than 300 or more than 5,000 detected genes. Cells with more than 20% of mitochondrial transcripts were also removed. We also confirmed that at least 200 cells per sample remained for downstream analysis (**Supplementary Figures 3.S1-E and 3.S2-E**).

Figure 3.10. Quality control assessment of the SPMS-CNS dataset before and after filtering. (*Next page*) (A-C) Representation of cells (dots) based on the sample of origin (X-axis) and according to the quality control metric value (Y-axis) before and after filtering: (A) library size, (B) number of detected genes and (C) mitochondrial gene expression ratio. (D) Scatter plot of the library size *versus* the number of detected genes. Each dot represents a cell colored by the sample of origin. (E) Bar plot for the number of evaluated cells per sample. *ID: identifier; MS: multiple sclerosis.*


A
Before quality control filtering
Library size
After quality control filtering
Sample ID
Group Control female MS female Control male MS male
B
Before quality control filtering
Detected genes
After quality control filtering
Sample ID
Group Control female MS female Control male MS male
C
Before quality control filtering
Mitochondrial gene ratio
After quality control filtering
Sample ID
Group Control female MS female Control male MS male
D
Before quality control filtering
Detected genes
Library size
After quality control filtering
Sample ID
E
Before quality control filtering
Number of cells
After quality control filtering
Sample ID
Group Control female MS female Control male MS male

For all three datasets, cell-specific scaling factors were computed using the deconvolution strategy. The raw filtered count matrices were normalized by the corresponding scaling factor and then log-transformed (**Supplementary Figure 3.S3)**. We did not identify any samples with anomalous distributions, so none of them were discarded.

From the normalized data, we determined the HVGs by selecting the genes with the highest biological variance. In all three cohorts, technical variance models were calculated for each batch independently, as all studies incorporated the capture batch and/or the sequencing batch as a metadata variable. To assess the fit of the model, we plotted total variance against mean expression for each processing batch and overlaid the fitted trend representing the technical variance component. For RRMS-PBMCs and PPMS-PBMCs cohorts, any anomalous pattern was detected. An example for each is illustrated in **Supplementary Figure 3.S4-A-B**. On the contrary, we identified overfitted trends in the SPMS-CNS dataset (**Supplementary Figure 3.S4-C**). This phenomenon was identified due to the trend being erratically disposed, generating a *path* between the dots (genes) of the plot. To obtain a better fit, the process of modelling the technical component of the variance was repeated by deactivating the density weights considered by the *fitTrendVar* function. In the standard process, a weight is given to each gene inversely proportional to the number of genes that have a similar average value. That is, the higher the number of genes with similar mean, the lower the weight, as we used more values to calculate the statistics. The resulting trend without providing gene density weights can be consulted in **Supplementary Figure 3.S4-D,** where we observed a better fit. Overall, as a result of HVGs selection, 2,093 genes were obtained for the SPMS-CNS dataset, 1,825 for RRMS-PBMCs and 1,768 for PPMS-PBMCs.

HVGs were used as input to perform the PCA. After applying the cut-off point by the elbow criteria, 7 principal components were selected in SPMS-CNS, 4 in RRMS-PBMCs, and 3 in PPMS-PBMCs (scree plots can be consulted in **Supplementary Figure 3.S5-A-C**). Cells were then projected onto the first two principal components and colored by their associated metadata. As a representative example, **Supplementary Figure 3.S5-D-F** shows the cell distributions by experimental group. Upon a visual inspection, we did not observe any distinct distribution patterns attributable to specific metadata categories (e.g., control males did not localize separately from other groups). To further assess the potential sources of variability, we quantified the proportion of variance explained by each reported variable across the principal components (**Supplementary Figure 3.S6**). We did not identify any variable as a dominant variance contributor to the first two principal components. Thus, we did not apply any data correction strategies prior to clustering, as we expect this source of variation to be

assigned to the different cell types. It should be noted that these variables do explain variability even if it is not predominant, so they were incorporated in the differential gene expression models.

Cell identities were next determined. We identified 18 clusters in the SPMS-CNS cohort, 20 in the RRMS-PBMCs cohort and 34 in the PPMS-PBMCs cohort. For each cluster, we identified marker genes and, after evaluating them, we did not detect significant genes that we could associate with technical artifacts (e.g. high proportion of ribosomal genes). Therefore, no clusters were excluded from further analysis. The cluster distribution on the UMAP coordinates can be visualized in **Supplementary Figure 3.S7-A-C**. We observed that the clusters for the CNS dataset are much further apart than those of PBMCs; this is to be expected since CNS profiles are more different from each other than those of PBMCs (generally speaking, an astrocyte and a neuron are less similar than a CD4+ cell and a CD8+ T cell). In addition, in CNS samples, most cells are observed to have a high purity (**Supplementary Figure 3.S7-D**). Thus, most groups are defined in a specific area and arranged contiguously (neighbors of the same cell belong to the same cluster as that cell). The clustering of PBMCs, on the other hand, is much more diffuse. This situation is also reflected in the purity diagram by observing how, in most cases, cells belonging to one group have as neighbors those of another (**Supplementary Figure 3.S7-E-F**).

Based on these results, we performed the annotation of cell types and, for each of them, we conducted biological inference approaches that are described in the following subsections.

3.4.3. Atlas of sex differences in secondary progressive MS post-mortem brain tissue

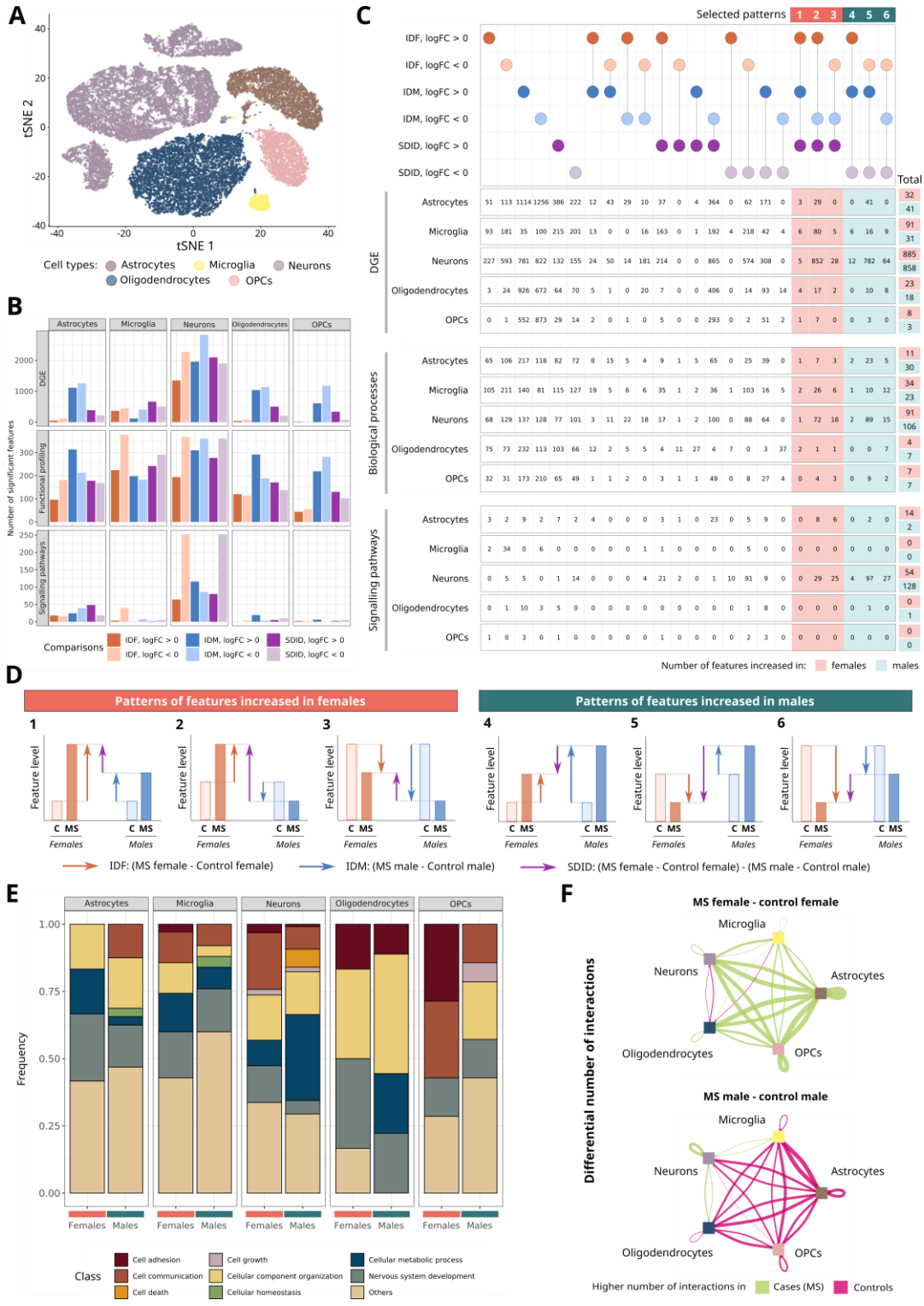
We identified neurons, astrocytes, microglia, oligodendrocytes, and OPCs within the SPMS-CNS dataset (**Figure 3.11-A**). **Supplementary Figure 3.S8-A** describes the expression pattern of the marker genes that confirmed the reference-based annotation, while **Supplementary Figure 3.S8-B** describes the number of cells by type, condition, and sex.

We encountered sex differences in all cell types (**Figure 3.11-B**). Although many genes, functions, and pathways remained specific to one sex, a considerable number presented simultaneous alterations in both sexes, creating a catalog of different patterns previously described in *Materials and Methods* section (**Figure 3.7**). We focused on the features significant in all three comparisons, that is, features that significantly change in the

disease in females (IDF comparison) and males (IDM comparison) but also present significant sex-differential expression in the SDID comparison (**Figure 3.11-C**). These alterations fell into different patterns, which we then grouped into two by their logFC sign in the SDID comparison: i) significant features increased in females (logFC > 0), and ii) significant features increased in males (logFC < 0). Henceforth, we used this classification to illustrate results for simplicity's sake. Further details of this interpretation can be found in the *Materials and Methods* section. **Supplementary Figures 3.S9-3.S13** summarize significant features for each cell type. We observed that most features display cell type-specificity or, at most, were identified significant in two different cell types (**Supplementary Figure 3.S14**).

In general terms, females displayed a higher proportion of functions related to *cell adhesion* (microglia, neurons, oligodendrocytes, and OPCs) and *nervous system development* (astrocytes, neurons, and oligodendrocytes) compared to males (**Figure 3.11-D**). Conversely, males exhibited a higher proportion of functions implicated in *cell death* (neurons), *cellular homeostasis* (astrocytes and microglia), and *cell growth* (OPCs) compared to females. Concerning significant pathway effectors, the most frequent terms increased in females related to *infection diseases*, *cell growth and death* (astrocytes), and *signal transduction* (neurons). Meanwhile, male effectors are mainly related to *signal transduction* (astrocytes) and *nervous system* (neurons) (**Supplementary Figure 3.S15-A**).

Figure 3.11. Transcriptomic landscape of sex differences in secondary progressive MS central nervous system. (*Next page*) (A) Cell type distribution in tSNE dimensions. (B) Number of significant features by cell type, analysis, and comparison. (C) Significant features for each cell type separated by comparison and direction of change (logFC). Dots display the significance of the feature, with specific colors indicating the comparison and logFC sign. Numbers indicate the number of significant features corresponding to the evaluated column in the dot map for each cell type and specific analysis. Colored squares highlight significant features in IDF, IDM, and SDID comparisons corresponding to (D) six qualitatively detailed patterns. (E) Relative frequency distribution of GO terms significantly overrepresented in females (orange) and males (green) by cell type. Individual terms were classified into general categories (see legend). (F) Differential cell-cell communication networks between MS females and control females (top) and MS males and control males (bottom). Colors indicate more interactions in MS (green) or controls (pink); the thickness of the interaction is proportional to the magnitude of change. *C: controls; DGE: differential gene expression; GO: Gene Ontology; IDF: impact of disease in females; IDM: impact of disease in males; MS: multiple sclerosis; OPCs: oligodendrocyte precursor cells; SDID: sex differential impact of disease; tSNE: t-Distributed Stochastic Neighbor Embedding.*

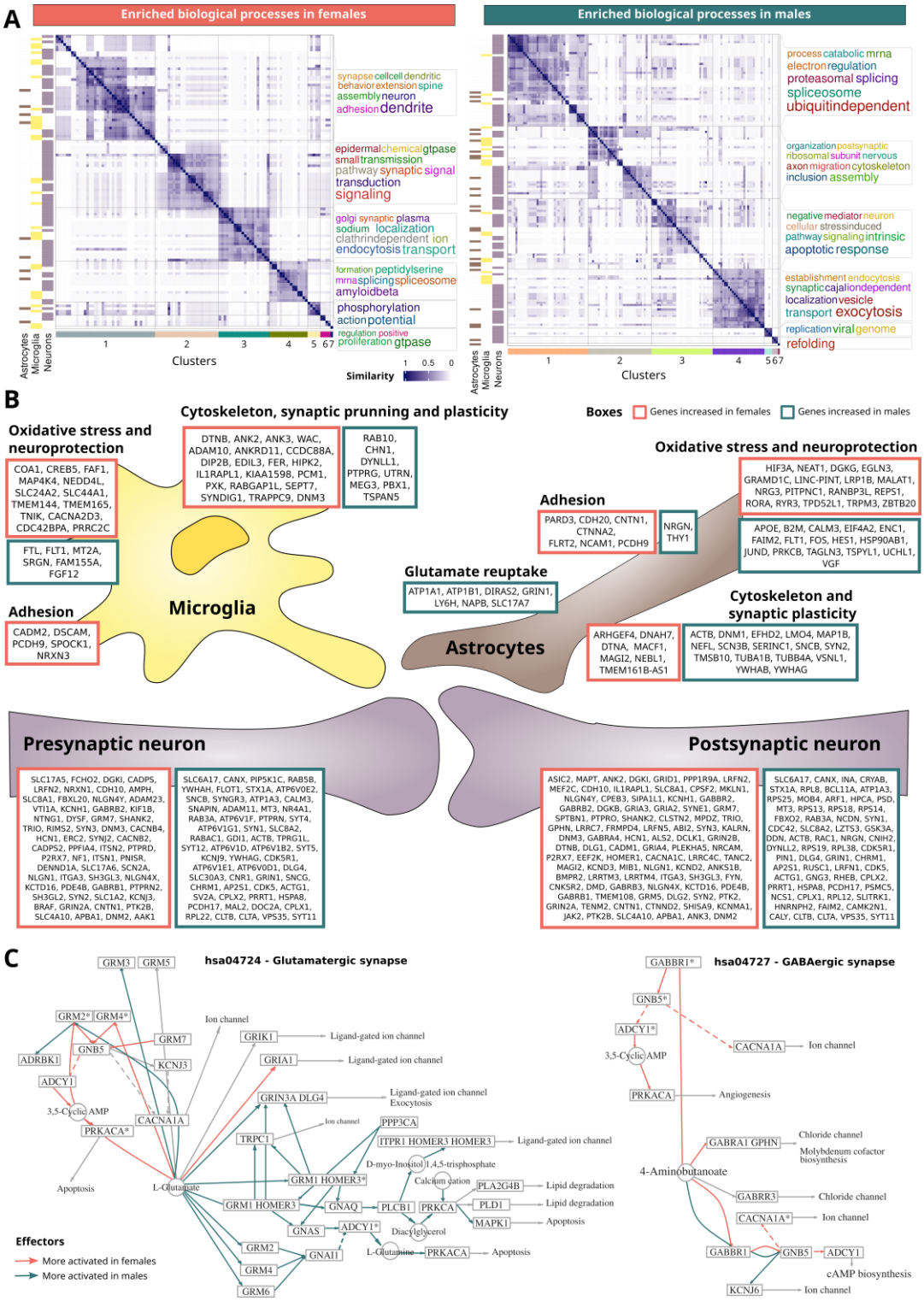

A
tSNE 2
tSNE 1
Cell types: Astrocytes Microglia Neurons Oligodendrocytes OPCs
B
Astrocytes Microglia Neurons Oligodendrocytes OPCs
DGE
Functional profiling
Signalling pathways
Number of significant features
Comparisons
IDF, logFC > 0
IDF, logFC < 0
IDM, logFC > 0
IDM, logFC < 0
SDID, logFC > 0
SDID, logFC < 0
C
Selected patterns 1 2 3 4 5 6
IDF, logFC > 0
IDF, logFC < 0
IDM, logFC > 0
IDM, logFC < 0
SDID, logFC > 0
SDID, logFC < 0
Total
DGE
Biological processes
Signalling pathways
Astrocytes
Microglia
Neurons
Oligodendrocytes
OPCs
Number of features increased in: females males
D
Patterns of features increased in females
Patterns of features increased in males
Feature level
C MS
Females
Males
IDF: (MS female - Control female)
IDM: (MS male - Control male)
SDID: (MS female - Control female) - (MS male - Control male)
E
Astrocytes Microglia Neurons Oligodendrocytes OPCs
Frequency
Females Males
Class
Cell adhesion
Cell communication
Cell death
Cell growth
Cellular component organization
Cellular homeostasis
Cellular metabolic process
Nervous system development
Others
F
MS female - control female
MS male - control male
Microglia
Neurons
Astrocytes
Oligodendrocytes
OPCs
Differential number of interactions
Higher number of interactions in Cases (MS) Controls

Cell-cell communication analyses complete our atlas of sex differences. **Supplementary Figure 3.S15-B** describes the number of inferred interactions. Notably, males and females displayed practically opposite patterns (**Figure 3.11-E**): females presented a greater number of cell-cell interactions in MS, while males in healthy controls. Remarkably, we observed increased neuron-neuron interactions in MS males and MS females, although the value for males exhibits a larger magnitude of change.

3.4.3.1. Sex differential alterations in the astrocyte-microglia-neuron triad implicate synaptic components and stress responses

We performed a semantic clustering analysis for each sex over the enriched biological functions in astrocytes, neurons, and microglia. This approach allowed us to determine whether similar functions were enriched in different cell types or, on the contrary, if most of them were cell type-specific functional changes. Results showed that functional alterations in both males and females were distributed across similar semantic areas, without a clear distinction in clustering patterns pointing to a specific cell type (**Figure 3.12-A**). Notably, the comparison of the word cloud between males and females revealed certain functional similarities. Biological functions related to RNA processing were found to be significant in both sexes (female cluster 4 and male cluster 1), although males specifically presented enriched functions related to protein degradation and apoptosis (male clusters 1 and 3). **Supplementary Table 3.S4 and Supplementary Table 3.S5** provide detailed descriptions of these functions, inferring that males may exhibit a broader stress-related and catabolic processes during MS.

Figure 3.12. Sex differences in secondary progressive MS post-mortem brain tissue synapses. (*Next page*) (A) Degree of similarity between functions overrepresented in females (left) and males (right) for astrocytes, microglia, and neurons. Each row and column correspond to a significant function. Blue intensity indicates the degree of similarity. Left horizontal lines represent term significance in astrocytes, microglia, and/or neurons. Clusters at the bottom are associated with the word cloud on the right of the plot. (B) Atlas of sex differences in synapse-related genes. Significant genes involved directly/indirectly in synapses were selected and classified in broad categories based on their associated functions. Colored boxes indicate sex patterns: genes increased in females (orange) and males (green). (C) Signaling pathways of glutamatergic (left) and GABAergic (right) synapses. Nodes represent proteins of the signaling pathway and edges the interactions between nodes. Effector proteins are the last node in each subpathway (arrow point), pointing to the biological functions they exert. *GABA: Gamma-aminobutyric acid.*


A
Enriched biological processes in females
Enriched biological processes in males
synapse cellcell dendritic behavior extension spine assembly neuron adhesion dendrite
epidermal chemical gtpase small transmission pathway synaptic signal transduction signaling
golgi synaptic plasma sodium localization clathrindependent ion endocytosis transport
formation peptidylserine mrna splicing spliceosome amyloidbeta
phosphorylation action potential
regulation positive proliferation gtpase
process catabolic mrna electron regulation proteasomal splicing spliceosome ubiquitindependent
organization postsynaptic ribosomal subunit nervous axon migration cytoskeleton inclusion assembly
negative mediator neuron cellular stressinduced pathway signaling intrinsic apoptotic response
establishment endocytosis synaptic cajaliondependent localization vesicle transport exocytosis
replication viral genome
refolding
Astrocytes
Microglia
Neurons
Clusters
Similarity
B
Cytoskeleton, synaptic prunning and plasticity
Boxes
Genes increased in females
Genes increased in males
Oxidative stress and neuroprotection
COA1, CREB5, FAF1, MAP4K4, NEDD4L, SLC24A2, SLC44A1, TMEM144, TMEM165, TNIK, CACNA2D3, CDC42BPA, PRRC2C
FTL, FLT1, MT2A, SRGN, FAM155A, FGF12
DTNB, ANK2, ANK3, WAC, ADAM10, ANKRD11, CCDC88A, DIP2B, EDIL3, FER, HIPK2, IL1RAPL1, KIAA1598, PCM1, PXK, RABGAP1L, SEPT7, SYNDIG1, TRAPPC9, DNM3
RAB10, CHN1, DYNLL1, PTPRG, UTRN, MEG3, PBX1, TSPAN5
Adhesion
CADM2, DSCAM, PCDH9, SPOCK1, NRXN3
Microglia
Glutamate reuptake
ATP1A1, ATP1B1, DIRAS2, GRIN1, LY6H, NAPB, SLC17A7
Astrocytes
Oxidative stress and neuroprotection
HIF3A, NEAT1, DGKG, EGLN3, GRAMD1C, LINC-PINT, LRP1B, MALAT1, NRG3, PITPNC1, RANBP3L, REPS1, RORA, RYR3, TPD52L1, TRPM3, ZBTB20
APOE, B2M, CALM3, EIF4A2, ENC1, FAIM2, FLT1, FOS, HES1, HSP90AB1, JUND, PRKCB, TAGLN3, TSPYL1, UCHL1, VGF
Adhesion
PARD3, CDH20, CNTN1, CTNNA2, FLRT2, NCAM1, PCDH9
NRGN, THY1
Cytoskeleton and synaptic plasticity
ARHGEF4, DNAH7, DTNA, MACF1, MAGI2, NEBL1, TMEM161B-AS1
ACTB, DNM1, EFHD2, LMO4, MAP1B, NEFL, SCN3B, SERINC1, SNCB, SYN2, TMSB10, TUBA1B, TUBB4A, VSNL1, YWHAB, YWHAG
Presynaptic neuron
SLC17A5, FCHO2, DGKI, CADPS, LRFN2, NRXN1, CDH10, AMPH, SLC8A1, FBXL20, NLGN4Y, ADAM23, VTI1A, KCNH1, GABRB2, KIF1B, NTNG1, DYSF, GRM7, SHANK2, TRIO, RIMS2, SYN3, DNM3, CACNB4, HCN1, ERC2, SYNJ2, CACNB2, CADPS2, PPFIA4, ITSN2, PTPRD, P2RX7, NF1, ITSN1, PNISR, DENND1A, SLC17A6, SCN2A, NLGN1, ITGA3, SH3GL3, NLGN4X, KCTD16, PDE4B, GABRB1, PTPRN2, SH3GL2, SYN2, SLC1A2, KCNJ3, BRAF, GRIN2A, CNTN1, PTK2B, SLC4A10, APBA1, DNM2, AAK1
SLC6A17, CANX, PIP5K1C, RAB5B, YWHAH, FLOT1, STX1A, ATP6V0E2, SNCB, SYNGR3, ATP1A3, CALM3, SNAPIN, ADAM11, MT3, NR4A1, RAB3A, ATP6V1F, PTPRN, SYT4, ATP6V1G1, SYN1, SLC8A2, RABAC1, GDI1, ACTB, TPRG1L, SYT12, ATP6V1D, ATP6V1B2, SYT5, KCNJ9, YWHAG, CDK5R1, ATP6V1E1, ATP6V0D1, DLG4, SLC30A3, CNR1, GRIN1, SNCG, CHRM1, AP2S1, CDK5, ACTG1, SV2A, CPLX2, PRRT1, HSPA8, PCDH17, MAL2, DOC2A, CPLX1, RPL22, CLTB, CLTA, VPS35, SYT11
Postsynaptic neuron
ASIC2, MAPT, ANK2, DGKI, GRID1, PPP1R9A, LRFN2, MEF2C, CDH10, IL1RAPL1, SLC8A1, CPSF2, MKLN1, NLGN4Y, CPEB3, SIPA1L1, KCNH1, GABBR2, GABRB2, DGKB, GRIA3, GRIA2, SYNE1, GRM7, SPTBN1, PTPRO, SHANK2, CLSTN2, MPDZ, TRIO, GPHN, LRRC7, FRMPD4, LRFN5, ABI2, SYN3, KALRN, DNM3, GABRA4, HCN1, ALS2, DCLK1, GRIN2B, DTNB, DLG1, CADM1, GRIA4, PLEKHA5, NRCAM, P2RX7, EEF2K, HOMER1, CACNA1C, LRRC4C, TANC2, MAGI2, KCND3, MIB1, NLGN1, KCND2, ANKS1B, BMPR2, LRRTM3, LRRTM4, ITGA3, SH3GL3, FYN, CNKSR2, DMD, GABRB3, NLGN4X, KCTD16, PDE4B, GABRB1, TMEM108, GRM5, DLG2, SYN2, PTK2, GRIN2A, TENM2, CNTN1, CTNND2, SHISA9, KCNMA1, JAK2, PTK2B, SLC4A10, APBA1, ANK3, DNM2
SLC6A17, CANX, INA, CRYAB, STX1A, RPL8, BCL11A, ATP1A3, RPS25, MOB4, ARF1, HPCA, PSD, MT3, RPS13, RPS18, RPS14, FBXO2, RAB3A, NCDN, SYN1, CDC42, SLC8A2, LZTS3, GSK3A, DDN, ACTB, RAC1, NRGN, CNIH2, DYNLL2, RPS19, RPL38, CDK5R1, PIN1, DLG4, GRIN1, CHRM1, AP2S1, RUSC1, LRFN1, CDK5, ACTG1, GNG3, RHEB, CPLX2, PRRT1, HSPA8, PCDH17, PSMC5, NCS1, CPLX1, RPL12, SLITRK1, HNRNPH2, FAIM2, CAMK2N1, CALY, CLTB, CLTA, VPS35, SYT11
C
hsa04724 - Glutamatergic synapse
GRM3
GRM5
GRM2*
GRM4*
GRM7
GNB5
KCNJ3
ADRBK1
ADCY1
3,5-Cyclic AMP
PRKACA*
CACNA1A
Apoptosis
L-Glutamate
Ion channel
GRIK1
Ligand-gated ion channel
GRIA1
Ligand-gated ion channel
GRIN3A DLG4
Ligand-gated ion channel
Exocytosis
Ion channel
PPP3CA
TRPC1
ITPR1 HOMER3 HOMER3
Ligand-gated ion channel
D-myo-Inositol 1,4,5-trisphosphate
GRM1 HOMER3*
Calcium ation
PLA2G4B
Lipid degradation
GRM1 HOMER3
GNAQ
PLCB1
PRKCA
PLD1
Lipid degradation
MAPK1
Apoptosis
GNAS
ADCY1*
Diacylglycerol
GRM2
GNAI1
L-Glutamine
PRKACA
Apoptosis
GRM4
GRM6
Effectors
More activated in females
More activated in males
hsa04727 - GABAergic synapse
GABBR1*
GNB5*
ADCY1*
3,5-Cyclic AMP
CACNA1A
Ion channel
PRKACA
Angiogenesis
GABRA1 GPHN
Chloride channel
Molybdenum cofactor biosynthesis
4-Aminobutanoate
GABRR3
Chloride channel
CACNA1A*
Ion channel
GABBR1
GNB5
ADCY1
cAMP biosynthesis
KCNJ6
Ion channel

The remaining clusters, both in female- and male-enriched scenarios, were predominantly associated with synaptic processes, suggesting that sex-related differences may affect the synaptic transmission. To further characterize these differences, we examined the genes directly or indirectly linked to synaptic features that were differentially expressed between sexes. As shown in **Figure 3.12-B**, neurons exhibited a wide range of sex-biased changes in both presynaptic and postsynaptic associated genes. Interestingly, some gene families presented higher expression in one sex. Of note, several voltage-gated calcium channel subunits (CACNA1C, CACNB2, and CACNB4) were upregulated in female neurons, potentially affecting the calcium-dependent synaptic activity. In contrast, male neurons exhibited increased expression of various ATPase subunits (ATP6V0E2, ATP1A3, ATP6V1F, ATP6V1G1, ATP6V1D, ATP6V1B2, ATP6V1E1, and ATP6V0D1), which may be involved in maintaining ion gradients, vesicle acidification, and neurotransmitter packaging. Furthermore, female neurons showed increased expression of genes involved in modulating excitability, including both glutamate receptors (GRIA2 and GRIA3) and GABA receptors (GABRB1 and GABBR2). Pathway activation analyses revealed that glutamatergic effectors, typically associated with increased neuronal excitability, were higher activated in males (**Figure 3.12-C, left**). Meanwhile, GABAergic effectors, linked to inhibitory signaling, showed greater activation in females (**Figure 3.12-C, right**). Cholinergic, serotonergic, and dopaminergic synapses displayed further significant differences (**Supplementary Figure 3.S16**), sparking an intricate network of sex disparities in neuronal excitability. Altogether, these findings highlight a complex sex-dependent modulation of the synaptic activity.

Astrocytes and microglia presented significant differences in gene expression profiles associated with neuroinflammation and tissue homeostasis (**Figure 3.12-B,** *Oxidative stress and neuroprotection* and *Glutamate reuptake* panels). Intriguingly, male astrocytes expressed increased levels of glutamate reuptake-related genes, suggesting enhanced capacity for extracellular glutamate clearance. Sex-differential stress responses in astrocytes also included the increased expression of genes related to calcium homeostasis (RYR3, TPD52L1, and TRPM3 in females; CALM3 in males), oxidative stress (REPS1 in females; ENC1, HSP90AB1, PRKCB, and UCHL1 in males), hypoxia (HIF3A, EGLN3, and ZBTBW in females; EIF4A2 in males) and neuroprotection (NRG3, RORA, NEAT1, RANBP3L, MALAT1, and LINC-PINT in females; JUND, FOS, FLT1, FAIM2 and HES1 in males).

Meanwhile, female microglia notably increased the expression of genes associated with transmembrane transporters involved in metabolism and homeostasis maintenance, such as SLC24A2 (calcium/cation antiporter), SLC44A1 (choline transporter), TMEM144

(carbohydrate transporter), TMEM165 (cation/proton antiporter) and CACNA2D3 (calcium transporter), while male microglia increased the expression of genes associated with intracellular metal homeostasis (FTL and MT2A).

Additionally, female and male microglia displayed increased expression levels of other oxidative stress-related genes (COA1, CREB5, and FAF1 in females; SRGN and FGF12 in males). These glial cells also exhibited the dysregulated expression of genes potentially involved in myelin recovery, which are illustrated in the following section.

3.4.3.2. Sex differential alterations in secondary progressive MS post-mortem brain tissue also affect lipid metabolism and myelin recovery

We next explored oligodendrocytes and OPCs to gain insight into the sex differential potential for myelin repair and its maintenance. Neuronal-related functions and myelination represented the major altered categories in both cell types (**Figure 3.13**); however, the associated genes differ between sexes (**Supplementary Table 3.S6**).

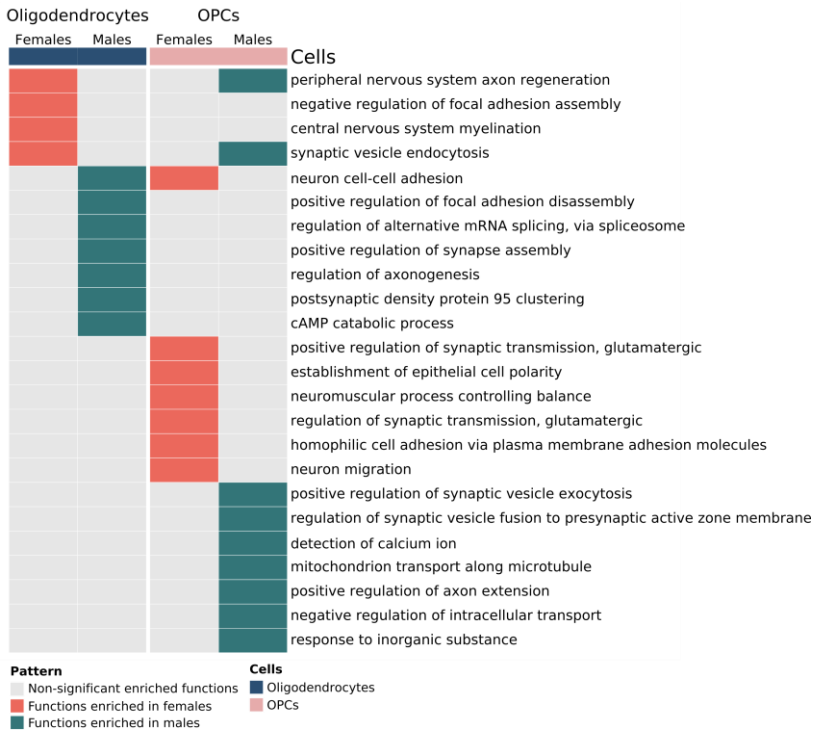


Figure 3.13. Significant enriched functions in oligodendrocytes and oligodendrocyte precursor cells in secondary progressive multiple sclerosis. *OPCs: oligodendrocyte precursor cells*.

Specifically, oligodendrocytes differed in genes encoding neuronal adhesion and synaptic molecules that support the maintenance of myelin targeting to axons, with the majority of genes increased in females (NCAM2, NLGN1, CADM2, CLDN11, ANK3, IL1RAPL1, CTNND2, and LPHN3); and CAMK2B and NRGN in males. Females also displayed the increased expression of genes to promote myelin repair, such as AMER2 (negative regulator of the Wnt/ß-catenin pathway) and OMALINC (lncRNA marker of oligodendrocyte maturation that regulates myelination-associated gene expression). Interestingly, males displayed the increased expression of marker genes for oligodendrocyte differentiation (CAMK2B and MIR219A2) and myelinating activity (MAPB1, BCAS1, and MGAT5).

Among the genes with higher expression in female OPCs, we encountered genes involved in differentiation and myelin repair processes (GPNMB, PARD3, QKI, and TNR) and neuroprotection (HIF3A, and VCAN), while male OPCs only display one gene (MAP1B) related with these functions.

We next characterized platelet-derived growth factor (PDGF) and fibroblast growth factor (FGF) communication signaling to OPCs (**Figure 3.14**), given that these two pathways drive OPCs to growth and differentiate into oligodendrocytes to promote myelin repair. The PDGF and FGF signaling displayed higher interaction strengths in MS males and MS females compared to corresponding controls (**Figure 3.14-A-B, left**). However, the contribution of donor cells and the strength of ligand-receptor pairs differed according to sex, with more pronounced alterations observed in MS females (**Figure 3.14-A-B, right**). In PDGF signaling, the MS female PDGFC (OPCs) - PDGFRA (OPCs) interaction increased proportionally compared to control females, to the detriment of interactions with neurons-OPCs. In FGF signaling, ligands provided by oligodendrocytes increased in interaction strength proportionally in MS females compared to female controls, while astrocyte ligand interactions decreased. Notably, we did not observe these changes in PDGF and FGF signaling when comparing MS males to male controls.

Together, these results suggest that females may have a greater myelin repair potential through the expression of numerous myelin-related genes, whereas males appear to exhibit a more differentiated oligodendrocyte profile.

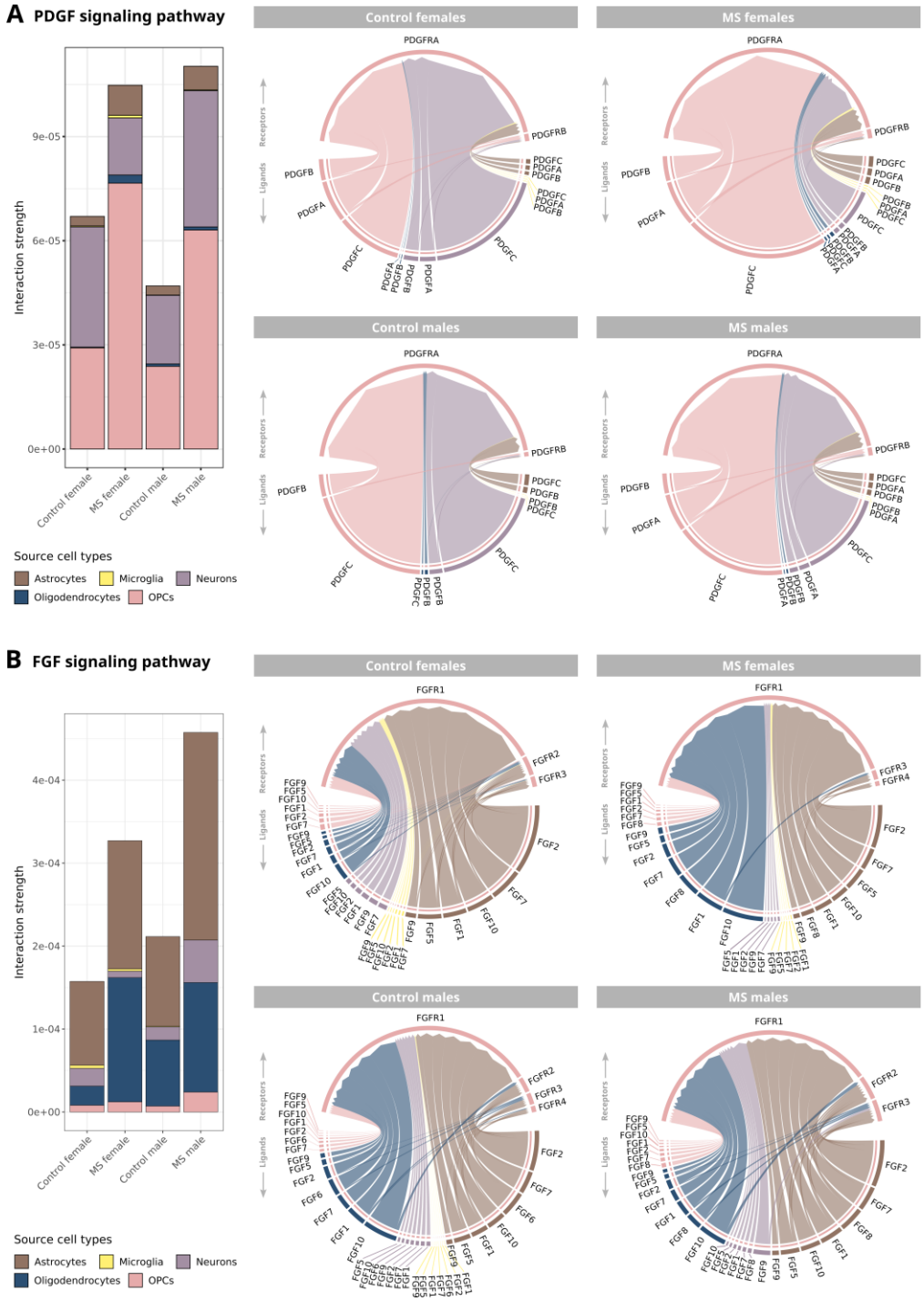


Figure 3.14. (A) PDGF and (B) FGF inferred signaling received by oligodendrocyte precursor cells to promote their growth and differentiation. (Right) Total interaction strength by cell type for each group. (Left) Interaction strength frequency of ligand-receptors pairs. Outer

circumference labels the cell type. Inner flux represents the proportional interaction strength colored by the sender cell type. Sender cells: astrocytes, microglia, neurons, oligodendrocytes, and OPCs. Receiver cells: OPCs. *FGF: fibroblast growth factor; MS: multiple sclerosis; OPCs: oligodendrocyte precursor cells; PDGF: platelet-derived growth factor*.

Astrocytes and microglia also displayed significant sex differences in the expression of lipid metabolism-associated genes, which may be involved in myelin clearance. We observed more genes with increased expression in female than male astrocytes (e.g., DGKG, GRAMD1C, LRP1B, and PITPNC1). However, males displayed APOE increased expression, which could potentiate lipid droplet accumulation. Concurrently, female microglia displayed the increased expression of lipid-related genes such as CELF2, CHST11, AGO3, ATG4C, GPM6B, LPAR1, LPGAT1, OSBPL1A, PIP4K2A, PLD1, QKI, SOX2-OT, and TMEM131, while male microglia increased the expression of RAB10, TNS3, COLEC12, PLSCR1, SCARB1, and VIM.

We propose these results may constitute adaptive differences in lipid metabolism and debris clearance from damaged myelin in MS, where females appear to show a more pronounced adaptive response.

3.4.4. ATLAS OF SEX DIFFERENCES IN RELAPSING-REMITTING MS PERIPHERAL BLOOD MONONUCLEAR CELLS

Next, we focused on the PBMCs in RRMS subtype (**Figure 3.15-A**). **Supplementary Figure 3.S17** contains the expression of marker genes by cell type, cell distribution by group and sex, and the number of statistically significant features; **Supplementary Figures 3.S18-3.S22** include an overview of the results for each cell type. In this scenario, we illustrated significant features in our three comparisons (IDF, IDM and SDID) as previously defined for SPMS in the CNS. Due to the limited number of results obtained for B cells and dendritic cells, we focused on CD4+ T cells, CD8+ T cells, NK cells and monocytes.

In all cell types, females exhibited a higher number of significant functions than males in mostly all categories (**Figure 3.15-B**), which agrees with the higher number of genes with significantly increased expression compared to males (**Figure 3.15-C**). We noted an exception in the *Metabolism and bioenergetics* category for male CD4+ T cells, CD8+ T cells, and NK cells, which mainly relates to the mitochondrial electron transport. Lastly, the differential activation of signaling pathways and cell-cell communication interactions comprised an intricate framework of sex differences (**Supplementary Figure 3.S23**).

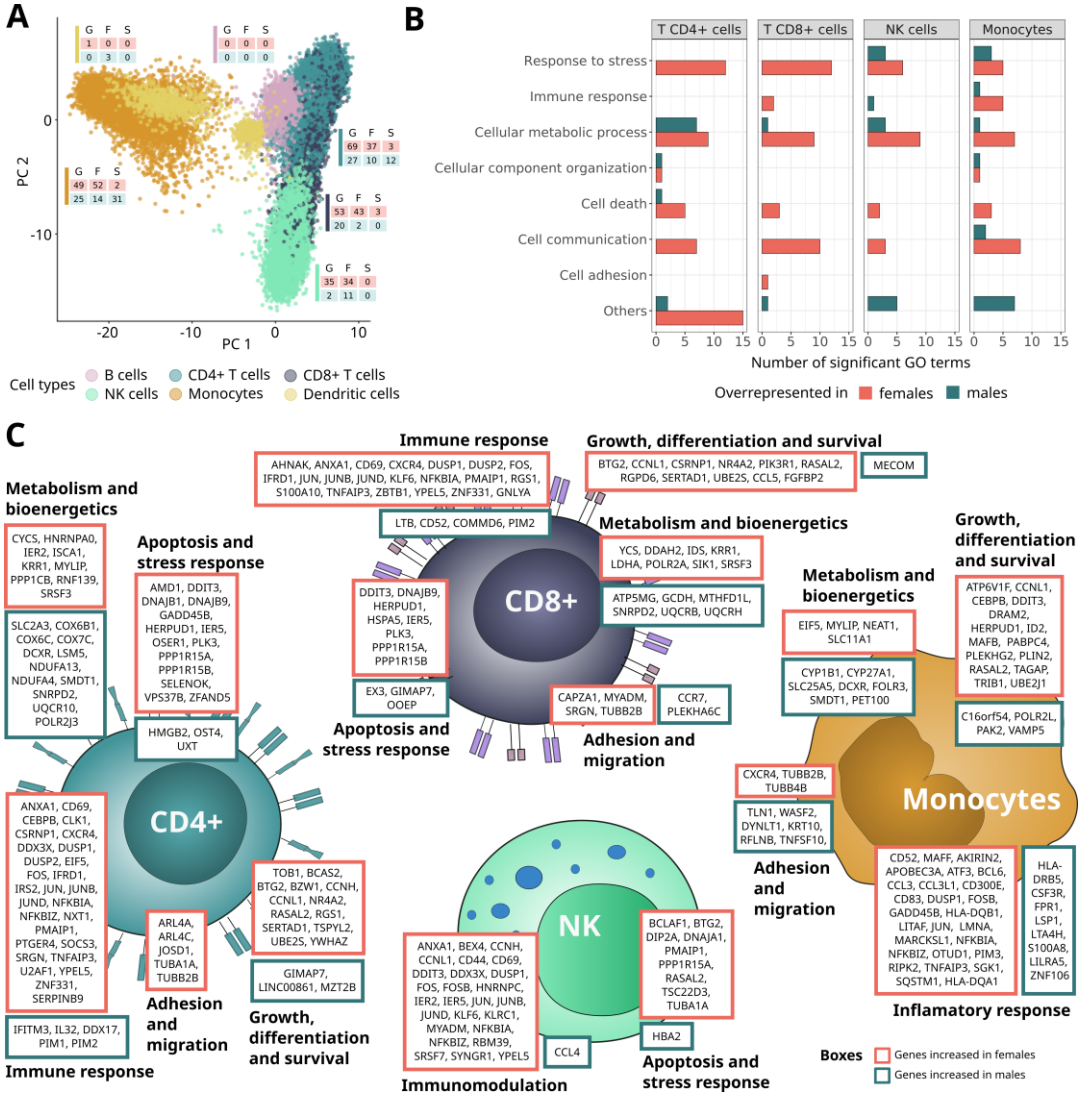


Figure 3.15. Transcriptomic landscape of sex differences in relapsing-remitting MS peripheral blood mononuclear cells. (A) Cell type distribution. Colored boxes indicate that the number of genes (G), functions (F), and signaling pathways (S) significantly increased in females (orange) and males (blue). (B) Number of significant GO terms by cell type classified under broader biological categories. (C) Atlas of sex differences in gene expression patterns. Genes classified in general biological terms. Colored boxes indicate the sex patterns: genes increased in females (orange) and males (green). *GO: Gene Ontology; MS: multiple sclerosis; NK: natural killer; PC: principal component.*

3.4.4.1. Intersection analysis reveals an immune signature core in relapsing-remitting MS females with the implication of the AP-1 transcription factor

In RRMS, we identified 21 genes with increased expression in females compared to males that were significant in at least three of the four evaluated cell types (**Figure 3.16-A**), providing a general insight into the status of the female innate and adaptive immune systems. Indeed, these genes constitute a significantly connected network (p-value: < 1.0e-16) (**Figure 3.16-B**), where the prominent hub comprises the AP-1 transcription factor complex (FOS, JUN, JUNB, and JUND). However, the definition of the inflammatory-related differences is complex, as we also identified the inhibitors NFKBIA and NFKBIZ, which may limit NF-κB driven inflammation, and the increased expression of the biomarker CD69, indicative of a pro-inflammatory scenario.

We did not identify a common gene core for males. The C16orf54 gene —associated with immune infiltration— represents the only significantly increased gene in male CD4+ T cells, CD8+ T cells and monocytes (**Supplementary Figure 3.S24**).

Alterations in the transcriptomic profile were also reflected in the inferred differential interaction strength of ligand-receptors pairs in MS females compared to control females, where we detected the increased interaction of signaling pathways mediated by proinflammatory cytokines (e.g., IL6) and cell adhesion molecules (e.g., CD99 and cadherins) that do not become reinforced in males (**Supplementary Figure 3.S23-B**).

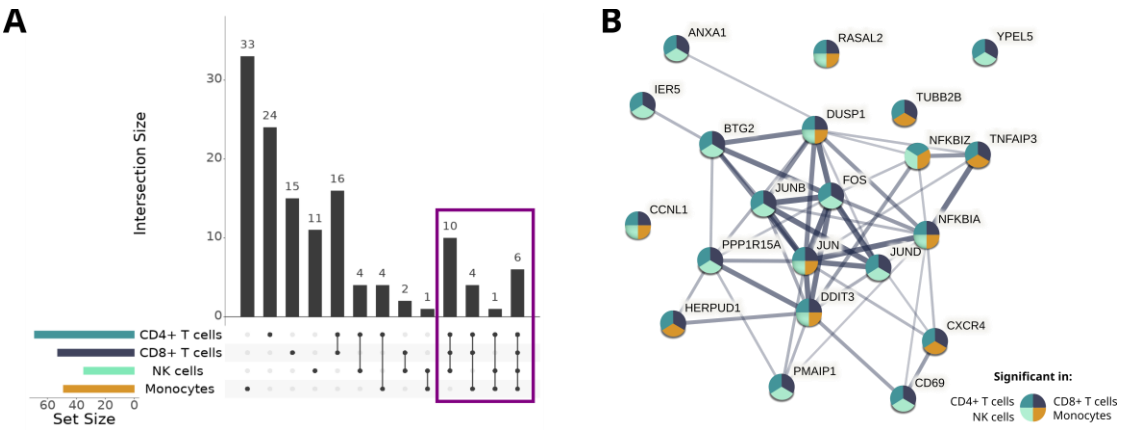


Figure 3.16. Female immune signature in RRMS. (A) Upset plot of genes with increased expression in females compared to males by cell type. Horizontal bars represent the number of significant genes in each cell type. Dots indicate the combinations of intersections evaluated, with the top vertical bars denoting the number of significant genes of the corresponding intersection. Purple square: significant genes in at least three evaluated cell types. (B) Protein-

protein interaction networks of significant genes with increased expression in females compared to males in at least three cell types evaluated: CD4+ T cells, CD8+ T cells, NK cells and monocytes. Edge thickness indicates the structural and functional confidence of the interaction. *NK: natural killer*.

3.4.4.2. The adaptive immune response in relapsing-remitting MS males exhibits exacerbated mitochondrial dynamics compared to females

We previously noted the absence of a central hub of genes with increased expression in the immune system of MS males; however, we found intriguing results by separating the innate and adaptive immune systems. The adaptive immune system presents different genes functionally related to the mitochondrial electron transport chain (**Figure 3.15-C**). CD4+ T cells displayed the increased expression of genes from all steps: NADH to ubiquinone (NDUFA13 and NDUFA4), ubiquinol to cytochrome c (UQCR10), and cytochrome c to oxygen (COX6B1, COX6C, and COX7C), while CD8+ T cells displayed the increased expression of genes for two additional ubiquinol-cytochrome C reductases (UQCRB and UQCRH) and ATP5MG, a gene involved in ATP synthesis coupled with proton transport. Meanwhile, male monocytes also increased the expression of genes that favor shape-shifting infiltration (DYNLT1, TLN1, and RFLNB) and mediate innate inflammation (LTA4H, S100A8, and LILRA2) (**Figure 3.15-C**). Male monocytes also display an elevated activation of pathways primarily involved in cell adhesion and inflammatory modulation compared to females (**Supplementary Figure 3.S23-A**). Notably, we observed an increase in the interaction of receptor-ligand pairs considering all cell types, which involved response modulation via semaphorins and the T-lymphocyte co-stimulatory molecules CD80 and CD86 (**Supplementary Figure 3.S23-C**).

3.4.5. Atlas of sex differences in primary progressive MS peripheral blood mononuclear cells

We also explored the sex differential transcriptomic profiles of peripheral immune cells in PPMS (**Figure 3.17-A**). We observed significant differences across all cell types, with CD8+ T cells showing the largest number of significant findings. **Supplementary Figure 3.S25** reports the expression of marker genes by cell type, cell distribution by group and sex, and the number of statistically significant features, while **Supplementary Figures 3.S26-3.S30** summarize the results obtained for each cell type. We identified significant functions in broad biological categories with similar proportions between

males and females (**Figure 3.17-B**), revealing sex alterations in the immune transcriptomic profile that impact several functional categories (**Figure 3.17-C**).

We also encountered sex differences affecting the activation of signaling pathway effectors (**Supplementary Figure 3.S31**) and the strength of interaction between ligand-receptor pairs (**Supplementary Figure 3.S32**). In the latter, we noted a lack of commonly activated pathways in MS males (*vs.* control males) and MS females (*vs.* control females).

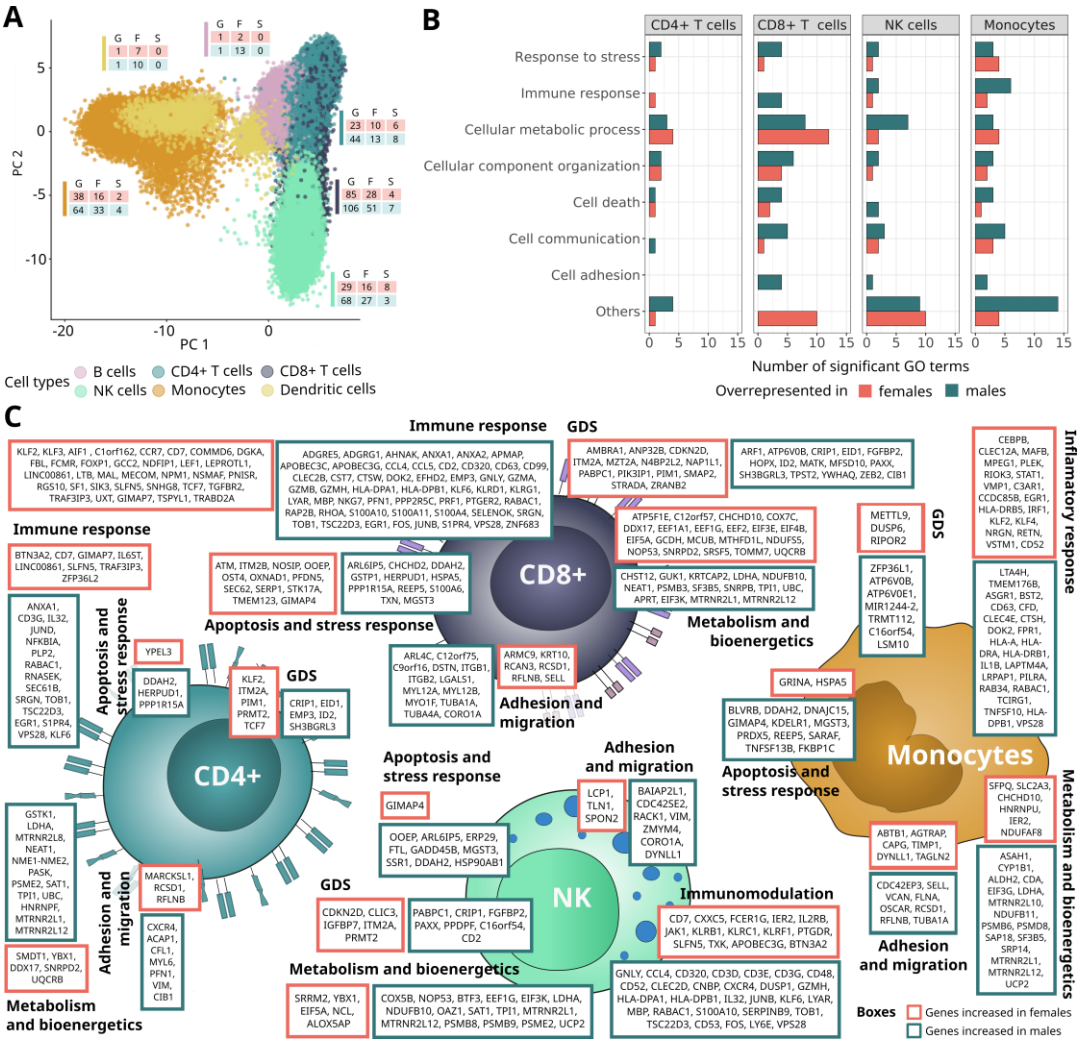


Figure 3.17. Transcriptomic landscape of sex differences in primary progressive MS peripheral blood mononuclear cells. (A) Cell type distribution. Colored boxes indicate that the number of genes (G), functions (F), and signaling pathways (S) significantly increased in females (orange) and males (blue). (B) Number of significant GO terms by cell type classified under

broader biological categories. (C) Atlas of sex differences in gene expression patterns. Genes classified in general biological terms. Colored boxes indicate the sex patterns: genes increased in females (orange) and males (green). *GDS: growth, differentiation and survival; GO: Gene Ontology; MS: multiple sclerosis; NK: natural killer; PC: principal component.*

3.4.5.1. Marked sex differences in CD8+ T cells describe predominant cytolysis in males and homeostatic processes in females

CD8+ T cells exhibited the most pronounced sex differences as previously described (**Figure 3.17**). The genes with sex-differential increased expression formed highly connected protein-protein interaction networks for females and males, with p-values of 3.84×10^{-14} and 1.0×10^{-16}, respectively (**Figure 3.18**). We performed functional enrichment analysis of these genes to identify the biological processes potentially underlying the sex-differential immune responses. We inferred that female CD8+ T cells may restore cellular homeostasis to a greater degree than males, as females displayed the increased expression of genes related to the regulation of protein translation (red cluster), the differentiation and survival processes of T lymphocytes (ochre cluster), and energy production through oxidative phosphorylation and mitochondrial maintenance (yellow cluster) (**Figure 3.18-A**). Conversely, male CD8+ T cells may exhibit a more activated state thanks to higher cytolytic responses through granzyme and perforin-mediated apoptotic processes (purple cluster), calcium regulation (light green cluster), and the formation of extracellular vesicles (park green cluster) (**Figure 3.18-B**).

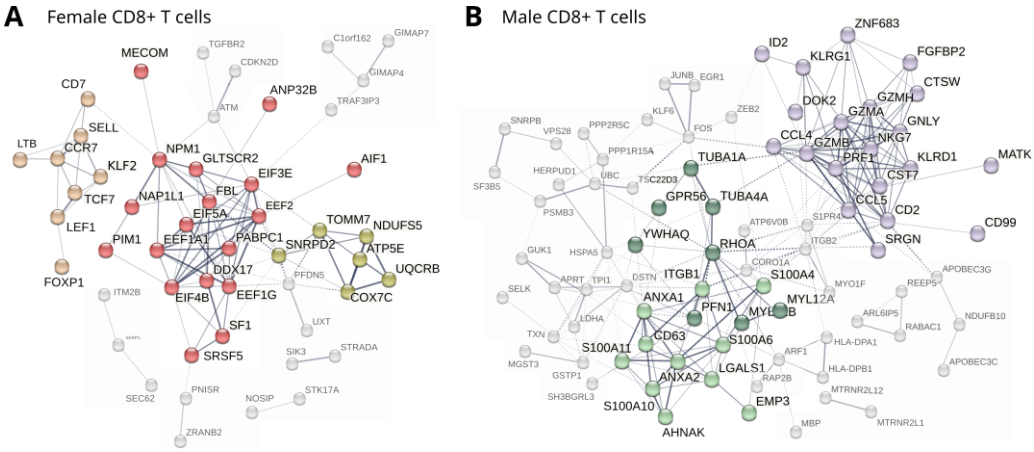


Figure 3.18. Protein-protein interaction networks for (A) female and (B) male CD8+ T cells. Edge thickness indicates the structural and functional confidence of the interaction, representing the intra-group (solid line) and inter-group (dotted line) connections. Clusters of interest are highlighted with colors.

3.4.6. SEX DIFFERENCES IN IMMUNE SYSTEM STATUS CLUSTER RELAPSING-REMITTING MS AND PRIMARY PROGRESSIVE MS CELL TYPES

After independently delineating sex-differential patterns in immune cells for RRMS and PPMS, we sought to explore the significant disparities across these MS subtypes. To this end, we clustered the immune cells from both subtypes ($CD4^+$ T cells, $CD8^+$ T cells NK cells and monocytes) based on the expression profiles of the genes previously identified as significantly different between sexes. This approach allowed us to determine which cell populations exhibited greater similarity, evaluating if they grouped by MS subtype, by cell type or without any known distribution.

We performed the clustering analysis using the gene expression profiles with broader sex differential patterns, that is, genes significantly different at least in three of the eight cell types evaluated. This analysis revealed an initial stratification of cells primarily by disease subtype (RRMS or PPMS), followed by a secondary separation based on their immunological classification as part of either the adaptive or innate immune system (**Figure 3.19**). These findings suggest that critical disparities in the immune system impact the clinical variability observed between MS subtypes, rather than subtle variations within a specific cell type. The 67 genes that supported clustering formed a highly connected interaction network (p-value: < 1.0e-16, **Supplementary Figure 3.S33**) primarily related to stimuli responses such as reactive oxygen species, cytokines, lipids, and leukocyte differentiation.

We took a closer look at these genes. RRMS significant genes displayed an increased expression in females, with a specific core that does not display significant dysregulation in PPMS (**Figure 3.19, gray box**). This RRMS female gene set reflected an activated state (CD69) driving responses, especially to stress and apoptosis (e.g., IER5, DDIT3, and BTG2). Conversely, the majority of significant genes in PPMS were increased in males, with some exhibiting an increase in RRMS females (**Figure 3.19, yellow box**). A high proportion of PPMS genes increased in males related to the modulation of inflammation, along with processes not highlighted in the RRMS female core, such as glycolytic metabolism (LDHA and TPI1) and mitochondrial protein translation (MTRNR2L1 and MTRNR2L12). Interestingly, the adaptive immune system presented a gene set patterned inversely based on disease subtype, a scenario we did not observe in the innate immune system (**Figure 3.19, orange box**). Of note, ANXA1 and SRGN, which are involved in CNS infiltration, showed increased expression in RRMS females and PPMS males.

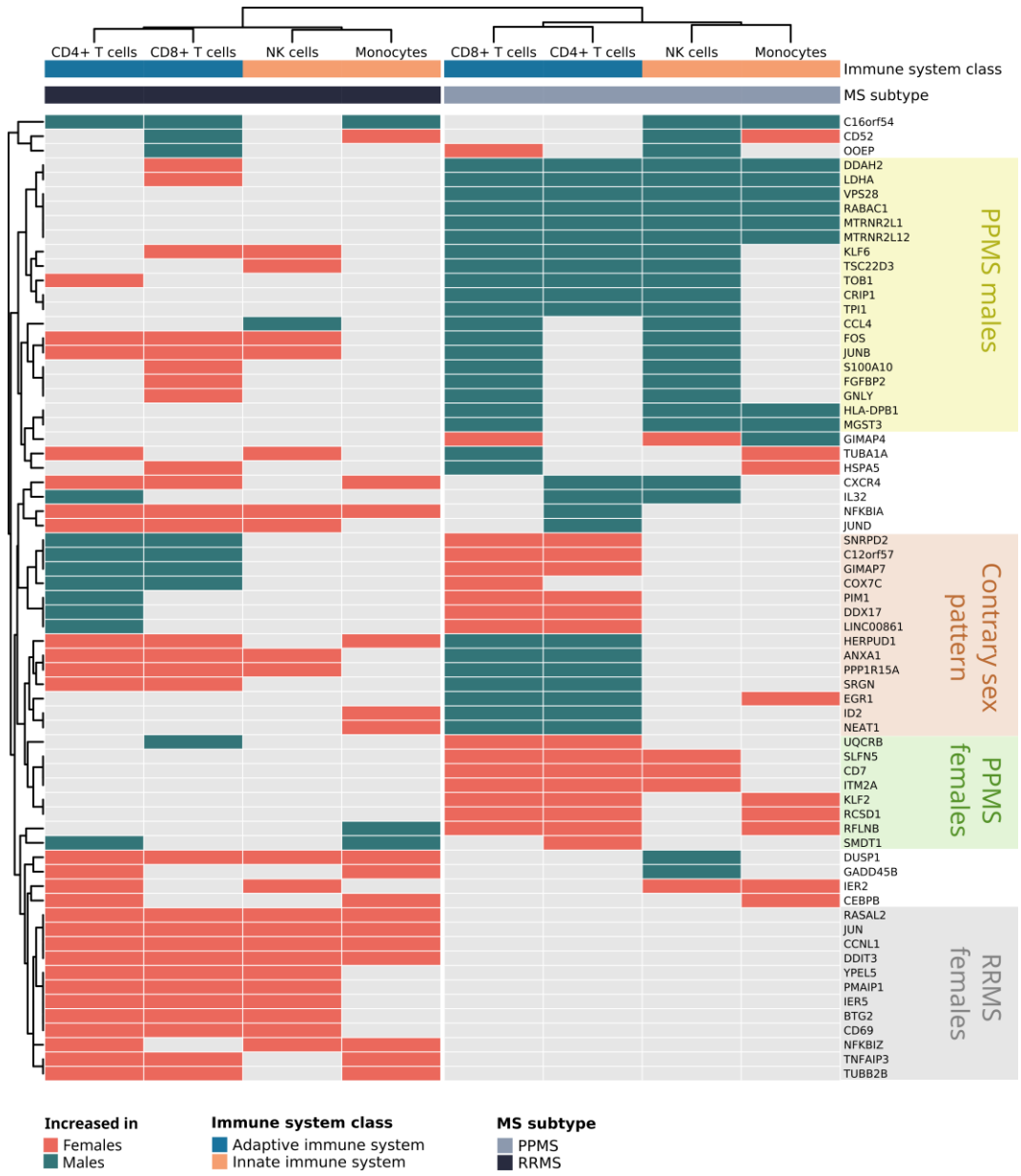


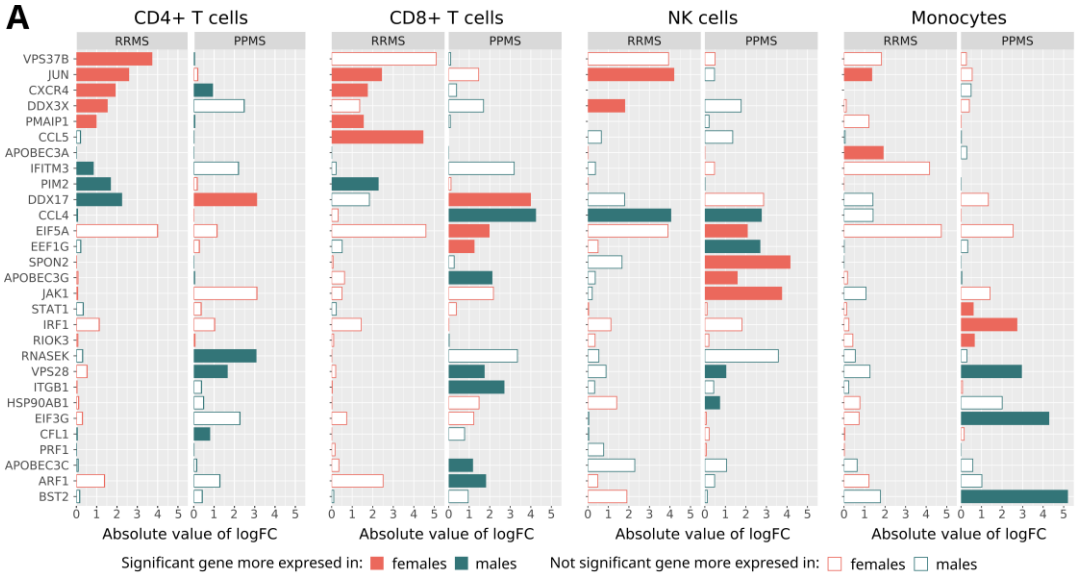
Figure 3.19. Clustering of relapsing-remitting MS and primary progressive MS immune system cell types based on their sex differential profile. Heatmap showing the classification of CD4+ T cells, CD8+ T cells, NK cells, and monocytes in RRMS and PPMS subtypes (columns) from genes with significant expression by sex in at least three of the evaluated cell types (rows). Disease subtype, immune system class, and the sex in which the significant increment is observed are specified. *MS: multiple sclerosis; NK: natural killer; PPMS: primary progressive MS; RRMS: relapsing-remitting MS.*

3.4.7. SEX DIFFERENTIAL VIRAL RESPONSES AND ANTIGEN PRESENTATION BY MS SUBTYPE

Considering the extensive association of MS susceptibility with viral responses, we also investigated the potential existence of sex differences in this regard. RRMS females, compared to males, presented more significant functions related to viral infection, while the opposite pattern occurred for PPMS. In detail, RRMS females exhibited the significant functions *positive regulation by host of viral transcription* (GO:0043923) and *regulation of viral process* (GO:0050792) in CD4+ T cells, *cellular response to virus* (GO:0098586) in NK cells, and *negative regulation by host of viral transcription* (GO:0043922) in monocytes. We found no significant viral-related function in RRMS males. However, PPMS males displayed the significant functions *positive regulation of defense response to virus by host* (GO:0002230) in CD8+ T cells, *response to virus* (GO:0009615) in NK cells, and *defense response to virus* (GO:0051607) in monocytes. The latter represents the only virus-related function overrepresented in PPMS females, also in monocytes. The same tendency was observed when evaluating the differential expression of significant genes belonging to these functions (**Figure 3.20-A**).

HLA genes play a critical role in the immune response by presenting foreign antigens to T cells; furthermore, they can influence MS susceptibility/protection by regulating the immune system's response to self-antigens. We observed sex-differential expression of genes from the HLA family, primarily HLA-DP and HLA-DR subfamilies. The highest number of significant results was found in PPMS male cell types. (**Supplementary Table 3.S7**).

MHCs are the protein products of HLA genes. Here, we inferred sex differences in their communication strengths (**Figure 3.20-B**). We encountered the broadest sex difference in MHC-I for RRMS and PPMS between the HLA-E ligand with the CD94:NKG2A and KLRC1 receptors (**Figure 3.20-B, top**). These interactions became significant for RRMS females and PPMS males compared to the rest of the groups. Regarding MHC-II, we also observed stronger interactions in female controls than male controls at the baseline (**Figure 3.20-B, bottom**). Considering this baseline, RRMS males displayed increased interaction strengths in MHC-II to reach the same level as RRMS females (**Figure 3.20-B, bottom**). Conversely, ligand-receptor interactions disappeared in PPMS females compared to control females, while in PPMS males maintain a state similar to control males (**Figure 3.20-B, bottom**). Notably, the differential pattern identified between RRMS and PPMS controls for MHC-II interaction could be age-dependent, as the mean age for the PPMS cohort (48.26) is higher than the RRMS cohort (35.57).


A
CD4+ T cells
CD8+ T cells
NK cells
Monocytes
RRMS
PPMS
VPS37B
JUN
CXCR4
DDX3X
PMAIP1
CCL5
APOBEC3A
IFITM3
PIM2
DDX17
CCL4
EIF5A
EEF1G
SPON2
APOBEC3G
JAK1
STAT1
IRF1
RIOK3
RNASEK
VPS28
ITGB1
HSP90AB1
EIF3G
CFL1
PRF1
APOBEC3C
ARF1
BST2
0 1 2 3 4 5
Absolute value of logFC
Significant gene more expresed in: females males
Not significant gene more expresed in: females males

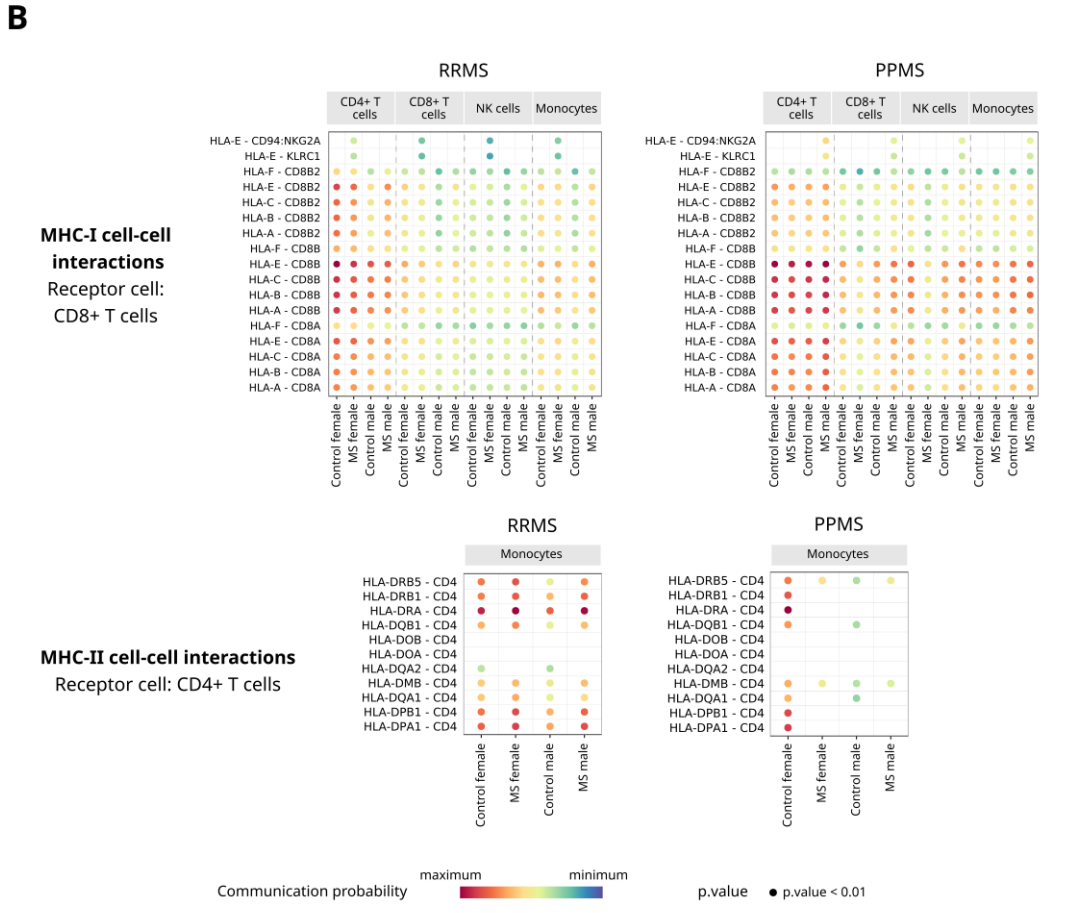

B
MHC-I cell-cell interactions
Receptor cell:
CD8+ T cells
RRMS
PPMS
CD4+ T cells
CD8+ T cells
NK cells
Monocytes
HLA-E - CD94:NKG2A
HLA-E - KLRC1
HLA-F - CD8B2
HLA-E - CD8B2
HLA-C - CD8B2
HLA-B - CD8B2
HLA-A - CD8B2
HLA-F - CD8B
HLA-E - CD8B
HLA-C - CD8B
HLA-B - CD8B
HLA-A - CD8B
HLA-F - CD8A
HLA-E - CD8A
HLA-C - CD8A
HLA-B - CD8A
HLA-A - CD8A
Control female
MS female
Control male
MS male
MHC-II cell-cell interactions
Receptor cell: CD4+ T cells
HLA-DRB5 - CD4
HLA-DRB1 - CD4
HLA-DRA - CD4
HLA-DQB1 - CD4
HLA-DOB - CD4
HLA-DOA - CD4
HLA-DQA2 - CD4
HLA-DMB - CD4
HLA-DQA1 - CD4
HLA-DPB1 - CD4
HLA-DPA1 - CD4
Communication probability
maximum
minimum
p.value
p.value < 0.01

Figure 3.20. Sex differences in viral responses and antigen presentation between relapsing-remitting MS and primary progressive MS. (*Previous page*) (A) Patterns of gene expression alterations associated with viral responses by cell type and disease subtype. The magnitude of change in the SDID comparison (X-axis) for each gene extracted from the response to the virus and viral process (Y-axis) functions. Genes were selected as being significant in at least one cell type. (B) Significant ligand-receptor pairs for MHC-I (top) and MHC-II (bottom) interactions. Dot presence indicates a significant interaction. Color reflects the ligand-receptor pair's interaction strength. *MHC: Major histocompatibility complex; NK: natural killer; PPMS: primary-progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis; SDID: sex differential impact of disease.*

3.4.8. FINDINGS RECAPITULATION AND WEB PLATFORM

Throughout this *Results* section, we have described how sex differences manifest across the three MS subtypes, considering both CNS and PBMCs samples. Our findings focused on the description that females increased potential protective mechanisms against neurodegeneration. Specifically, female neurons may potentiate GABAergic synapses; female microglia may increase genes potentially involved in myelin metabolism, clearance and recovery; and female OPCs may present pronounced alterations to promote myelin repair during MS.

Regarding the RRMS peripheral immune system, which is driven by inflammation, we identified a female immune signature core with potential role in intensifying the inflammatory response. Moreover, the adaptive immune response of males may exhibit exacerbated mitochondrial dynamics. Meanwhile, in PPMS, which is driven by neurodegeneration, we noted pronounced sex differences in CD8+ T cells. Considering the PPI networks, female cells were enriched for homeostatic regulatory genes, whereas male cells exhibited genes associated with cytolytic activity.

We acknowledge that the volume and complexity of results may be substantial, and numerous biological patterns remain to be explored in more detail. To encourage a complete coverage of this transcriptomic landscape, we developed the interactive web tool https://bioinfo.cipf.es/cbl-atlas-ms/ (**Figure 3.21**). This application offers a user-friendly interface that supports the free access and visualization of research outcomes. The user can choose the biological inference approach and, within this analysis, select the disease subtype and cell type of interest to explore the complete results.

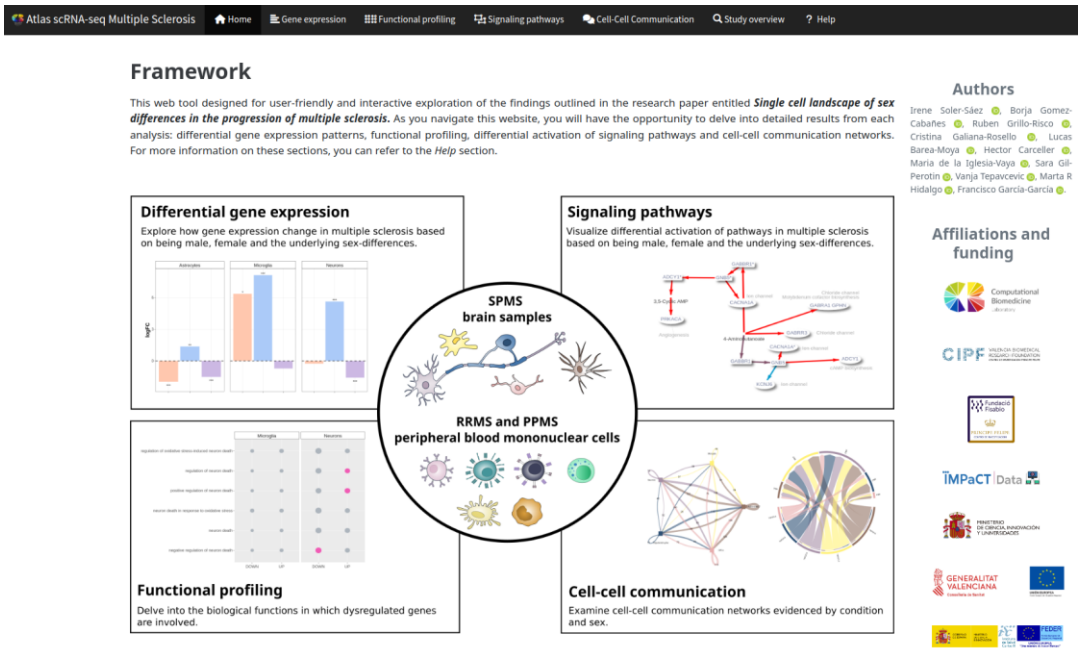


Figure 3.21. Home page of the interactive web tool. The top navigation bar provides access to the different modules of the platform: Home, Gene Expression (differential gene expression results), Functional Profiling (over-representation analysis), Signaling Pathways (differential pathway activation), Cell–Cell Communication, Study Overview (dataset and metadata description), and Help. The central section displays the framework of the tool, while the right panel includes author information, institutional affiliations, and funding acknowledgements. *PPMS: primary-progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis.*

3.5. DISCUSSION

Despite the well-documented sex differences in the clinical outcomes of MS, the specific underlying molecular mechanisms and the impact of determinant factors, such as sex, remain to be fully characterized. This thesis outlines the sex differential single-cell transcriptomic landscape in MS subtypes. With this work, we hope to provide a better comprehension of sex differential molecular mechanisms within each cell type.

We first uncovered CNS samples from SPMS subtype. Our results suggest that females may activate distinct protective responses against neurodegeneration, involving diverse biological processes that are mediated by specific cell types. In particular, we focused on mechanisms related to excitotoxicity, cellular stress responses, and myelin recovery.

The combination of these findings points to sex-biased molecular alterations that could contribute to explain the faster and more severe neurodegenerative progression observed in males.

Excitotoxicity, that is, neuronal damage due to exposure to elevated glutamate levels, represents a major MS hallmark[263]. We observed that female neurons showed increased expression of genes that may act to compensate for such excitotoxicity, such as glutamate and GABA receptors and Ca^{2+} channels to modulate excitability. They also increased the activation of GABA signaling effectors. In contrast, male neurons appeared more susceptible to excitotoxicity due to the increased expression of genes encoding ATPases subunits that provide energy for vesicle biogenesis, while exhibiting a higher activation of glutamatergic signaling effectors.

On another note, the glial-driven MS stress response represents a complex set of processes to cope with different stimuli that account for sex differences; however, studies currently differ in describing their sex differential impact[66]. We reported that male astrocytes expressed increased levels of glutamate reuptake-related genes, which could modulate excitotoxicity in the synaptic cleft[264]. Conversely, female astrocytes presented key genes (such as HIF3A and EGLN3) associated with hypoxic stress, which is characteristic of MS demyelinating lesions[265]. Female astrocytes also appear to balance their profile towards neuroprotection or reactivity in MS by regulating the expression of multifaceted genes such as NEAT1 - a negative regulator of neuronal excitability in neurodegenerative diseases[266] - and LINC-PINT - a potential regulator of oxidative stress in Parkinson's disease substantia nigra[267].

Previous studies have demonstrated more effective remyelination in aged female rodents than males[268], and a hormonal sex-dependent regulation of myelin repair markers in experimental autoimmune encephalomyelitis[220]. Expanding on these findings, our research revealed sex differences in myelin-related genes, which could be crucial for developing effective remyelination therapy strategies. Female oligodendrocytes expressed a greater number of neuronal-like genes than males, which would be relevant in establishing axon-myelin contacts[269]. Female OPCs also increased the expression of QKI, a pivotal gene that promotes oligodendrocyte differentiation[270] and maintains myelin sheaths by regulating lipid homeostasis[271]. To counteract myelin damage, PDGF and FGF represent important signals. PDGF promotes the growth, differentiation, and survival of OPCs[272]. PDGFA (OPCs) - PDGFRA (OPCs) interaction emerged in MS males when compared to male controls, potentially involved in recovery from chronic demyelination[273]. While this interaction becomes established in both control females and MS females, the PDGFC (OPCs) - PDGFRA (OPCs) interaction intensifies in MS females, which relates to reduced neuroinflammation in MS[274]. For FGF signaling,

which promotes myelination, MS females presented an increased proportion of OPC-neuron interactions that could support myelin sheath establishment.

We also found that MS female microglia displayed more genes related to phagocytosis and lipid metabolism compared to males, which could favor myelin clearance[275]. Meanwhile, male microglia presented increased SCARB1 gene expression, which clears Aβ deposition in Alzheimer's disease[276]. Of note, male astrocytes presented increased levels of APOE4, which potentially promotes lipid droplet accumulation, compromising their functionality to offer metabolic neuronal support[277]. Interestingly, APOE4 is associated with an increased tendency of MS progression[278] and cognitive impairment[279].

Previous efforts have been made to understand the role and status of the peripheral immune system in MS. The Multiple Sclerosis Competence Network underscored the importance of its characterization by reporting three different blood endophenotypes linked to distinct disease outcomes[180]. However, evidence related to sex-dependent disturbances remains inconclusive. In our analysis, we reported RRMS sex-differential changes in features that participate in critical processes such as inflammation, cell signaling, and cell-cell adhesion. These alterations may imply mechanisms by which exacerbated immune responses and greater susceptibility are detected in females with RRMS compared to males[280]. Specifically, we identified that genes encoding the AP-1 subunits (FOS, JUN, JUNB, and JUND) formed the central hub of the female immune network, reflecting their importance as master regulators of a broad range of immune system processes such as proliferation, CNS infiltration, and inflammation[281]. This set of genes could serve as a starting point to explore immune targets differentially regulated between sexes. We also identified NFKBIA and NFKBIZ increased in females, genes encoding inhibitors of the MS-associated NF-κB - a pathway that may counteract the proinflammatory status[282]. Interestingly, these findings are likely driven by sexual hormones, as estrogen promotes AP-1 complex activation[283] and NFKBIA gene and protein expression[284].

Our findings also suggest that the adaptive immune system of RRMS males suffers from the enhanced activity of the mitochondrial electron transport chain. Mitochondrial DNA variants have also been proposed as susceptibility biomarkers of RRMS[285]. Regarding PBMCs, RRMS patients exhibited mitochondrial dysfunction compared to controls, which becomes more pronounced in patients experiencing a more aggressive disease course[177]. These findings led us to hypothesize that exploring the mitochondrial status of CNS-infiltrating peripheral T lymphocytes may provide insight into the more rapid neurodegeneration observed in MS males.

The role of the peripheral immune system in PPMS remains less well documented. PPMS differs from other subtypes: treatments effective in RRMS have reduced efficacy in PPMS[40], and the expression of clinical MS biomarkers remains lower in PPMS compared to SPMS[286]. These findings suggest the necessity to comprehensively characterize the PPMS immune profile. In contrast to RRMS, where alterations in females predominate, we found that PPMS males and females have a similar proportion of enriched functions involving broad immunophenotypic features. Interestingly, a higher number of functional and gene expression differences according to sex was detected in CD8+ T cells compared to other cell types. Intrathecal CD8+ T cell subtypes represent critical elements in the immunopathogenesis of PPMS and are associated with white matter injury and thalamic atrophy[183]. They hold particular interest in PPMS due to the significant decline in peripheral CD8+ T cell number observed with age in MS patients compared to healthy controls, suggesting an accelerated aging effect in the disease[287]. We reported sex-biased CD8+ T cell responses, with PPMS females more likely to restore cellular homeostasis than males and PPMS males exhibiting a more activated state characterized by heightened cytolytic responses. These results may represent an additional gene set to explore why MS males trigger more rapid disease progression.

We also identified a genetic signature composed of 67 genes that, based on their sex differential profile, classified the MS immune system primarily by disease subtype and secondly by their immune class. This signature may enable the characterization of biomarkers according to MS subtypes. 12 genes that undergo significantly increased expression in female RRMS but remain not significant in PPMS mainly relate to modulating inflammation, a key differentiating factor between RRMS (higher inflammation) and PPMS (lower inflammation). Of them, CD69 protein has been detected in MS lesions in both brain and infiltrated immune cell types[288,289]. Therefore, increased CD69 expression in RRMS females holds promise for further investigation. Meanwhile, PPMS-associated genes were primarily increase in male patients. The glycolytic metabolism stood out, particularly through the expression of the TPI1 (catalyzes the isomerization of glyceraldehyde 3-phosphate and dihydroxy-acetone phosphate) and LDHA (catalyzes the reversible conversion of pyruvate to lactate) genes. MS peripheral immune system cells exhibit metabolic dysfunction, partly due to their activation during immune responses and the mitochondrial damage they undergo[290]. Given our findings, we emphasize the importance of elucidating sex-specific metabolic profiles to accomplish a better understanding of the disease.

Some of the strongest MS genetic associations in immune cells locate to the HLA gene loci[169]. This gene family has been implicated in several autoimmune conditions[291].

Furthermore, distinct sex-specific differences in HLA expression have been documented across various contexts, including peripheral immune responses to lipopolysaccharide[292] and generalized aggressive periodontitis[293]. MS is no exception, as we observed significant variations in HLA gene expression levels and ligand-receptor interaction strengths. Demyelinated lesions are characterized by increased expression of HLA genes[294]. Specifically, HLA-DRB5 plays an intricate role in MS as it may mitigate disease severity while actively presenting antigens derived from myelin promoting autoimmunity[295,296]. Additionally, we found that HLA-E protein interactions with CD8+ T cells became significant in RRMS females and PPMS males, which may play a pivotal role in presenting autoantigens[16]. If such interactions differ between subtypes, ultimately they may contribute to disease progression. Regarding HLA class II genes, which also influence genetic susceptibility to MS[297], RRMS males exhibited heightened interaction strengths, reaching the levels observed in RRMS females. In contrast, ligand-receptor interactions disappeared in PPMS females compared to female controls. This loss of intensity in connections in female PPMS could represent another area of further exploration. Moreover, it may point to a potential mechanism allowing that, although females have a more active immune system in healthy conditions, this is attenuated to have a neurodegenerative centered progression in the PPMS subtype.

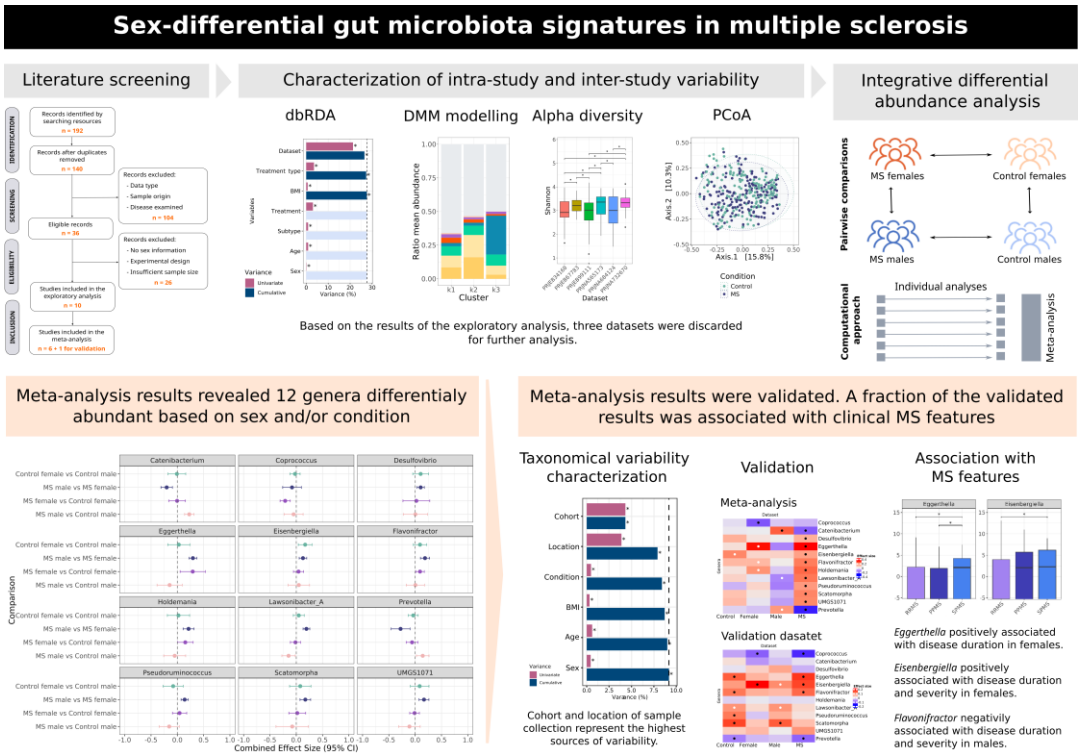
We acknowledge the limitations associated with this study. Although this strategy captured the state-of-the-art scRNA-seq MS data in humans through the systematic review process, the heterogeneous characteristics among the retrieved datasets hindered the analysis. We wish for additional data to analyze novel combinations of tissue and MS subtypes. Additional data will also improve the robustness of our current findings, which are limited by the low number of cells in some populations like microglia, OPCs, and oligodendrocytes. Moreover, it would enable the inclusion of relevant cell types, such as B cells and dendritic cells, that were excluded due to insufficient representation. Despite these limitations, we shed light on cell type-specific responses. The continued dissection of differential expression profiles in additional subpopulations remains of great interest to refine our understanding of sex differential molecular mechanisms.

Although we attempted to cover diverse areas of disease pathology in the CNS and immune system, addressing all encountered differences within a single manuscript remains unfeasible. To provide broader and open access to our findings, we developed a user-friendly website (https://bioinfo.cipf.es/cbl-atlas-ms/), enabling the interactive exploration of the complete results.

To the best of our knowledge, we generated the first single-cell sex-differential transcriptomic atlas of MS, providing results stratified by tissue and disease subtype. Our investigation identified genes, functions, and pathways relevant for understanding

the molecular basis of MS considering the sex of the individual. We hope this work may guide future research and advance knowledge of disease mechanisms, ultimately contributing to the description of potential biomarkers and targeted therapeutic approaches.

4. STUDY II: Integrative analysis to identify sex-differential gut microbiota signatures in multiple sclerosis



4.1. INTRODUCTION

4.1.1. DEFINITIONS AND CONSIDERATIONS OF MICROBIAL-RELATED TERMINOLOGY

Before proceeding with the content of this chapter, in this section we defined a set of terms related to the metagenomic analysis.

Firstly, the term *microbiota* corresponds to the community of microorganisms that inhabit a specific environment, such as the human gut. Meanwhile, the term *microbiome* defines the genetic material of all microorganisms in the given environment, constituting the genes and genomes of the corresponding microbial community. As this doctoral thesis includes metagenomic analyses, the terms *microbiota* and *microbiome* are used interchangeably throughout the text to refer to the entire microbial composition of a given sample or condition.

Additionally, the term *16S* refers to the 16S rRNA gene, a highly conserved genetic marker found in bacteria and archaea microorganisms as introduced in the *General Introduction* chapter. It contains conserved and hypervariable regions, the latter enabling the identification and taxonomic classification of microorganisms. The data analyzed in this doctoral thesis derived from the sequencing of this gene. For brevity, *16S* is used to denote the 16S rRNA gene in this manuscript.

Next, taxonomy is the scientific discipline that involves the characterization, classification, and nomenclature of all living organisms, including the microorganisms that comprise a given microbiota. This classification follows a hierarchical structure, progressing from the most general to the most specific taxonomic ranks: domain, phylum, class, order, family, genus, and species[298]. The term *taxa* (in singular *taxon*) designates units within the hierarchy as a general definition, regardless of their rank. Given the 16S sequencing resolution, the present study evaluated microbial community composition at the genus level. Therefore, the terms *taxa* and *genera* are used as synonyms unless otherwise specified.

As an example of the taxonomical classification, the bacterial species *Bacteroides faecis*, commonly found in human faeces, belongs to the genus *Bacteroides*, within the family *Bacteroidaceae*, order *Bacteroidales*, class *Bacteroidia*, phylum *Bacteroidota*, and domain *Bacteria*. Meanwhile, *Bacteroides intestinalis* belongs to the same genus and higher taxonomic ranks, whilst *Faecalibacterium prausnitzii* only shares the domain *Bacteria* with the two previous species (ranks: genus *Faecalibacterium*, family

Ruminococcaceae, order *Oscillospirales*, class *Clostridia*, phylum *Bacillota*). These examples illustrate how distinct taxa can be close (or far away) within the taxonomic hierarchy. These classifications were retrieved from the Genome Taxonomy Database (GTDB) website (https://gtdb.ecogenomic.org/, last accessed August 12, 2025).

Historically, microbial classification was limited to cultured strains, since only microorganisms that could be grown in laboratory conditions could be phenotypically characterized and taxonomically assigned. However, metagenomic approaches enabled the classification of uncultured microorganisms based on genomic data[298]. This scientific advance boosted the development of curated databases like GTDB[299], SILVA[300] and GreenGenes[301]. Notably, microbial taxonomy is constantly updated by the development of new taxonomic concepts, improved phylogenetic methods, and the incorporation of novel genomic data[298,302]. Following with the previous exemplification, recent revisions restructured the genus *Bacteroides*, with a proportion of taxa reclassified under the genus *Phocaeicola*[303]. Consequently, in microbiome research, it is essential to specify the reference database used and its version, as the taxa classification can be updated over time and different taxonomic designations may refer to the same organisms.

4.1.2. DYNAMICS IN MICROBIAL COMMUNITIES

Microorganisms inhabit complex and dynamic communities rather than co-existing in isolation. They execute diverse biochemical processes that result in the production of metabolites with ecological and physiological relevance. The microbial communities are determined by the interactions within the microbiota and with external factors at a given time and under specific conditions[304]. The external factors that can influence the microbial composition include temperature, pH, oxygen availability, moisture, and nutrient availability, among others. For microbiota inhabiting living hosts, additional influences such as diet and antibiotic use can modulate the community diversity. Moreover, interactions among microorganisms within the same ecological niche may be mediated by the exchange of metabolites, competition for shared resources, and horizontal gene transference[305–308].

Microbial communities can be described from their taxonomic composition and their functional potential, both of which can evolve over time. The taxonomic component refers to the identification and abundance quantification of the taxa present in the microbial community. Meanwhile, the functional potential is established by the repertoire of metabolic pathways and biochemical reactions carried out by the community, regardless of the specific taxa that execute the functions. Although the

taxonomic composition and its functional capacity are related, they are not equivalent, as distinct microbial taxa can perform similar functions[309,310].

4.1.3. General overview of the human gut microbiota

In this doctoral thesis, we investigate the gut microbial taxonomic composition through the analysis of human fecal samples. The human gut microbiota represents the microbiota that inhabits the distal part of the gastrointestinal tract, being one of the most complex, dense and diverse microbial communities within the human body. Specifically, fecal samples provide a representative approximation of the colon composition[311]. The colonic microbiota is composed of around 10^{13} to 10^{14} microorganisms, approximately 99% of which are bacteria[312].

The systematic characterization of the human gut microbiota started with large-scale sequencing initiatives during the first decade of the 2000s. In the United States, the Human Microbiome Project (HMP) aimed to characterize microbial communities across multiple human body sites and different environments. It provided one of the first references for studying the diversity and functional potential of the human gut microbiota as a complete entity[313]. In Europe, the Metagenomics of the Human Intestinal Tract (MetaHIT) consortium also focused on the characterization of the human gut microbiome, with the aim of generating a catalog of microbial genes present in human feces. Notably, the majority of these genes were previously unknown[314].

As a result of these efforts, together with other studies[315], significant progress was made in the description of the human gut microbiota composition. It is composed of a relatively small core set of bacteria present in the majority of individuals in different proportions, with the phyla *Bacteroidetes* (now restructured to *Bacteroidota*) and *Firmicutes* (now restructured to *Bacillota*) being the most abundant. Despite the presence of this core, approximately different 200 genera reside simultaneously in each individual, presenting high inter-individual variability both in terms of genera diversity and their abundance[309,313,314]. Moreover, the functional profiling of the MetaHIT gene catalog identified functions common to all bacteria, including central carbon and RNA metabolism, and amino acid biosynthesis. Functions that adapt the microbiota to the gut environment were also reported, such as adhesion proteins that interact with gastrointestinal host cells and proteins that metabolize compounds from the human diet[314].

Currently, there is a more detailed understanding of the composition and metabolic potential of the human gut microbiota[316,317]. Among the microbial metabolic activities,

the production of short-chain fatty acids (SCFAs) stand out. SCFAs are microbial fermentation products of dietary fibers and represent important signaling metabolites in host–microbe interactions. They are absorbed primarily by colonocytes, and subsequently enter the systemic circulation. Once in the host, SCFAs contribute to multiple physiological processes, including energy production, lipid and glucose metabolism, regulation of immune and inflammatory responses, maintenance of gut barrier integrity, and modulation of gut–brain communication[318].

Regarding the taxonomic composition, the phyla *Firmicutes* and *Bacteroidota* can constitute up to 90% of the bacterial population in a human fecal sample. *Firmicutes* include numerous genera such as *Faecalibacterium*, *Roseburia*, *Ruminococcus*, and *Clostridium*, many of which are butyrate producers. Butyrate is one of the most important SCFA to maintain the gut barrier integrity[317,319]. Meanwhile *Bacteroidota*, represented mainly by the genus *Bacteroides*, are important taxa for metabolizing complex polysaccharides from the diet. As a result, they also produce SCFAs like propionate and acetate[320].

Other phyla, although less abundant, contribute to the functional diversity of the human gut microbiota. For instance, the phylum *Actinobacteria,* which is represented primarily by the genus *Bifidobacterium*, plays an important role in carbohydrate metabolism and immune modulation[321,322]. Meanwhile, the phylum *Proteobacteria,* which includes genera such as *Escherichia* and *Klebsiella*, is generally present at low abundance in healthy individuals, but their abundance can increase in conditions of disease[323].

4.1.4. HOST MODULATORS IN THE HUMAN GUT MICROBIOTA

The human gut microbiota exhibits high variability, both within an individual over time and across different individuals. Interestingly, Lianmin Chen *et al.* 2021[324] characterized longitudinal intra-individual changes in the gut microbiota during four years. They found that almost 60% of taxa and more than 40% of metabolic pathways displayed significant within-individual variation over time. However, the variability was lower within individuals than between individuals, indicating that a person's gut microbial composition remains more similar to their own profile than to the microbiota from other individuals. This finding is consistent with other studies reporting high intra- and inter-individual variability[200,325–327].

Such pronounced variability is originated from an intricate contribution of different factors, including lifestyle, habits, diet, geographic location, environmental exposures, host genetics, immune status, and interactions between microbial taxa themselves, among others (**Figure 4.1**)[328]. These determinants are not independent from one to

another, generating a dynamic multifaceted ecosystem. We now describe the most relevant factors to addressing the biological questions explored in this doctoral thesis.

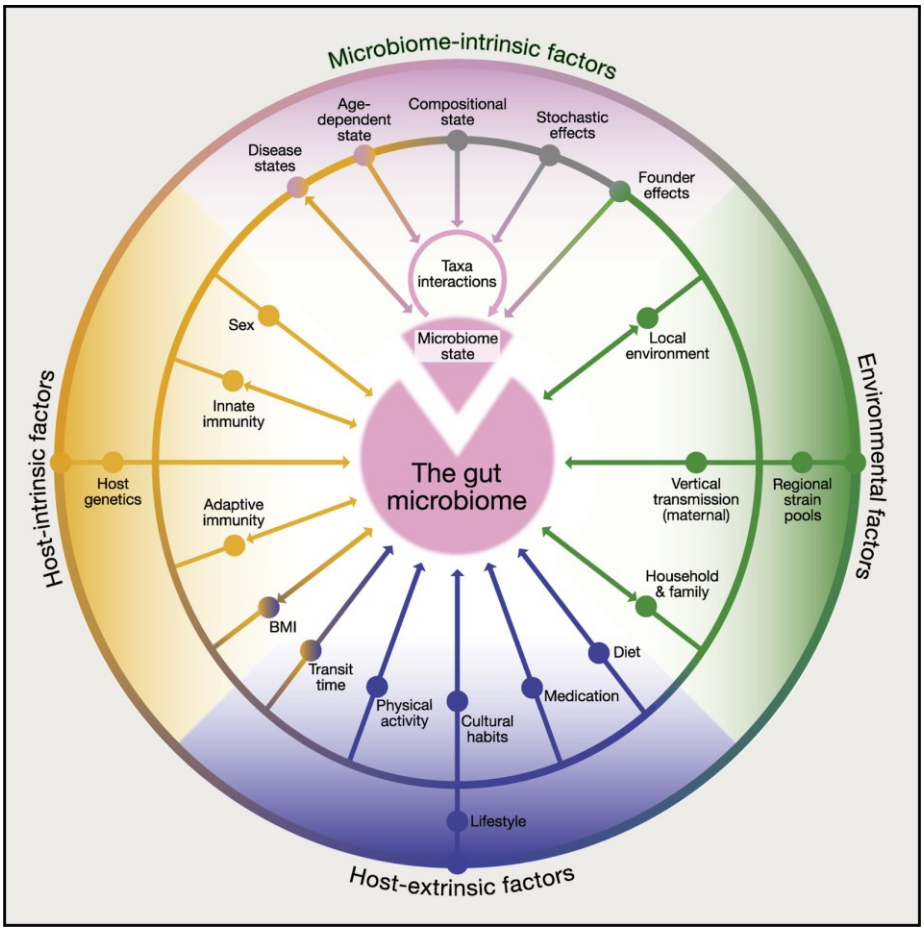


Figure 4.1. Representation of factors influencing the composition and function of the human gut microbiota. The wheel illustrates the multifactorial nature of microbiota regulation, integrating host-intrinsic, host-extrinsic, microbiome-intrinsic and environmental factors. *BMI: body mass index.* Figure from Thomas S.B. Schmidt et al. 2018[328].

4.1.5. Impact of diet and transit time

One of the most influential factors shaping the human gut microbiota is diet, which can account for approximately 5–20% of the variability in microbial composition[329]. Therefore, diet partly determines the microbial community structure and the functional

capacity of the gut microbiota, including its ability to produce bioactive metabolites[330]. The work from Ascinar *et al.* 2021[327] characterized associations between specific microbial taxa and distinct nutrients, food groups, and overall dietary patterns. These associations were also linked to host metabolic parameters related to health status, such as glucose and triglyceride homeostasis.

The main components of the human diet include fibers, fats, carbohydrates and proteins[331]. These dietary constituents reach the intestine with varying degrees of absorption from the host, and are further processed according to the metabolic capacity of the resident microbiota. The non-digested fraction reaches the colon, mainly composed of complex carbohydrates and dietary fibers. These compounds serve as primary substrates for microbial fermentation resulting in the production of SCFAs[332]. Meanwhile, the proteins that reach the colon are metabolized by the microbiota through its proteolytic activity. As a result, bioactive compounds such as ammonia and polyamines are produced, whose detrimental or beneficial effects on the host are determined by their concentration and the physiological context[333].

The microbial variability is also influenced by the gastrointestinal transit time. This term refers to the duration required for ingested material to pass through the digestive tract, from ingestion to excretion. Thus, it is reflected in the stool consistency (**Figure 4.2**). During this process the ecological dynamics within the gut are determined by the dominant metabolic profile. Under fast transit time, the saccharolytic fermentation predominates, favoring the expansion of fast-growth taxa, which is reflected in loose stool consistency. In contrast, slower transit leads to depletion of carbohydrates. Consequently, the microbiota suffers a metabolic shift towards proteolysis, being a slower and more resource-demanding process. The shift alters the competitive relationships among taxa, promoting greater community diversity. This scenario is associated with harder stools, as prolonged transit allows for increased water reabsorption[334,335].

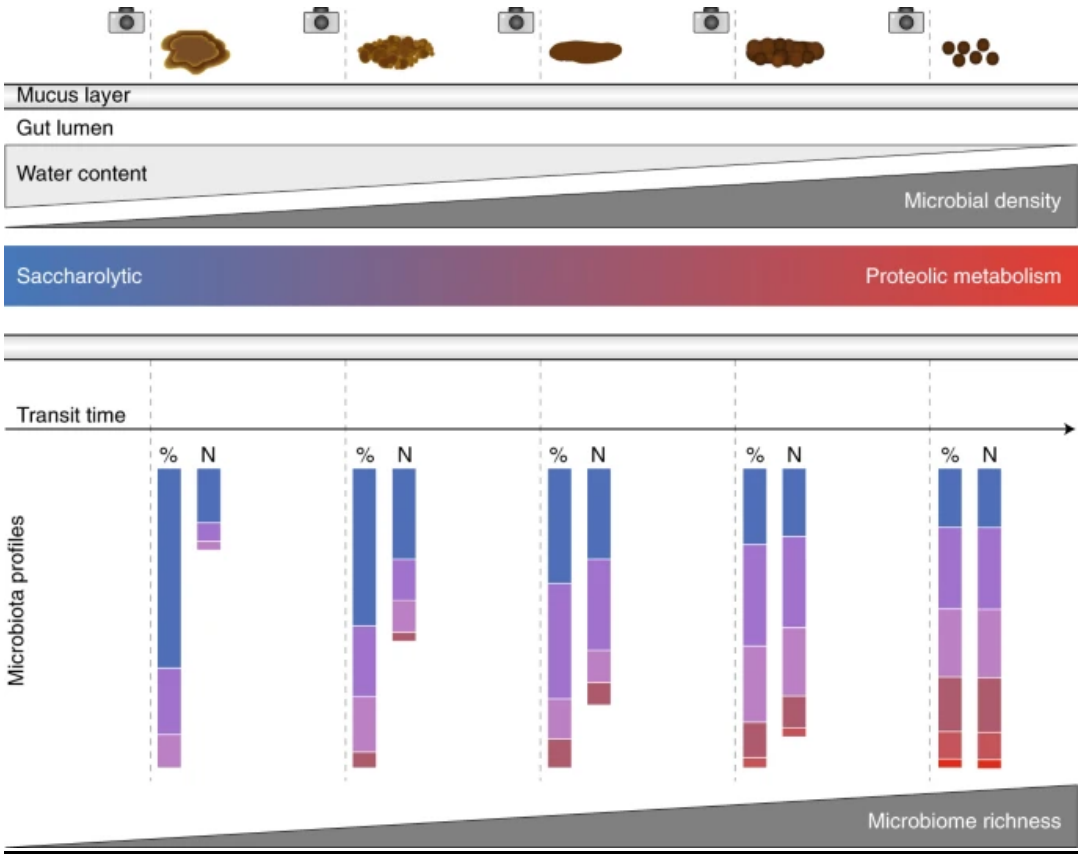


Figure 4.2. Schematic representation of the relationship between gastrointestinal transit time, stool consistency, and microbial metabolism. The upper panel shows stools along the Bristol Stool Scale[336], ranging from loose (fast transit) to hard (slow transit). The color gradient indicates the predominant microbial metabolism: bluer tones reflect carbohydrate-based (saccharolytic) activity, whereas redder tones indicate proteolytic activity. The lower bar plots show changes in microbial richness (number of taxa, N) and the relative abundance (%) of taxa across the transit time gradient. Figure from Gwen *et al.* 2018[334].

4.1.6. Definition of enterotypes

In the previous section, it was described how different successional stages can occur within the gut microbiota. These ecological dynamics promote competition for available nutrients, enabling the definition of taxonomic ratios that reflect functional differences in the microbiota. One of the most widely studied is the *Firmicutes*/*Bacteroidetes* ratio. While these phyla are not mutually exclusive, usually one can dominate over the other. *Firmicutes* are typically associated with the production of butyrate and other SCFAs by the degradation of complex polysaccharides, whereas *Bacteroidetes* tend to generate

SCFAs via the processing of different proteins and fibers. Variations in this ratio have been repeatedly associated with host metabolic traits, such as obesity[337].

When considering the overall microbiota composition rather than focusing on a single phylum, we can define the concept of enterotypes. This term was first proposed in 2011. Enterotypes describe the classification of human gut microbiota in discrete clusters dominated by particular taxa. The original description identified three main clusters: the *Bacteroides*-dominated enterotype, often linked to diets high in animal protein and saturated fats; the *Prevotella*-dominated enterotype, associated with diets rich in carbohydrates and fibers; and the *Ruminococcus*-dominated enterotype, marked by mucin-degrading activity and high microbial diversity[338].

More recent metagenomic analyses have refined the enterotype concept, proposing a four-cluster model in which the *Bacteroides*-dominated enterotype is partly subdivided into two distinct groups: *Bacteroides* 1 (Bact1) and *Bacteroides* 2 (Bact2). This classification is supported by several studies in independent cohorts[339–344]. Bact1 is characterized by complex carbohydrate metabolism and associated with diets high in fiber and intermediate transit time. On the contrary, Bact2 is characterized by lower microbial diversity, higher prevalence of opportunistic pathogens, and functional signatures linked to pro-inflammatory states. In both enterotypes, a high proportion of the genera *Bacteroides* and *Phocaeicola* is identified. Then, the *Ruminococcus* (Rum) enterotype is associated with slow transit time, high diversity, and the greater abundance of slow-growing taxa such as *Akkermansia*, *Alistipes*, *Methanobrevibacter*, and various *Firmicutes* genera. Lastly, the *Prevotella* enterotype (Prev) is the most distinct, with high proportions of the genus *Prevotella* and associations with fast transit time and diets high in fiber and carbohydrates[342].

However, the enterotype concept continues to be debated, with some authors proposing to study the gut microbiota as a continuum rather than through discrete categories[345]. Nonetheless, the enterotype-based characterization of the human gut microbiota remains as a valuable tool in microbiome research, enabling the identification of consistent ecological structures and functional trends within gut microbial communities[342].

4.1.7. GUT MICROBIOTA STATES

We generally identify two states of the microbiota in relation to health and disease: eubiosis and dysbiosis (**Figure 4.3**). The term *eubiosis* describes a balanced microbiota characterized by high diversity, functional redundancy, and stable interactions that support host physiological homeostasis. In a eubiotic state, the microbiota participates

in the metabolism of nutrients, reinforces the intestinal barrier, modulates immune activity, and prevents the colonization of pathogens[346,347]. In contrast, *dysbiosis* denotes a disrupted microbial configuration in which these beneficial functions are compromised. It is often characterized by impaired barrier integrity, the presence of commensals and opportunistic pathogens, and functional alterations that promote host inflammatory responses[347,348].

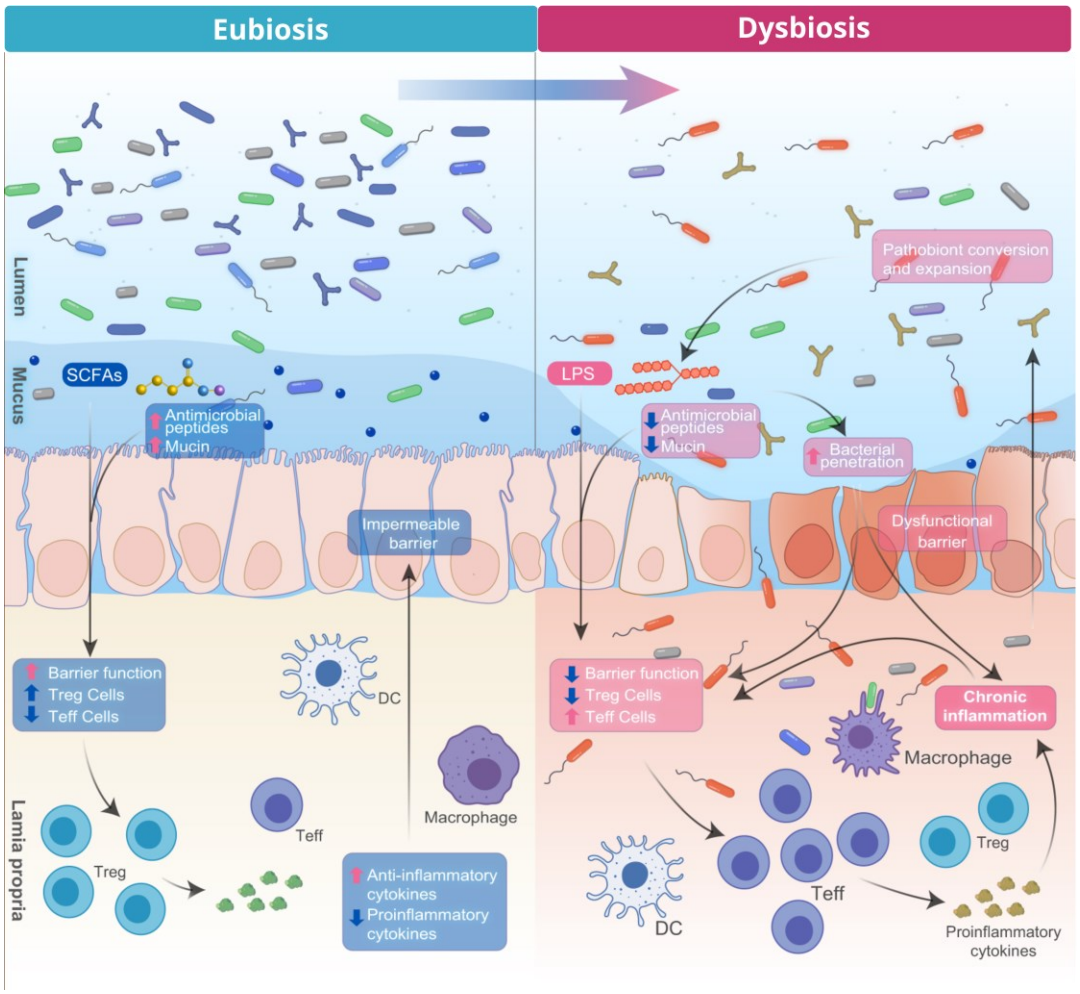


Figure 4.3. Overview of structural and functional differences between the gut microbiota states of eubiosis (left) and dysbiosis (right). The illustration summarizes changes occurring at the intestinal lumen, the mucosal barrier, and host tissue. Upward pink arrows indicate an increase, while downward blue arrows indicate a decrease of the corresponding compound, cell or function. *DC: dendritic cells; LP: lipopolysaccharide (as an inflammation mediator); SCFAs: short-chain fatty acids; Teff: effector T lymphocytes; Treg: regulatory T lymphocytes*. Figure adapted from Kaijian Hou *et al.* 2022[349].

There is still no consensus on the definition of a *healthy* gut microbiota[350]. Initial descriptions focused on taxonomic composition, pointing that the presence of specific beneficial bacteria could be taken as a biomarker of good health. However, the scenario is more complex, as microbial communities can differ between individuals while still supporting the host well-being[351,352]. More recent perspectives define a healthy microbiota by its ecological and functional properties. These include high richness, metabolic versatility, and resilience to perturbations, taking into account host-related variables such as age and diet[349,353]. Moreover, a healthy gut microbiota should be able to ferment complex carbohydrates into SCFAs, maintain epithelial barrier integrity, regulate immune tolerance, and prevent pathogen colonization as previously defined in the concept of eubiosis. However, it is important to note that apparently healthy individuals can present microbiota profiles with properties linked to dysbiosis. This reflects how microbiota health is context-dependent, influenced by inflammation levels, transient changes, and host genetics among other factors[346].

Although defining a healthy gut microbiota remains challenging, certain microbial compositions are commonly linked to disease states. One example is Bact2 enterotype, which has been associated with different diseases. Specifically, it has been linked with inflammatory bowel disease, metabolic syndrome, type 2 diabetes, colorectal cancer, and reduced quality of life. It has also been associated with elevated systemic and chronic inflammatory markers such as C-reactive protein and interleukin-6[340,341,354].

Additionally, previous studies have reported consistent associations between specific gut microbial taxa and human diseases. Inflammatory bowel diseases are one of the most widely studied. In these diseases, there were consistently identified a reduction in *Faecalibacterium prausnitzii* abundance compared to healthy individuals. This butyrate-producing bacterium presents anti-inflammatory properties, whose depletion is linked to increased mucosal inflammation and a worsening of the symptomatology[355–357]. Conversely, opportunistic taxa such as *Ruminococcus gnavus* were identified as enriched in these diseases and associated with active inflammatory states[358,359].

Together, the gut microbial signatures may serve as potential biomarkers for diagnosis and disease monitoring. They also reflect the interplay between gut microbiota composition, host immunity, and the potential disruption of homeostasis.

4.1.8. SEX INFLUENCES THE GUT MICROBIOTA

Gut microbiota features associated with health or disease are further modulated by individual host characteristics. Within this context, sex is an important host-related factor influencing both the microbial composition and its functional potential. Differences between males and females have been reported across the human lifespan[360]. These differences may arise from a combination of hormonal, immunological, behavioral, and lifestyle-related factors that interact with the gut microbiota[361,362].

Notably, the term *microgenderome* was introduced by Flak *et al*. 2013[363], to describe the sexual dimorphism of human microbiomes pointing to the bidirectional interactions between microbiota, sex hormones, and the immune system. Building on this concept, it has been proposed to redefine the term as *microsexosome*, to integrate both the biological sex and gender-related factors into the study of host-microbiome interactions[364]. Its relevance was confirmed by the reanalysis of existing large scale datasets stratifying samples by sex. A representative example were the reanalyses of the HMP data, which revealed significantly differences between male and female gut microbiota[365].

In this section, we will focus on three scenarios through which sex-related differences can influence the human gut microbiota: gastrointestinal tract microbial composition, immune system interactions, and the gut–brain axis.

4.1.8.1. Gastrointestinal tract microbial composition.

Sex-based differences in the gastrointestinal tract have been reported at different levels. The research elaborated by Sender *et al.* 2016[366] suggested that the total number of bacteria in the colon differs between males and females. This difference is reflected in the ratio of microbial to human cells, estimated to be 2.3:1 in females and 1.3:1 in males. In addition, gastrointestinal transit time may be shorter in males than females[335], although in females also varies according to the phase of the menstrual cycle[367].

Sex differential abundance patterns have been reported in different investigations[368]. Some studies described a higher abundance of the phylum *Bacteroides* in males compared to females[369]. Other studies point to differences at the species level, where there could be taxa specific to one sex (not present in the other), and taxa enriched in one of the two sexes[370]. Specifically, the genus *Prevotella* may be more prevalent in males, although steroid metabolism in females appears to promote its growth[371]. However, other studies did not find significant differences in the overall composition once potential confounders are taking into account[372].

Importantly, there is evidence that microbial taxa contribute to the metabolism of sex steroid hormones. The gut microbiota can modulate circulating levels of estrogens and testosterone by altering their reabsorption rates[373,374]. Furthermore, some gut bacteria appear to express receptors for sex steroids. Estrogen receptor β has been identified as directly modulating bacterial metabolism in response to hormone signaling[373,374]. Beyond hormonal influences, evidence from animal model research suggests that sex chromosomes can also modulate gut microbiota composition[375].

4.1.8.2. Immune system interactions

The microbiota interacts directly with the immune system in the gastrointestinal tract. Within the gut-associated lymphoid tissue, the epithelial cells and immune cells maintain a continuous crosstalk with the gut microbiota. As a result, the immune system plays an important role in maintaining homeostasis with the resident microbial communities while the microbiota shapes the immune regulation[376]. Disruption of these interactions can enhance inflammatory signaling and increase susceptibility to chronic inflammatory and autoimmune diseases that affect at systemic level[377].

The immune system exhibits sexual dimorphism in the host, with females generally disposing stronger innate and adaptive immune responses than males. These differences arise from both hormonal and chromosomic influences[224]. Importantly, a bidirectional feedback loop between sex hormones and gut microbes modulates the expansion of specific microbial lineages, which in turn can influence autoimmunity by promoting either inflammatory or tolerogenic immune responses[378, 379].

Specifically, different studies reported sex-dependent associations between gut microbiota composition and autoimmune diseases. In the non-obese diabetic mouse model of type 1 diabetes, the transference of microbial content from males to females increased circulating testosterone, reduced pancreatic inflammation and decreased autoantibody production. Thus, the microbiota transference conferred protection against the development of this autoimmune disease. Similarly, in a murine model of another autoimmune disease called primary biliary cholangitis, researchers identified that more severe lymphocytic infiltration in females compared to males was associated with their microbiota composition[380].

4.1.8.3. Gut-brain axis

Similar to the immune system, the gut microbiota maintains a bidirectional communication with the CNS, commonly referred to as the microbiota–gut–brain axis. Despite the anatomical separation between the gut and the brain, multiple mechanisms

of interaction have been identified: signaling through the vague and enteric nervous systems, regulation of the neuroendocrine axis, and influences on the circulatory system via the release of neuroactive compounds, metabolites, and hormones[381]. A well-known example is serotonin, as around 90% of which is synthesized in the gastrointestinal tract[382].

Different studies have reported sex differences in the gut–brain axis, although the specific molecular mechanisms remain incompletely understood. One communication route involves enteroendocrine cells in the gastrointestinal epithelium, which express receptors for sex steroid hormones[383]. In addition, sex-specific structural brain associations have been observed. Specifically, MRI-based studies have identified links between gut microbiota composition and the morphology of cortical regions in males but not in females[384]. The gut microbiota can also influence neuroinflammation through metabolites such as SCFAs and tryptophan derivatives, which can cross the blood–brain barrier and modulate the activity of the microglia and the astrocytes. Some of these effects showed sex-dependent patterns[385].

The role of sex in the microbiota-gut-brain axis has been investigated in various neurological conditions. In autism spectrum disorder, studies in animal models reported sex differences in gut microbiota composition, microbial metabolite profiles, and the corresponding influence in the immune responses[386]. In children with this condition, Wang *et al.* 2019[387] identified sex-dependent differences in microbiota-mediated immune regulation. Sex differences were also reported in schizophrenia, which may be associated with alterations in glutamatergic neurotransmission[386].

Neurodegenerative diseases are no exception to the influence of the microbiota in the brain considering the sex of the individual[374]. Piyali Saha *et al.* 2024[388] recently reviewed that sex-dependent gut microbiome perturbations that may influence Alzheimer's disease pathophysiology. In Parkinson's disease, both experimental models[389] and human studies[390] identified sex-related differences in microbiota composition that may contribute to the disease via the gut-brain axis.

4.1.9. INHERENT CHARACTERISTICS OF 16S METAGENOMIC DATA

Microbiome sequencing datasets present specific properties that must be considered when analyzing and interpreting the results[391].

Firstly, this data type is compositional. The biological material obtained from a sample represents a fraction of the total microbial community present in the host, given the vast number of microorganisms that inhabit it. This fraction is further subsampled during the processing steps before sequencing, meaning that the number of final reads is delimited by saturation effects. If the microbial load or cell density is not measured, sequencing outputs cannot be quantified, providing relative abundance estimates[340]. This is of particular importance because the proportion of sequenced cells relative to the total in the sample can vary due to both technical factors and biological variability, such as differences in stool consistency. As a result, metagenomic data provides information about the composition rather than the absolute number of microorganisms present in the corresponding sample[392].

In addition to compositionality, metagenomics datasets are characterized by high sparsity, with a large proportion of taxa showing zero counts across many samples. Given the compositional nature of these data, it is not possible to distinguish whether zeros in the count matrix reflect the true absence of a taxon —due to biological factors such as host-specific conditions or diet— or whether they result from undersampling, in which case the taxon is present but undetected due to insufficient sequencing coverage[393].

Another statistical feature of microbiome sequencing data is overdispersion, where the variance in taxon counts is higher than the corresponding mean values. This property is defined partly from the uneven structure of microbial communities, where a small number of taxa dominate in abundance, while the remaining majority are present in low proportion. Moreover, these abundance distributions vary considerably between individuals, reflecting both biological heterogeneity and stochastic sampling effects[391].

Variability in sequencing library sizes represents an additional technical aspect to consider. These sizes can differ by orders of magnitude both within and across studies, introducing potential biases in the estimation of taxonomic abundances. Since sequencing captures only a fraction of the total microbial community, larger libraries sizes present more probability of detecting low-abundance taxa, whereas smaller libraries sizes are more susceptible to not detect low abundant taxa[394].

Given these characteristics, different computational normalization approaches have been developed to consider the aspects of metagenomic data structure[394,395]. Moreover,

statistical analyses frequently rely on tests that do not assume normality, including non-parametric methods, permutation-based approaches, and models that incorporate explicit dispersion parameters to account for the high variability. Additionally, preprocessing steps often involve the removal of low-prevalence taxa to reduce noise and improve the robustness of downstream analyses[391,396–398].

4.1.10. BIOINFORMATIC STRATEGIES FOR ANALYZING 16S DATA

An illustration of the 16S metagenomic strategies is outlined in **Figure 4.4.** As a result of the sequencing process, the generated FASTQC files contain the raw reads for each sample. The previous experimental steps are defined in the *General Introduction* chapter. After sequencing, a quality control step is performed to filter and trim low quality reads. Next, the different versions of 16S rRNA gene sequences are identified, quantified, and taxonomically classified based on reference databases. Once the microbial taxa present in each sample are identified and quantified, downstream analyses can be conducted to explore the objectives of the study[121,399].

Different metrics can be computed to characterize the variability of the microbial communities. Specifically, alpha diversity metrics quantified within-sample diversity, whereas beta diversity evaluates compositional differences between samples. To explore relationships between microbiota composition and explanatory variables (e.g., host-related factors), we can apply redundancy distance based analyses. Additionally, we can identify discrete community types (**Figure 4.4, pink, orange, grey, and green boxes**).

To further investigate compositional changes, differential abundance analyses can be conducted to identify taxa whose abundances varied significantly between the experimental groups of interest. Moreover, association analyses can be performed to explore the relationships between different taxa and host or environmental metadata. Additional approaches not conducted in this work, such as functional inference analyses, can be performed to predict the metabolic potential of the microbiota (**Figure 4.4, blue, purple and yellow boxes**).

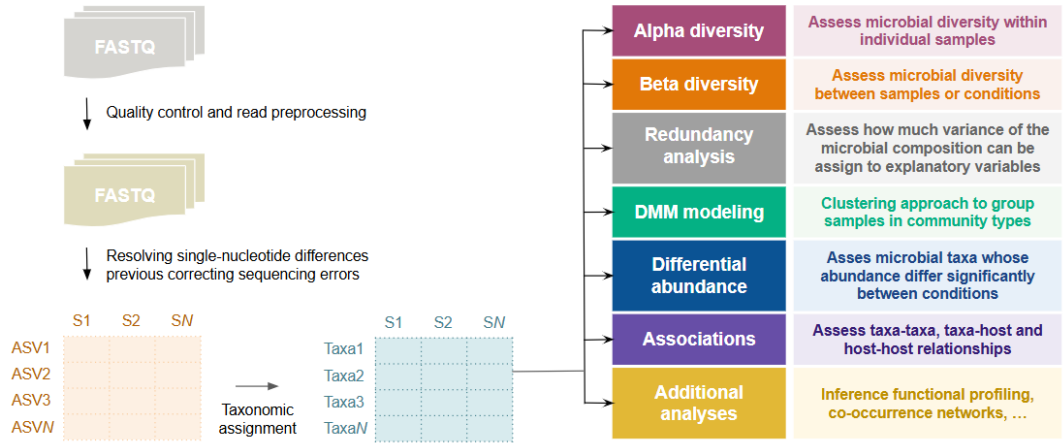


Figure 4.4. Overview of the bioinformatic analyses that can be used for 16S metagenomic analysis. Raw FASTQ sequencing files are processed to remove low quality reads and sequencing artefacts. Error-corrected distinct sequences are identified as amplicon sequence variants, which are quantified and then assigned to taxonomic categories. The phylogenetic matrix is evaluated in downstream analyses, including alpha diversity, beta-diversity, redundancy analysis, Dirichlet multinomial mixture modeling, differential abundance and association analyses to explore the microbial composition and its relationships with environmental and host-related factors. A*SV: amplicon sequence variant; DMM: Dirichlet multinomial mixture.*

4.2. CONTEXTUALIZATION, MOTIVATION AND OBJECTIVES

The gut microbiota has emerged as an important factor in the study of MS. Studies in animal models for MS determined that the presence of the gut microbiota is necessary for the development of the disease. Both germ-free mice and mice treated with antibiotics to deplete the complete gut microbiota did not develop MS-like symptomatology. Notably, when the gut microbiota was restored, the susceptibility to surfer MS was increased[400,401]. Moreover, fecal microbiota transplantation from MS patients into germ-free mice promoted the development of MS and exacerbated disease progression[402,403]. This microbial-dependent modulation of MS pathogenesis has been attributed primarily to the regulation of T cell autoimmune responses within the gut-associated immune system, whose alterations extend to the systemic level. Specifically,

the presence of certain microbial communities influences the conversion from proinflammatory to anti-inflammatory T cell profiles and vice versa[404].

In human studies, distinct patterns of gut microbial dysbiosis have been identified among MS patients[404,405]. Several investigations reported alterations in the abundance of specific bacterial taxa compared with healthy controls, although the exact taxa differ between studies[406–409]. However, a recurring observation is the depletion of taxa with anti-inflammatory potential and the enrichment of taxa with pro-inflammatory capacity[410,411]. Moreover, individuals with milder forms of MS tend to present a gut microbiota composition more similar to that of healthy controls than to those with severe MS[412]. Certain taxa also appear to be associated with increased disease severity during progression, potentially reflecting a more stressful microbial environment and a reduced production of SCFAs[413]. Interestingly, a recent meta-analysis integrating both preclinical and clinical studies of probiotic interventions in MS revealed that probiotic supplementation can reduce disease severity, delay disease progression, and improve motor impairments, highlighting the pivotal role of the gut microbiota in MS pathophysiology[414].

Beyond the gut and the immune system, these microbial alterations may also have an influence in the CNS. Precisely, gut dysbiosis in MS was linked to increased permeability of both the gastrointestinal and blood-brain barriers, facilitating the translocation of bacterial metabolites into the systemic circulation and ultimately into the CNS. Once in the CNS, these metabolites may modulate the activity of glial cells and influence the differentiation of OPCs, thereby impacting the synthesis and maintenance of the myelin sheath[415]. Experimental studies in model organisms further support this link, reporting that butyrate, one of the SCFAs produced by the microbiota, can promote remyelination[416].

The individual efforts of the scientific community have been compiled in systematic reviews, trying to summarize different literature findings. The work from Ali Mirza *et al.* 2020[417] reported that gut microbiota diversity did not differ between MS patients and healthy controls in the majority of studies. Additionally, no consistent differences were observed when stratifying by MS clinical characteristics. Some studies identified differential abundances of specific taxa, although these were rarely reproducible across multiple cohorts. The systematic reviews from Alba Ordoñez-Rodriguez *et al.* 2023[418] and Sophia Jette *et al.* 2024[419] reached comparable conclusions. The former also highlighted that, although taxon-level findings remain inconsistent and often contradictory, the majority point to alterations in SCFA metabolism, which may play a modulatory role in MS-associated inflammation[418].

To the best of our knowledge, at the time of writing this doctoral thesis two meta-analyses were identified. They integrate microbiome data to perform new analyses rather than summarizing the results of the individual manuscripts. Qingqi Lin, *et al.* 2024[420], found no significant differences in alpha and beta diversity metrics between MS patients and healthy controls. Interestingly, only 20% of taxa were identified consistently across all analyzed studies, yet these accounted for 86.2% of the total microbial abundance. Among these, a reduced proportion of *Prevotella* was consistently observed in MS patients, reaching statistical significance in more than half of the analyzed cohorts. Similarly, Xiaoyun Zhang *et al.* 2024[421] conducted a meta-analysis of multiple alpha-diversity indices and also reported no significant differences between MS patients and controls. Their microbial composition analyses likewise failed to identify consistent taxonomic changes across studies.

To our understanding, the interaction between sex and MS remains poorly characterized. Marianna D'Anca *et al.* 2023[422] summarized current evidence on sex-microbiota and MS-microbiota interactions, highlighting the need for studies that explicitly evaluate both factors together. Evidence from model organisms supports this link. Patrick G. Miller *et al.* 2015[423] generated transgenic mice with autoimmune responses against myelin sheath proteins. In this model, they observed a sex-dependent immune response. Moreover, the analysis of microbial profiles revealed protective bacteria enriched in males, whereas females exhibited higher abundance of bacteria with pathogenic potential. In another mouse model of MS, Gil Benedek *et al.* 2017[424], identified that estrogen treatment modified the gut microbiota, enriching its composition and promoting immune regulation while the severity of the disease was reduced. Similarly, female mice expressed higher levels of specific receptors to interact with the microbiota in intestinal epithelial cells, which promoted autoimmunity and exacerbated MS-like symptoms[425,426]. In humans, quantification of SCFA levels in MS patients and controls revealed no overall differences. However, female patients presented significantly lower SCFA concentrations compared to male patients, suggesting a sex-specific metabolic imbalance[427].

Overall, the gut microbiota contributes to multiple layers of human physiology, including the regulation of both the immune system and the CNS, playing an important role in autoimmune and neurological diseases. In MS, the gut microbiota has been shown to influence disease onset and progression, as summarized in the previous paragraphs. However, current findings remain inconsistent across studies, partly due to differences in methodological approaches and population heterogeneity. Importantly, although sex is a well-established factor influencing MS risk and clinical course, the interaction between sex and microbiota composition in MS remains poorly defined. Thus, the main

objective of this chapter is to characterize the contribution of the gut microbiota to sex differences in MS. To achieve this, we developed an integrative *in silico* strategy to analyze publicly available 16S rRNA sequencing datasets. The analysis is structured around the following specific objectives:

1. To identify 16S datasets that explore MS based on predefined inclusion and exclusion criteria, retrieving eligible datasets from the corresponding repositories.
2. To process each selected dataset individually through standard bioinformatic workflows to obtain processed data suitable for downstream analyses.
3. To analyze inter- and intra-study variability in gut microbiota datasets, in order to evaluate their suitability for inclusion in the subsequent integration approach.
4. To perform pairwise comparisons considering the condition (MS or control) and sex (female or male) within each selected study, obtaining individual differential abundance profiles.
5. To integrate the individual study results into a meta-analysis, identifying consistent microbial signatures associated with sex differences in MS.
6. To validate computationally the robustness of the meta-analysis results with an independent cohort.
7. To associate the significant taxa with MS clinical characteristics, aiming to uncover sex-related microbial markers linked to disease course and severity.
8. To develop an open-access, user-friendly web resource for dissemination of complete results to the scientific community.

4.3. MATERIALS AND METHODS

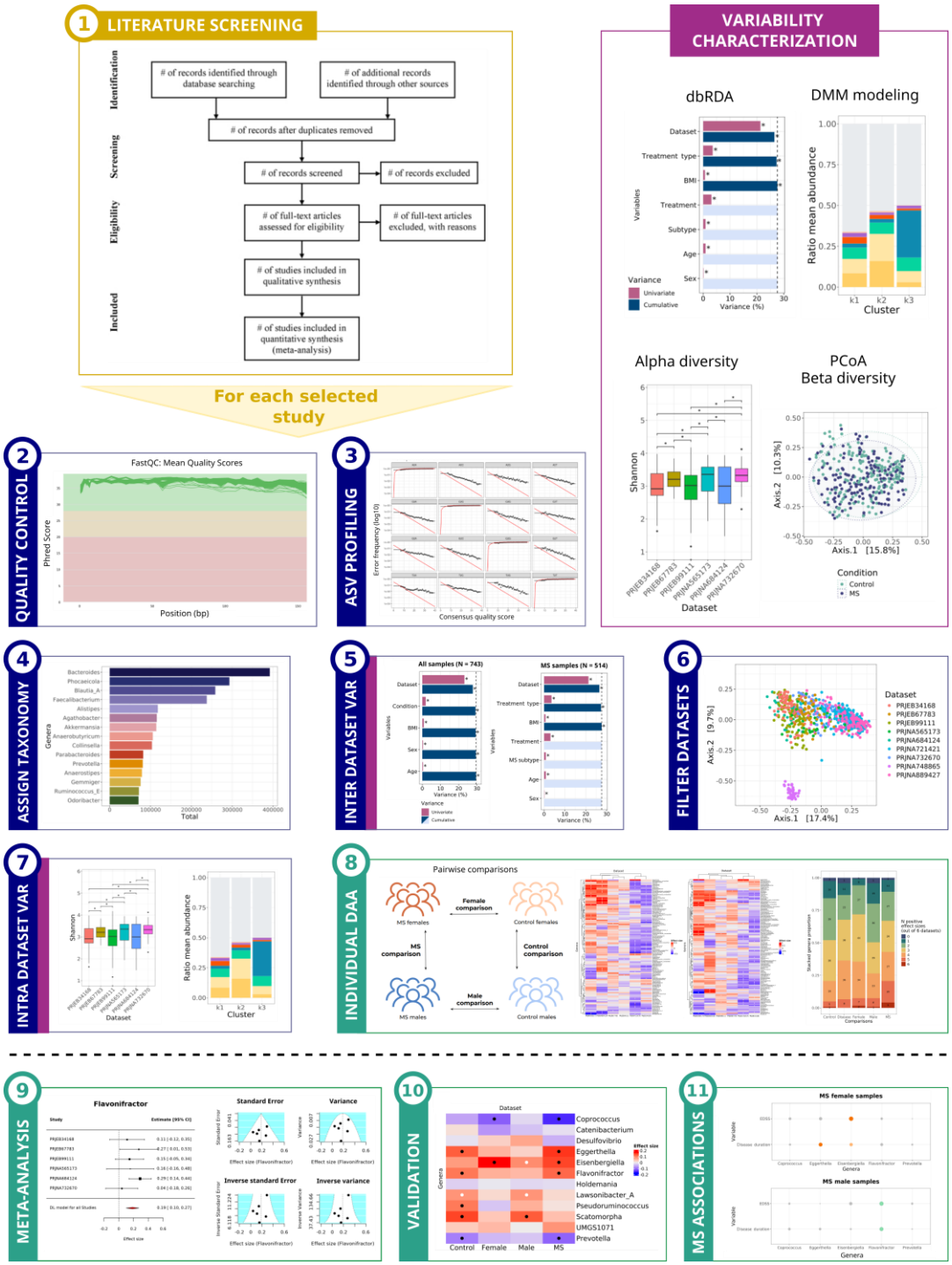
4.3.1. WORKFLOW DESCRIPTION

In this work, microbiome datasets composed of sequencing reads were identified through a systematic review (**Figure 4.5, yellow box**). The upstream experimental procedures to obtain the sequencing reads were not conducted; for further details consult the *General Introduction* chapter.

The computational workflow for each selected study was performed following different steps (**Figure 4.5, blue boxes**): i) quality control, ii) identification and quantification of exact biological sequences, and iii) their taxonomical annotation. As a result, we obtained the count matrices. Then, we characterized the intra-dataset variability to evaluate potential sources of heterogeneity, which allowed us to exclude datasets exhibiting excessive variability. For the remaining datasets, we further assessed the intra-study variability and conducted individual differential abundance analyses, focusing on four pairwise comparisons considering the condition (MS or control) and sex (female or male) of the individuals (**Figure 4.5, green box above the dashed line**). Results by comparison were integrated through a meta-analysis approximation. Finally, the significant results were validated in an independent cohort, and further associated with MS clinical features (**Figure 4.5, green boxes below the dashed line**).

The bioinformatics code developed in this chapter was generated using a combination of bash executions and R programming (version 4.3.2). **Supplementary Table 4.S1** provides a record of the versions of corresponding programs and R packages. Moreover, the bioinformatic code is available at https://github.com/IrSoler/cbl-metaanalysis-16S-MS.

Figure 4.5. Workflow of this research. (*Next page*) Steps followed in the metagenomic analysis: literature screening to identify datasets, computational processing of microbial 16S reads, characterization of taxonomical variability, and biological inference approaches. The dashed line separates the computational analyses on individual datasets (top) from integration and validation steps (bottom). *ASV: amplicon sequencing variant; BMI: body mass index; bp: base pairs; DAA: differential abundance analysis; dbRDA: distance-based redundancy analysis; DMM: Dirichlet multinomial model; MS: multiple sclerosis; PCoA: principal coordinates analysis; QC: quality control; VAR: variability.*


1 LITERATURE SCREENING
Identification
Screening
Eligibility
Included
# of records identified through database searching
# of additional records identified through other sources
# of records after duplicates removed
# of records screened
# of records excluded
# of full-text articles assessed for eligibility
# of full-text articles excluded, with reasons
# of studies included in qualitative synthesis
# of studies included in quantitative synthesis (meta-analysis)
For each selected study
VARIABILITY CHARACTERIZATION
dbRDA
DMM modeling
Alpha diversity
PCoA Beta diversity
2 QUALITY CONTROL
FastQC: Mean Quality Scores
3 ASV PROFILING
4 ASSIGN TAXONOMY
5 INTER DATASET VAR
6 FILTER DATASETS
7 INTRA DATASET VAR
8 INDIVIDUAL DAA
Pairwise comparisons
9 META-ANALYSIS
Flavonifractor
10 VALIDATION
11 MS ASSOCIATIONS

4.3.2. SYSTEMATIC REVIEW

To identify human gut metagenomic data, we followed PRISMA guidelines[135]. Literature screening was conducted up to January 2025 in the ENA and SRA public repositories. Additionally, we searched for peer-reviewed publications with associated data in PubMed. In these three resources, we applied the following inclusion criteria:

- For the ENA database, we used the keywords *multiple sclerosis microbiome* in the *Enter text search terms* field. We explored the resulting entries at the project level.
- For the SRA database, we constructed the following query using the advanced options:

  ((multiple sclerosis) AND "human gut metagenome"[orgn:__txid408170]) AND bioproject_sra[filter]

  This query retrieved MS studies involving the human gut metagenome, specified with the NCBI taxonomic identifier 408170.
- In PubMed browser, we employed the following query:

  ("multiple sclerosis"[All Fields] AND ((micro[All Fields]) AND (human[All Fields] OR "homo sapiens"[All Fields]) AND (metagenom*[All Fields] OR metatranscript*[All Fields] OR 16S[All Fields])))*

  We used wildcards (*) to include related terms with different endings. Similarly, we attempted to identify the different sequencing approaches with "metagenom* OR metatranscript* OR 16S" keywords. We also restricted the results to studies involving humans, with the terms "human" or "*Homo sapiens*". The [All fields] tag enabled the identification of the keywords in all searchable fields, including title, abstracts, keywords, etc.

After this initial search, all the identified studies were individually reviewed. Studies were excluded if they met any of the following criteria:

- Methodology: studies that did not generate metagenomic or metratranscriptomic omics data.
- Sample origin: studies not based on human fecal samples.
- Disease examined: studies not based on MS.

- Experimental design: studies that included only control individuals or only MS patients, but not both within the same dataset.
- Sex information not reported.
- Sample size: data from at least three different individuals per condition and sex (control females, MS females, control males, MS males).
- Unavailability of the FASTQ and/or associated metadata files.

After applying inclusion and exclusion criteria, we identified nine studies based on 16S rRNA sequencing and one based on WGS metagenomics. Given the broader availability of the 16S studies, we focused on the integration of this type of data, discarding the WGS record. The corresponding FASTQ files and associated metadata from the selected studies were downloaded and stored within the computational infrastructure of the CIPF. To facilitate their reference, in this doctoral thesis each study is labelled with its accession identifier.

4.3.3. STANDARDIZATION OF METADATA NOMENCLATURE

Due to the lack of standardization when reporting sample-associated variables, a prior step to the bioinformatic analysis was the assessment of the available metadata. This process included, for each selected dataset, the identification of which variables were present and which were not reported.

When working with different datasets, it is also essential to ensure that variables are directly comparable across studies. For numerical variables, it is necessary the verification that they are reported using the same units or scale (e.g., disease duration can be reported in years or in months), and the homogenization of the nomenclature of categorical variables.

Specifically, in this study we aimed to collect at least the following variables:

- Technical metadata: sequencing technology, targeted 16S rRNA hypervariable region(s), sample collection date, batch information for sample collection and/or sequencing, geographic origin of the study (country), fecal collection process, and DNA extraction kit.
- Biological metadata: clinical condition (healthy individuals or MS patients), MS subtype, body mass index (BMI), age and MS-specific treatments.

When the variable was not included in a given study, we recorded it as missing value.

The information regarding MS-specific treatments stood out for being highly heterogeneous across datasets, not only considering how and which treatments were reported, but also due to the use of brand names or active compounds. Thus, a two-step approach was developed. Firstly, treatment names were homogenized to the corresponding active compounds; and secondly treatments were categorized according to their mechanism of action based on current classifications[31]. The assignments can be consulted in **Supplementary Table 4.S2,** where we defined the following categories:

- Never treated: individuals who have never received any treatment for MS.
- Present untreated: individuals who had not received treatment at least three months prior fecal sample collection.
- Classic immunomodulator: treatments that modulate the balance between pro-inflammatory and anti-inflammatory cytokines (e.g., interferon beta).
- Lymphocyte retention: treatments that retain lymphocytes in secondary lymphoid organs, reducing their circulation in the peripheral blood (e.g., fingolimod).
- CNS adhesion inhibitors: treatments that block the migration of lymphocytes into the CNS without significantly altering their levels in the peripheral blood (e.g., natalizumab).
- B-cell depleting (anti-CD20): treatments that specifically reduce the population of B cells in the peripheral blood (e.g., ocrelizumab).
- Lymphocyte depleting: treatments that induce the depletion of both T and B cells (e.g., cladribine).

This variable is the one evaluated when referring to the term *Treatment type* in this chapter, unless otherwise indicated. Similarly, when we use the term *Treatment*, we refer to the experimental groups: control, untreated MS, and treated MS, regardless of the treatment that MS patients received.

4.3.4. PROCESSING OF RAW SEQUENCING READS

The first step of the analysis comprised the characterization of the quality and composition of the downloaded sequencing reads.

Fecal samples may contain residual host-derived DNA in addition to the DNA from the microbiota. Although 16S rRNA gene amplification is not expected to target human DNA, we ensured that no human contamination was present. We verified the absence of host derived reads using the *Kneaddata* pipeline developed by the Huttenhower Lab (https://github.com/biobakery/kneaddata, last accessed June 16, 2025), which performs an alignment against a reference human genome using *bowtie2* tool[428].

We next performed the quality control step of the microbial reads. Raw sequencing reads may contain technical artifacts, low quality bases and adapter sequences that must be removed. Quality metrics for each sample were calculated with the FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last access June 16, 2025). Then, results were summarized in a single report per database with MultiQC[429]. The quality metrics evaluated were: sequencing depth, per base and per sequence quality scores, per base and per sequence GC content, per base N content, and sequence length.

We found the sequencing reads were partially preprocessed. We reviewed the associated original publications and the corresponding supplementary materials, identifying that none of the studies explicitly reported de quality filtering and trimming parameters. Therefore, based on our quality control reports, we applied filtering and trimming cutoffs to remove low-quality bases and reads, adapter sequences (if present), and undetermined nucleotides. Detailed thresholds are defined in **Table 4.1**, which were established considering the specific characteristics of each dataset while maintaining the pipeline as standardized as possible. For filtering and trimming the reads we used the Cutadapt algorithm[430]. After processing, we recalculated the FASTQC metrics to confirm that all samples met the established quality standards.

Table 4.1. Overview of the preprocessing parameters by dataset. For each dataset, the following information is provided: dataset identifier, number of samples, sequencing mode (paired-end or single-end), targeted 16S rRNA region (V), sequencing technology, original read length (in base pairs), trimming cutoffs applied at the 5′ and 3′ ends, minimum read length (in base pairs) retained after quality trimming, and the minimum fold of parent reads over total abundance to remove quimeras.

Dataset	N	Type	Region	Technology
PRJNA684124	30	Paired-end	V3-V4	Illumina MiSeq
PRJNA721421	456	Paired-end	V4	Illumina MiSeq
PRJEB99111	143	Single-end	V4	Illumina MiSeq
PRJNA748865	38	Single-end	V2:V4-V6:V9	Ion Torrent
PRJNA889427	143	Paired-end	V4	Illumina MiSeq
PRJNA732670	56	Paired-end	V3-V4	Illumina MiSeq
PRJEB34168	65	Paired-end	V4	Illumina MiSeq
PRJNA565173	30	Paired-end	V3-V4	Illumina MiSeq
PRJEB67783	47	Paired-end	V3-V4	Illumina MiSeq
PRJEB32762	2927	Single-end	V4	Illumina MiSeq

Dataset	Original length	Trimming 5’	Trimming 3’	Min length	Min quimera
PRJNA684124	300	10, 10	10, 50	125	8
PRJNA721421	Variable	10, 10	10, 10	125	8
PRJEB99111	150	10	0	75	8
PRJNA748865	Variable	trunQ = 25		50	8
PRJNA889427	250	10, 10	0, 0	125	8
PRJNA732670	300	10, 10	10, 30	125	8
PRJEB34168	300	20, 20	20, 40	125	8
PRJNA565173	280	10, 10	20, 50	125	8
PRJEB67783	300	20, 20	20, 50	125	8
PRJEB32762	150	10	0	75	8

4.3.5. QUANTIFICATION OF AMPLICON SEQUENCE VARIANTS

The next step in the individual analysis was the inference of amplicon sequence variants (ASVs). ASVs represent the unique DNA sequencing reads that are present in the biological sample resolved at single-nucleotide level. Their inference enables the identification of true biological sequences, which are expected to appear repeatedly as consistent PCR amplifications, in contrast to sequencing errors that occur randomly[431].

We inferred ASVs implementing DADA2 algorithm[432] based on the tutorial version 1.8 (https://benjjneb.github.io/dada2/tutorial_1_8.html, last access June 18, 2025). The analysis was composed of the following steps: dereplication, error rate learning, ASV inference, merging paired-end sequencing reads (when necessary), chimera removal, statistic evaluations, construction of ASV abundance matrix, and filtering non-target-length sequences.

Using the filtered and trimmed sequences as input, the first step was dereplication, which consists of collapsing identical sequencing reads into unique sequences counting their abundance within each sample. Next, we generated a model of sequencing error rates. The model estimated the probability that each nucleotide was misidentified as another (e.g., adenine to guanosine). The estimated probability is based on the Phred score resulting from the sequencing process. As a result, the DADA2 pipeline infers the subset that most likely represent true biological sequences, distinguishing them from the sequences containing errors that may have been created during the sequencing process or PCR amplification steps.

For the datasets that contained paired-end sequencing reads (defined in **Table 4.1**), forward and reverse reads were merged to generate the complete amplicon sequence. DADA2 default parameter established a minimum overlap of 20 base pairs between reads, which was used for all datasets in our study.

At this point we inferred the error rates at specific positions of the reads. However, we also identified chimeric sequences, which are technical artifacts introduced during the PCR amplification. Chimeras are commonly generated when the DNA polymerase stops before finishing the synthesis, producing an incomplete fragment. In the next amplification cycles, this fragment may act as a *primer*, hybridizing with a different although similar read that acts as a DNA template. The result is a hybrid sequence from two real ones, which can be confused with a novel variant not present in the biological sample. To identify chimeras, we defined a minimum abundance of biological sequences relative to the potential quimera in the *minFoldParentOverAbundance* parameter. We set the value to 8 for all datasets, meaning that the real sequences had to be at least 8

times more abundant than the candidate chimera to be removed. Finally, we verified that the length of the final ASVs fell in the expected range considering the sequenced 16S hypervariable region(s).

Throughout this pipeline reads were discarded if they were not identified as ASVs, if forward and reverse reads could not be successfully merged (in paired-end datasets), if they were identified as chimeras or if the read length fell outside the expected range for the amplified region. After completing the process, we evaluated the proportion of reads that remained at each step, confirming that none of the samples in any dataset needed to be excluded for downstream analysis.

As a result of all the processing steps described above, we obtained ASV abundance count matrices, with rows representing samples, columns corresponding to the individual ASVs, and each row-column position indicating the abundance of the given ASV in the corresponding sample.

4.3.6. PHYLOGENETIC ANNOTATION

Next, we assigned the taxonomic identities to the ASVs. We used the taxonomic annotations from the GTDB[299], a resource developed by the University of Queensland. GTDB is a manually curated and periodically updated database, constructed using the sequences available in the RefSeq[433] and GenBank[434] repositories. For the taxonomic assignment, GTDB concatenates 120 bacterial and 122 marker genes, respectively. These gene markers are derived from single-copy proteins conserved across microbial genomes, minimizing the difficulties occasioned by gene duplications in the phylogenetic tree construction. Moreover, they are also ubiquitous, highly conserved with low mutation rates. Thus, the concatenated final sequences are informative for resolving phylogenetic relationships, especially polyphyletic taxa annotation[435,436].

Among the selected marker genes is the 16S rRNA gene. As the GTDB annotations are not integrated in DADA2 pipeline, Dr. Claus Christophersen's team at Edith Cowan University periodically processes the GTDB database to extract 16S-based annotations, enabling its direct use with DADA2. In this work, we used the version corresponding to the GTDB release 220, which can be found in the Zenodo repository https://zenodo.org/records/13984843 (last access June 25, 2025).

The taxonomic assignment was performed with the *assignTaxonomy* function from the DADA2 R package[432]. Its algorithm divides the ASV into k-mers and estimates the probability, with a naive Bayes classifier, that a given sequence belongs to a particular taxon based on its k-mer profile. To improve the annotation, the method also implements

a bootstrap approach to assess the confidence of each assignment, only labelling the reliable ones. As a result, each ASV was taxonomically classified at different levels: *Kingdom*, *Phylum*, *Class*, *Order*, *Family*, and *Genus*. For ASVs lacking confident classification at a given level, we assigned them as *unclassified* from the last known taxonomic rank. For instance, if a sequence could be confidently assigned to the *Lachnospiraceae* family but not to any genus, we annotated its genus level as *Lachnospiraceae_unclassified*.

4.3.7. FILTERING CRITERIA AT SAMPLE LEVEL

Once the phylogenetic count matrices were generated for each individual study, a quality assessment of the samples was performed to exclude those unsuitable for analysis. This included the samples with poor or incomplete preprocessing, as well as those potentially contaminated.

As part of this evaluation, samples that remained of low quality in the filtering and trimming steps were discarded. Moreover, two quality metrics were calculated for each sample: the total number of counts and the number of distinct genera detected. These values were visualized to explore their distribution across samples. Samples with fewer than 1,000 total counts or fewer than 30 identified genera were excluded.

Rarefaction curves were also computed to determine the taxonomic richness relative to sequencing depth. Rarefaction is a statistical procedure that estimates the number of taxa expected at a given subsampling depth. The resulting values are plotted in refraction curves, with the number of identified taxa (Y-axis) against the number of sequencing depth (X-axis). The maximum reached by the curve is the real sequencing depth. The curve usually exhibits a steep initial slope, representing the rapid identification of frequently prevalent taxa. Then, the curve levels off as fewer additional taxa are identified reaching a *plateau* value. A sample whose rarefaction curve ends in the exponential phase may indicate incomplete detection of taxonomic diversity, suggesting insufficient sequencing depth. In contrast, curves that reach a plateau suggest that most taxa present in the sample have been captured, with only rare low-abundance taxa potentially undetected[437,438].

In addition, the relative abundances of the top 15 most abundant genera were visualized for each sample. We discarded samples containing genera present at much higher proportions than expected, or containing genera not typically present in human gut microbiota.

In summary, within each individual study, samples were removed if they met any of the following exclusion criteria: i) total read counts below 1,000; ii) fewer than 30 distinct taxa identified; iii) rarefaction curves that did not reach a clear *plateau*, suggesting insufficient sequencing depth; or iv) indications of contamination or poor processing based on the most abundant taxonomic composition. The remaining samples constituted the filtered abundance matrices, which served as the input for all subsequent analyses.

4.3.8. ALPHA DIVERSITY METRIC

Alpha diversity refers to the microbial diversity within a single sample[439]. Different metrics exist to quantify this diversity. In this study, we used the Shannon diversity index, as it simultaneously captures both richness (the number of different taxa present in the sample) and evenness (how homogeneously distributed are taxa in terms of their relative abundance). The Shannon index (H) is defined as[440]:

$$H = -\sum_{i=1}^{S} p_i \, ln\,(p_i)$$

Where $p_i = n_i/N$ represents the proportion of taxa belonging to the genus i (n_i) relative to the total number of taxa in the sample (N), ln denotes the natural logarithm, and S the total number of observed genera.

Higher values of Shannon index indicate greater diversity, considering both presenting many distinct taxa (high richness) with abundances more evenly distributed. In contrast, lower values suggest the sample is composed of few highly abundant taxa. We obtained the Shannon Index at genus level with the *estimate_richness* function from the phyloseq R package[441].

4.3.9. BETA DIVERSITY METRIC

Beta diversity metrics evaluate differences in the community composition among groups of interest[442]. Precisely, we quantified differences between microbial communities with the Bray-Curtis dissimilarity index. This metric compares two samples based on both the presence/absence of taxa and the abundance of shared taxa. Specifically, it is defined as[443]:

$$BC_{AB} = \frac{\sum_{i=1}^{S}|A_i - B_i|}{\sum_{i=1}^{S}|A_i + B_i|}$$

where A_i and B_i are the abundance of genus i in samples A and B, respectively. S is the total number of taxa observed across both samples. The Bray-Curtis index ranges from 0 to 1, where 0 represents complete similarity (identical community composition and abundances of the corresponding taxa), and 1 indicates complete dissimilarity (no shared taxa).

Beta-diversity dissimilarity distances between pairs of samples were explored with the dimensionality reduction approach *Principal Coordinates Analysis* (PCoA)[444]. The PCoA strategy presents the same objective as the described PCA method for transcriptomic data analysis: to identify linear combinations of the high dimensional data into a lower dimensional space, preserving the original distances between samples as accurately as possible. However, while PCA is based on correlation or variance matrices, PCoA is conducted with distance metrics such as Bray-Curtis dissimilarity values. Therefore, we decomposed the distance matrix into principal coordinates, which represent new orthogonal axes summarizing the variability of the samples. This procedure is also known as ordination analysis, independently on the selected distance metric. When visualizing samples distributed in the reduced (i.e., ordination) space, those samples located closer to each other present more similar microbial compositions than those located farther apart. In this study, the PCoA was computed using the function *ordinate* from the phyloseq R package[441], specifying "PCoA" as the method and "bray" as the distance metric.

4.3.10. dbRDA ANALYSIS

Distance-based redundancy analysis (dbRDA) was performed to quantify the proportion of phylogenetic variance that can be attributed to variables of interest[445]. The dbRDA is an extension of linear regression for modelling response variables. This method allows us to explain the variation in taxonomic composition (i.e., genus-level abundance matrix) using explanatory variables stored in the metadata files, such as disease condition, sex, or treatment.

Firstly, PCoA was constructed using the Bray-Curtis dissimilarity matrix, described in the previous section. As a result, we obtained the principal coordinates representing the distance across samples based on their microbial community profile, thus, providing a representation of how compositionally different samples are from one another. Next, we

fitted the model for each explanatory variable independently. The analysis was performed using the *capscale* function from the vegan R package[446]. The statistical significance was evaluated using a permutation-based ANOVA approach with the function *anova.cca* also from vegan R package[446].

From each model we extracted the following metrics:

- F: F-statistic as a measure of the strength of the association.
- R2: proportion of variance explained by the evaluated variable.
- Adjusted R2: proportion of variance explained adjusted by the number of observations and the number of degrees of freedom.
- N: number of samples.
- p-value as a measure of significance.

We applied the BH method[250] to correct for multiple testing across variables, considering significant results when FDR < 0.05. We handled missing values with the argument *na.action* set as *na.exclude*. This approach removes missing values during the model fitting, although it preserves the original structure of the data maintaining the excluded samples in the PCoA generation. Thus, we favored consistency when comparing or combining results from different variables.

More than one variable can account for a significant proportion of the variability in microbial composition. The variance explained by each of these variables may overlap meaning that it is redundant, or may represent different sources of variation[447]. To differentiate among these situations, we performed a stepwise dbRDA approach with the function *ordi2step* from vegan R package[446]. We first selected the metadata variables that represent significant associations resulting from the dbRDA univariate analysis. Then, we defined two models to explain the distance dissimilarity matrix: a null model with any explanatory variables, and a complete model including all significant variables from the dbRDA univariate analysis. Starting from the null model, we incorporated the variables one by one. At each step, the method evaluated whether including an additional variable significantly improved the model, i.e., whether we can explain more variability including the new variable. The process concluded when the model no longer increased the explained variability, resulting in a final model composed of the non-redundant subset of variables.

4.3.11. MICROBIAL COMMUNITY TYPING

As described in the *Introduction* section of this chapter, we can categorize microbial community profiles into groups called enterotypes. One of the most commonly used clustering methods to identify enterotypes is the Dirichlet Multinomial Mixture (DMM) model[448].

This method was developed specifically for metagenomic data, considering its inherent characteristics: the discrete nature of taxa counts, the fact that many taxa are low abundant or absent in most samples, and the high variability of microbial loads and sequencing depths. To account for these features, DMM relies on multinomial sampling rather than fixed distributions, estimating the probability of each taxon appearing in each sample (**Figure 4.6**). Therefore, each microbial community is described by a vector of taxa probabilities, representing both its average composition and the variability across samples. These vectors generated from a set of Dirichlet distributions, each representing a different metacommunity (i.e., a different group). By modeling the data as a mixture of the Dirichlet components, DMM clusterizes samples according to the metacommunity they most likely originate from. Overall, we are modelling the probability of belonging to a metacommunity based on the sampling probability of the sample[448].

The DMM model was applied to different numbers of metacommunities. We tested from 1 to 6, given that between 3 and 4 enterotypes are commonly identified in industrialized populations[449]. Each model was calculated at genus level using the *dmn* function from the R package DirichletMultinomial[450]. Then, we determined the most probable number of metacommunities identified with Laplace approximation[448] that quantifies how well the model explains the observed data. Specifically, Laplace estimated the correct fit of the model to the actual data while penalizing complexity to avoid overfitting. As a result, it calculates a negative log-marginal likelihood and consequently, the lower the value, the better the corresponding number of metacommunities. This model fit was also retrieved from the *dmn* function execution.

The identification of the optimal number of metacommunities does not necessarily imply that they correspond to enterotypes, since it results from an unsupervised clustering approach. To confirm this, we explored the taxa distribution and assessed whether it corresponds to the expected pattern. We also verified that no batch effect was present that could confound this assignment.

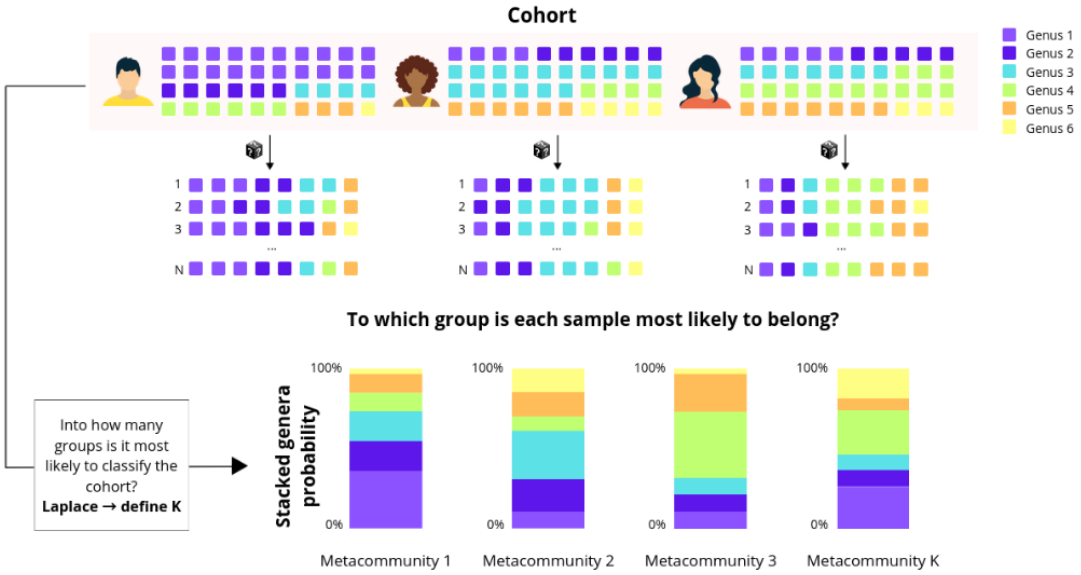


Figure 4.6. Illustration of Dirichlet Multinomial Mixtures algorithm. Top part represents the microbial composition of three individuals, where each square represents a genus and its color indicates its identity. The algorithm assigns samples to the most likely community group based on their taxonomic composition, thereby defining metacommunities. The optimal number of metacommunities is determined using the Laplace approximation, which evaluates the model fit**.**

4.3.12. STATISTICAL TESTS

As mentioned in the *Introduction* section of this chapter, the data generated in microbiome analyses are characterized by a non-normal distribution, upper outliers and being compositional. Given these peculiarities, non-parametric biostatistical strategies were selected.

4.3.12.1. χ^2 Goodness of fit

The χ^2 (Chi-squared) Goodness of Fit test is a statistical method used to evaluate whether the observed frequency distribution of a categorical variable significantly deviates from an expected distribution[451]. The levels of the categorical variable can be distributed in two or more groups, in which the test evaluates the following hypotheses:

- **H_0 (null hypothesis):** the observed frequencies are consistent with the expected frequencies under the uniform distribution, where all categories are expected to

belong to each group with equal probability. Any deviations are attributed to random variation.

- **H_a (alternative hypothesis):** the observed frequencies deviate significantly from the expected frequencies. The differences are greater than would be expected by chance.

The basis of the test involves constructing a contingency table that records the observed counts for each level of the categorical variable within each group. This test relies on the assumption of independence, as the observations must be independent from one another, and each observation contributes only to the frequency count of one group. The test calculates a χ^2 statistic by comparing the observed and expected counts under the null hypothesis with the formula:

$$\chi^2 = \sum_{i,j} \frac{(observed - expected)^2}{observed}$$

where χ^2 represents the Chi-squared statistic, i represents level of the categorical variable, and j represents the group.

Once the test statistic is obtained, the p-value is computed by comparing the χ^2 statistic to the chi-squared distribution with $(k - 1)$ degrees of freedom, where k is the number of categories. Small p-values suggest that the observed distribution significantly differs from the expected one, leading to the rejection of the null hypothesis. We calculated the test with the *chisq.test* function from the stats R package[452]. To control for multiple testing, p-values were adjusted by the BH method[250], considering significant results when FDR < 0.05.

4.3.12.2. Wilcoxon rank-sum test

The Wilcoxon rank-sum test, also known as the Mann-Whitney U test, determines whether the numerical variable of interest follows the same distribution across two independent groups[453,454]. In brief, samples are ranked from the highest to the lowest value of the corresponding numerical variable. If the samples of the two groups are evenly distributed in the ordered list, there is no significant difference between the groups. Conversely, a significant difference arises if one of the groups presents higher values, with most samples located at the top of the list. The samples from the other group, which show lower values, would be at the bottom of the ordered list. The Wilcoxon rank-sum test does not assume normality. This strategy provides robustness against outliers

and asymmetric distributions, being the non-parametric alternative to the *student* t-test. We calculated the test with the *wilcox.test* function from the stats R package[452]. Adjusted p-values were calculated by the BH method[250], considering significant results when FDR < 0.05.

4.3.12.3. Kruskal-Wallis and post hoc Dunn test

When comparing a numerical variable in more than two independent groups, we applied the Kruskall-Wallis test[455], which is an extension of the Wilcoxon rank-sum test previously defined. It is used for ordinal data or continuous data that would be converted to ranks.

If we obtained significant results with the Kruskall-Wallis test, we conducted pairwise comparisons with the *post hoc* Dunn test[456]. This test performs pairwise rank comparisons between all groups while adjusting for multiple testing, identifying which specific groups present significant differences.

Both tests were executed in R using the *kruskal.test* function from the stats package[452] and the *dunnTest* function from the dunn.test package[457]. When multiple testing was performed, adjusted p-values were calculated by the BH method[250], considering significant results when FDR < 0.05.

4.3.12.4. Spearman correlation

Correlations are statistical strategies used to evaluate the degree and direction of the association between two numerical variables. In this work, we applied Spearman's rank correlation coefficient, a non-parametric method suitable for data that do not meet the assumptions of normality[458]. Similar to the theory behind the Wilcoxon rank-sum test, Spearman's method operates on the ranked values of the data rather than directly using the magnitude of the values. Each variable is transformed by replacing its values with their respective ranks (first, second, third, etc.), preserving the order of the observations.

Spearman's coefficient evaluates the strength and direction of a monotonic relationship between the two variables, whose function consistently increases or decreases but not necessarily at a constant rate. The coefficient is computed by ranking the values of each variable, calculating the differences between the paired ranks, and then applying the formula:

$$\rho = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

where d_i is the difference between the ranks of each observation and n is the number of paired observations.

The resulting coefficient ranges from –1 to +1, where +1 indicates a perfect positive monotonic relationship, –1 indicates a perfect negative monotonic relationship, and 0 indicates no monotonic association.

Then, the hypotheses tested in Spearman's correlation analysis are:

- **H_0 (null hypothesis):** $\rho = 0 \rightarrow$ There is no monotonic association between the two variables.
- **H_a (alternative hypothesis):** $\rho \neq 0 \rightarrow$ There is a statistically significant monotonic relationship (positive or negative) between the variables.

To evaluate the statistical significance of ρ, a p-value is calculated. For large sample sizes such as those analysed in this thesis ($n > 20$), the statistical test can be approximated to a t-student distribution using the formula:

$$t = \frac{\rho \cdot \sqrt{(n-2)}}{\sqrt{(1 - \rho^2)}}$$

This statistic follows a *t*-distribution with $n - 2$ degrees of freedom. A p-value < 0.05 suggests rejecting the null hypothesis, supporting the existence of a monotonic relationship between the two variables. In this doctoral thesis, the correlation analyses were performed using the *cor.test* function from the R stats package[452], specifying the method "spearman". Adjusted p-values were calculated by BH method[250], and significant results were considered when FDR < 0.05.

4.3.13. IDENTIFICATION OF DIFFERENTIAL ABUNDANCE PATTERNS BY DATASET CONSIDERING THE CONDITION AND SEX OF THE INDIVIDUALS

We investigated potential shifts in microbial composition associated with the disease status and the sex of the individuals for each dataset. The computational workflow was structured in three main steps: i) filtering of low-abundance genera, ii) normalization of the raw abundance matrix, and iii) differential abundance analysis.

4.3.13.1. Filtering of low-abundance genera

The raw abundance matrix was filtered at genus level to retain only those taxa present at a relative abundance of at least 0.01% in a minimum of 10% of the samples, reducing noise introduced by rare or spurious taxa. Although differential abundance analyses were conducted independently for each dataset, the same filtering criteria were applied across all studies to avoid dataset-specific biases in taxonomic representation. After filtering, a total of 121 genera were retained for downstream analysis.

4.3.13.2. Normalization

The filtered matrix was normalized using the variance stabilizing transformation (VST) method from the DESeq2 R package[459]. With this strategy, the raw abundance matrix is first normalized according to the corresponding library size factors. Then, a scaled transformation is applied to account for the relationship between the dispersion and the mean of each taxon. This approach enabled the analysis of continuous values with approximately uniform variance, to mitigate the compositionality inherent to microbiome datasets. This strategy was selected based on benchmarking performed by Llorens *et al.* 2021[395], where VST was identified among the best-performing methods. Notably, it approximates the behavior of quantitative microbiome profiling, being recommended in case-control studies when direct experimental measurement of microbial load is not available.

4.3.13.3. Comparisons

Each dataset was then subjected to differential abundance testing. The evaluated comparisons were (**Figure 4.7**):

- **Impact of sex in the control group (Control comparison)**: differences among females and males being healthy individuals by the comparison:

Control females - Control males

- **Impact of sex in the MS group (MS comparison)**: differences among females and males being MS patients by the comparison:

MS females - MS males

- **Impact of disease in females (Female comparison):** differences among MS patients and healthy individuals being female by the comparison:

MS females - Control females

- **Impact of disease in males (Male comparison):** differences among MS patients and healthy individuals being male by the comparison:

MS males - Control males

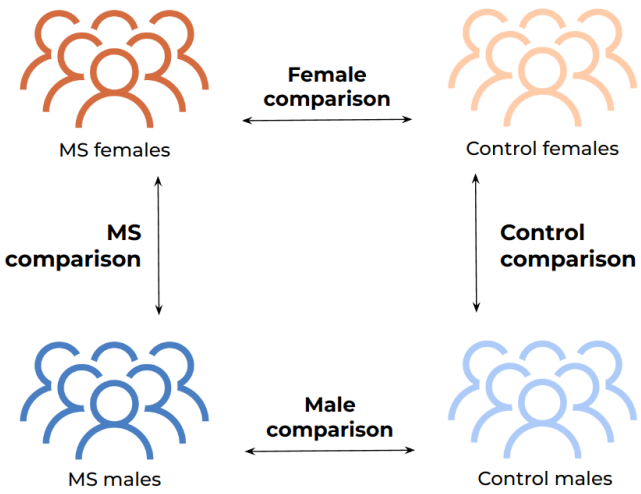


Figure 4.7. Pairwise comparisons for the differential abundance analyses. The four contrasts evaluated are illustrated: i) Control comparison: control females *vs.* control males, ii) MS comparison: MS females *vs.* MS males, iii) Female comparison: MS females *vs.* control females, and iv) Male comparison: MS males *vs.* control males. *MS: multiple sclerosis.*

4.3.13.4. Differential abundance analysis

For each comparison, non-parametric Wilcoxon rank-sum tests were used to determine differences in genera-level abundances between the corresponding groups. Further details on the implementation of this statistical test are provided in the *4.3.12.2* section of this chapter. Resulting p-values were adjusted using the BH method[250], with taxa being considered significantly differentially abundant when FDR < 0.05.

A notable batch effect was identified in the validation dataset (identifier: PRJEB32762), related to both the cohort and the city where the samples were collected. More details can be consulted in the *Results* section. To account for these sources of variability, a combined batch variable was created by merging these two factors. Differential abundance testing was then performed using a blocked Wilcoxon rank-sum test. Specifically, we used the *wilcox_test* function from the R coin package[460]. The formula was determined as $y \sim x \mid block$, where y represents the abundance of the genus being tested, x represents the grouping variable defining the comparison of interest, and *block* corresponds to the combined batch factor. Again, the resulting p-values were adjusted for multiple testing using the BH method[250], considering the taxa significantly differentially abundant when FDR < 0.05.

To provide a quantitative estimate of the magnitude of observed differences, effect sizes were calculated using the *wilcox_effsize* function form rstatix R package[461]. The effect sizes, denoted as r, were computed with the formula:

$$r = \frac{Z}{\sqrt{N}}$$

where Z is the standardized Z-statistic from the Wilcoxon test, and N the sample size. The resulting r values range from –1 to +1, where larger absolute values indicate stronger effect sizes. In addition to magnitude, the sign of the effect size reflects the direction of the change. Negative values ($r < 0$) indicate a higher abundance of the genus in the first group of the comparison (or a lower abundance in the second), whereas positive values ($r > 0$) indicate a higher abundance in the second group of the comparison (or lower in the first).

We also estimated the 95% confidence intervals for the effect sizes. For this purpose, bootstrap resampling of effect sizes was performed with 1,000 iterations, using the *boot* and *boot.ci* functions from the boot R package[462].

4.3.14. META-ANALYSIS

We applied a meta-analysis approach to integrate the results of the individual differential abundance analysis for each previously defined comparison. This strategy accounts for the variability between datasets providing a more robust estimate of whether the observed differences among groups of interest are consistent[141].

The data generated from each study presents inherent variability (i.e., different extraction protocols, cohort characteristics, etc.). To account for these sources of heterogeneity, we applied a random-effects model[463]. This model assumes that the *true* effect size may vary among studies, estimating a combined effect that reflects both within-study error and between-study variability. The effect size was used as the measure of the degree of change, as it captures both the direction and magnitude of the change between conditions of interest. Meanwhile, the standard error of the effect size was used as the measure of its variability. Given the number of iterations for calculating confidence intervals described in the previous section, we assumed a normal distribution, and consequently, we calculated the standard error (SE) with the formula:

$$SE = \frac{high\ confidence\ interval - low\ confidence\ interval}{2 \cdot 1{,}96}$$

We then integrated individual effect sizes with the DerSimonian-Laird (DL) method[464], executed via the metafor R package[465]. We obtained a weighted average of the effect size values for each feature, assigning less weight to studies with higher variability. Generally speaking, the resulting value is referred to as the combined effect size. For each taxon included in the meta-analysis, we calculated the combined effect size and the corresponding p-value. To control for multiple testing, p-values were adjusted by the BH method[250], considering significant results when FDR < 0.05.

In addition to the combined effect size, we retrieved different indicators that allowed the evaluation of the heterogeneity across studies for each taxon[464]:

- **QE and QEp**: represent the test statistic and its associated p-value, respectively, derived from the DL heterogeneity test. This test assesses whether the variability in effect sizes across studies exceeds what would be expected by chance. A significant QEp value suggests the presence of heterogeneity among studies.
- **Standard Error (SE):** reflects the uncertainty of the estimated combined effect size.

- **Tau-squared (τ^2)**: estimates between-study variance, quantifying how much true effect sizes differ across studies beyond chance.
- **I-squared (I^2):** represents the proportion of total variation in effect sizes that is due to between-study heterogeneity rather than chance.
- **H-squared (H^2):** measures the ratio of total observed variability to the expected sampling variability.

Heterogeneity was also evaluated at study level through influence plots and funnel plots. Influence plots help to identify individual studies that present a large impact on the combined effect size[466,467]. They are generated based on metrics calculated by sequentially removing one study and assessing how the DL model parameters are affected. Specifically:

- **rstudent**: quantifies how different is the effect size of each study from the predicted combined effect. Outlier values may indicate a strong impact of the study in the model.
- **dffits**: assesses how much a study influences the DL fitted model. Higher values indicate that the corresponding study has a strong impact on the model.
- **cook.d:** evaluates the overall impact of removing a study on the model parameters. High values suggest high influence on the combined estimate.
- **cov.r:** reflects the change in the covariance matrix of the model when a study is removed. Unusually high or low values may indicate an abnormal influence on the estimated precision.
- **tau2.del:** determines the change in τ^2 when each study is excluded from the model. Larger values suggest that the study contributes substantially to heterogeneity.
- **QE.del**: indicates how the QE statistic is affected by removing each study. Larger values may imply that the study is important for explaining the overall heterogeneity.
- **hat:** calculates how much a study influences the model by their specific characteristics. Higher values point to studies with atypical characteristics.
- **weight:** represents the statistical contribution of each study to the calculation of the combined effect size. Studies with greater weight present more influence on the integrated result.

Meanwhile, funnel plots were elaborated to visually assess the bias of the different effect sizes by displaying the estimated effect size on the X-axis against measures of its variability (SE and variance) on the Y-axis[468]. In the absence of dataset bias, studies were expected to be distributed symmetrically forming an inverted funnel shape. This pattern implies that more precise studies (lower variability) would be located close to the *true* effect, while less precise studies (higher variability) are distributed more broadly. Asymmetric distributions may also indicate different sources of heterogeneity.

4.3.15. WEB TOOL

To enable interactive exploration and visualization of the individual differential abundance and meta-analysis results, we developed the web tool https://irsoler.shinyapps.io/metaanalisis_16S_MS/. This web-based application was built using the Shiny R package[261]. It is hosted on *shinyapps.io*, a cloud platform for deploying Shiny applications. Users can interactively select the datasets and genera of interest, and adjust parameters to explore the outcomes of this work.

4.4. RESULTS

We now present the results obtained throughout the different computational phases of the study. Taken together, these analyses aim to uncover gut microbial variations that may help explain sex differences in MS pathophysiology. This section begins with the description of the 16S datasets identified during the systematic review. We then explore the heterogeneity across studies to determine which datasets are suitable for inclusion in the subsequent analyses. In addition, we characterize intra-study variability to define the most appropriate statistical approach. For each selected dataset, we performed four pairwise comparisons evaluating differential taxonomic abundance with respect to condition and sex: Control comparison, Female comparison, Male comparison, and MS comparison. The results obtained were computationally validated in an independent cohort. Finally, the significant taxa were associated with key clinical characteristics of MS, such as disease subtype and disability severity.

4.4.1. IDENTIFICATION OF SUITABLE DATASETS THROUGH LITERATURE SCREENING

The datasets evaluated in this study were identified through the systematic review described in the corresponding *Materials and Methods* section. The flow diagram summarizing the number of records retained at each stage of the review process is shown in **Figure 4.8**. Precisely, a total of 192 records were retrieved in the identification phase from the ENA and SRA repositories, as well as from peer-reviewed publications indexed in PubMed. After removing duplicates, 140 unique records remained. Following the individual review of each study, 104 records were excluded for one or more of the following reasons: the study did not contain metagenomic or metatranscriptomic data, the samples were not of human fecal origin, or the study did not include MS samples. This screening step resulted in 36 potentially eligible records, of which 26 were further excluded due to the absence of either healthy control or MS samples, the unavailability of sex information, or the lack of access to raw FASTQ files and/or associated metadata. Ultimately, 10 studies were included in the exploratory analysis. Technical and biological characteristics of these datasets are summarized in **Table 4.2**.

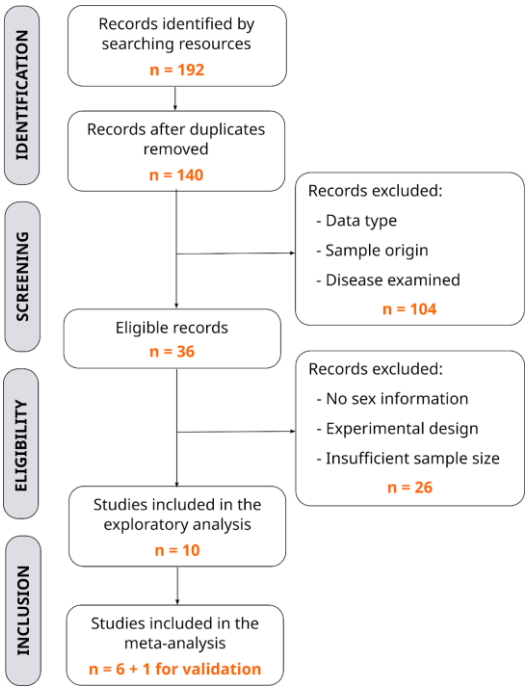


Figure 4.8. Systematic review results following PRISMA guidelines. Specification of remaining study number (n) through the identification, screening, eligibility, and inclusion phases, along with the corresponding exclusion justifications. *PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses*.

	PRJNA684124	PRJNA721421	PRJEB99111	PRJNA748865	PRJNA889427
Databases	ENA and SRA	ENA and SRA	ENA and SRA	ENA and SRA	ENA and SRA
Sequencing platform	Illumina MiSeq	Illumina MiSeq	Illumina MiSeq	Ion Torrent	Illumina MiSeq
Sequencing read type	Paired-end	Paired-end	Single-end	Single-end	Paired-end
16S hypervariable region(s)	V3-V4	V4	V4	V2:V4-V6:V9	V4
Collection date: from – to (years)	2020	2011 – 2019	2013 – 2016	2017 – 2018	2018 – 2019
MS subtypes	RRMS	RRMS, PPMS and SPMS	RRMS	RRMS	RRMS
Country	Italy	USA	USA	Spain	USA
N samples	30	456	116	38	143
N subjects	30	283	116	38	143
N subjects by group*	7:11:8:4	28:182:12:61	25:40:28:23	3:15:16:4	18:80:8:37
Treatments**	Classic immunomodulator (n = 3), Corticosteroid (n = 1), Lymphocyte depleating (n = 1), Lymphocyte retention (n = 1), Untreated (n = 9)	B-cell depleating (n = 26), Benzothiazoles (n = 1), Classic immunomodulator (n = 69), Classic immunomodulator and B-cell depleating (n = 1), Classic immunomodulator and Lymphocyte depleating (n = 1), CNS adhesion inhibitor (n = 31), Corticosteroid (n = 4), Lymphocyte depleating (n = 9), Lymphocyte retention (n = 49), Untreated (n = 52)	Untreated (n = 63)	Classic immunomodulator (n = 12), Lymphocyte retention (n = 3), Untreated (n = 4)	Treated*** (n = 84), Untreated (n = 33)
Other host-related reported variables	Age	Age, BMI	Age	Age, BMI	BMI
PMID	34206853	33876477	28893978	36325343	37758833

(*Continue next page*)

	PRJNA732670	PRJEB34168	PRJNA565173	PRJEB67783	PRJEB32762
Database	ENA and SRA	ENA and SRA	ENA and SRA	ENA and SRA	ENA and SRA
Sequencing platform	Illumina MiSeq	Illumina MiSeq	Illumina MiSeq	Illumina MiSeq	Illumina MiSeq
Sequencing read type	Paired-end	Paired-end	Paired-end	Paired-end	Single-end
16S hypervariable region(s)	V3-V4	V4	V3-V4	V3-V4	V4
Collection date: from – to (years)	2017 – 2019	2013 – 2017	2015	2021 – 2022	2015 – 2019
MS subtypes	RRMS	RRMS	PPMS	RRMS and SPMS	RRMS, PPMS and SPMS
Country	USA	USA	Russia	Italy	USA, Pensilvania, Spain, Scotland, Argentina
N samples	56	65	30	74	2925
N subjects	56	65	30	52	1152
N subjects by group*	28:17:6:5	27:22:12:4	9:6:6:9	8:20:6:18	201:400:375:176
Treatments**	B-cell depleating (n = 3), Classic immunomodulator (n = 13), Untreated (n = 6)	Untreated (n = 26)	Untreated (n = 15)	B-cell depleating (n = 2), Classic immunomodulator (n = 21), Classic immunomodulator and Lymphocyte depleating (n = 1), CNS adhesion inhibitor (n = 7), Lymphocyte depleating (n = 2), Lymphocyte retention (n = 3), Untreated (n = 2)	B-cell depleating (n = 28), Classic immunomodulator (n = 241), CNS adhesion inhibitor (n = 27), Lymphocyte retention (n = 71), Untreated (n = 209)
Other host-related reported variables	Age, BMI	Age, BMI	Age, BMI	Age, BMI	Age, BMI
PMID	35472144	32743517	31888483	39402079	36113426

(Table footnote on the next page)

Table 4.2. Description of the datasets incorporated into the analysis. (*Previous pages*) *Control females : MS females : Control males : MS males. **Treatments were defined as detailed in *Materials and Method* section - *Standardization of metadata nomenclature* subsection, ***Specific treatments types for PRJNA889427 were not reported. *BMI: Body mass index; CNS: central nervous system; ENA: European Nucleotide Archive; MS: multiple sclerosis; PMID: PubMed identifier; PPMS: primary progressive MS; RRMS: relapsing-remitting MS; SPMS: secondary progressive MS; SRA: Sequence Read Archive.*

After processing the datasets and performing the exploratory analyses, three studies were excluded, as detailed in the following sections. Of the remaining seven studies, six were selected to be included in the meta-analysis, while the seventh was used as an independent dataset to validate the meta-analysis results (**Figure 4.8**).

The studies included in the exploratory analysis exhibited a diverse objectives:

- PRJNA684124[407] aimed to identify differences in taxon abundance between MS patients and their household relatives. To minimize dietary variability, participants followed the same diet for two weeks prior to sample collection.
- PRJNA721421[469] characterized microbiota alterations in different MS subtypes and examined their associations with clinical features of the disease.
- PRJEB99111[402] investigated the effect of the gut microbiota on T-cell inflammatory responses. In vitro assays were performed by stimulating PBMCs from MS patients and healthy controls using microbial extracts from their own stool samples.
- PRJNA748865[408] explored the relationship between gut microbiota, SCFAs, diet, and MS.
- PRJNA889427[412] focused on investigating the combined contribution of genetic susceptibility and gut microbiome dysbiosis to RRMS by exploring genetic risk scores and microbiome profiles.
- PRJNA732670[470] focused on characterizing alterations in the gut mycobiome of MS patients compared to healthy individuals. Bacterial microbiome 16S data were also sequenced and analyzed, although its description was not the main goal of the manuscript.
- PRJEB34168[471] explored associations between the gut microbiota and inflammatory T cell subsets in RRMS patients.

- PRJNA565173[472] focused on characterizing the gut microbiota composition at different taxonomic levels comparing PPMS to healthy individuals.
- PRJEB67783[473] investigated the effect of *Mycobacterium avium subspecies paratuberculosis* infection on MS. This study includes MS patients with and without infection, as well as healthy controls. Since the infection status of the other cohorts was unknown, we considered all samples from this study as MS samples.
- PRJEB32762[409] (validation dataset): this study aimed to identify taxonomic and metabolic microbiome differences in MS patients compared to their household relatives, and to explore associations with MS clinical variables. It includes 16S rRNA gene sequencing data, through which the authors investigated alpha and beta diversity metrics. They also generated WGS to identify differential taxonomic and functional patterns. 16S dataset presents a large sample size (N = 1152) from two cohorts collected across seven cities worldwide: San Francisco (N = 328), Boston (N = 84), New York (N = 118), Pittsburgh (N = 24), Buenos Aires (N = 258), Edinburgh (N = 262), and San Sebastián (N = 78). Given these characteristics, the PRJEB32762 dataset was selected to perform an independent validation of the meta-analysis results.

Results derived from their individual processing are detailed in the next section.

4.4.2. COMPUTATIONAL DATA PROCESSING

This section outlines the results from the processing steps applied to each of the ten studies selected during the eligibility phase of the systematic review. The workflow included sequencing read processing, discarding low-quality samples, identifying potential sources of technical variability, and obtaining the taxonomic abundance matrices.

After downloading the FASTQ files, sequencing reads were filtered and trimmed. Next, DADA2 pipeline was applied for ASV inference. The outputs from this pipeline by dataset can be consulted in **Supplementary Figures 4.S1-4.S10**, which included the error rate modelling and the percentage of reads retained after each step of the pipeline. As a result, the ASV abundance matrix was obtained for each dataset, and ASVs were annotated at different taxa levels with the GTDB database. Further details of these steps can be consulted in the *Materials and Methods* section.

Most of the studies included an experimental design in which one sample per individual was processed. However, the datasets PRJNA721421, PRJEB67783, and PRJEB32762 presented a more complex design:

- In the PRJNA721421[469] study, two separate sequencing batches were established, labelled WL14 and WL17. 62% of the samples were sequenced in both runs, 19% only in WL14, and the remaining 19% only in WL17. After evaluating the distribution of library sizes and the number of observed taxa per sample, we identified that sequencing depth was higher and more homogeneous in run WL17 compared to run WL14 (**Supplementary Figure 4.S11**). Therefore, samples from run WL17 were selected for downstream analyses, while those from run WL14 were excluded.

- In the case of PRJEB67783[473], longitudinal sampling was performed for 57% of participants, with approximately one month between the first and the second time points. Since this is the only study that partially allows longitudinal analysis, samples from the second time point were excluded.

- The PRJEB32762[409] dataset presented the most complex design. Two sample collection methods were employed: Q-tip (dry) and snap-frozen (wet). Furthermore, different DNA extraction protocols were used across cohorts. Samples in the first cohort were processed using a QIAcube platform according to manufacturer protocols (QIAGEN), whereas samples from the second cohort were processed using the MagAttract PowerSoil DNA EP Kit (ref. 27100–4-EP). Some samples from each cohort were cross-processed with the alternative method to assess potential biases, and no substantial differences were reported. Additionally, samples from each individual were repeatedly sequenced ranging from 1 to 9 runs, which could be either Q-type or S-type. The original authors conducted the aggregation of the counts across all replicates per individual. In this doctoral thesis, we tested two strategies: selecting a single replicate for each individual and summing all replicates per individual. We found negligible differences in the results of the meta-analysis and their validation (data not shown). Consequently, we adopted the authors' approach and proceeded with the aggregated count matrix, summing all replicates per individual.

Once the final design for each dataset was established, low-quality samples were filtered out based on different criteria. Specifically, samples with fewer than 1,000 total counts or fewer than 30 distinct taxa were discarded for further analysis. Samples were also excluded if the rarefaction curve ended in the exponential phase failing to reach a *plateau,* indicating insufficient sequencing depth that led to incomplete taxonomic

detection. In addition, relative abundance plots of the top 15 taxa were evaluated to identify the presence of potential contamination or unexpected proportions of different taxa. The corresponding figures representing the distribution of the metrics by dataset can be consulted in **Supplementary Figures 4.S12-4.S21**.

For illustration purposes, **Supplementary Figure 4.S22** shows the rarefaction curves for a subset of samples from the PRJEB67783 dataset. Samples discarded for not reaching the *plateau* are marked in red. It should also be noted that replicates of the PRJEB32762 dataset were evaluated individually. The distribution of the number of counts and the number of distinct genera did not differ between sample types (Q-type and S-type), as represented in **Supplementary Figure 4.S23.**

Overall, the summary of discarded samples is provided in **Supplementary Table 4.S3,** and the final sample size and the number of distinct taxa per dataset are summarized in **Table 4.3**.

Table 4.3. Summary of the sample and genera sizes for each dataset. The table reports the number of samples after sample quality filtering, stratified by dataset, sex, and condition. Column 7 defines the number of distinct genera in each dataset. *Unique genera considering all datasets. PRJEB32762 validation dataset. *MS: multiple sclerosis.*

Dataset	Control females	MS females	Control males	MS males	N total	N genera
PRJNA684124	7	11	8	4	30	500
PRJNA721421	10	153	6	56	225	598
PRJEB99111	25	39	28	21	113	717
PRJNA748865	3	15	16	4	38	227
PRJNA889427	18	80	8	37	143	584
PRJNA732670	28	17	9	5	56	518
PRJEB34168	27	22	12	4	65	497
PRJNA565173	9	6	6	9	30	429
PRJEB67783	7	19	5	12	43	430
Total	134	362	95	152	743	1059*
PRJEB32762	201	400	375	176	1152	970

Additional technical effects were considered when evaluating the variability within each dataset, and its potential suitability for meta-analysis. Specifically:

- In the PRJEB99111 study, two potential sources of batch effects were identified: the sequencing set and the sequencing center. Regarding the sequencing batch, the first set exhibited a substantially larger library size compared to the others; however, this did not result in a higher number of taxa being identified (**Supplementary Figure 4.S24**). Considering the sequencing center, no additional differences were observed.

- In the PRJNA748865 study, the number of distinct taxa identified was considerably lower than in the other datasets. While most studies identified approximately 100 genera per sample, this dataset presented an average of 50 genera (**Supplementary Figures 4.S12-4.S20, panel B**). Furthermore, for PRJNA748865, a large proportion of reads were classified under the generic label *Bacteria unclassified*, likely due to the reduced resolution of the amplicons resulting from combining primers targeting V2 to V9 hypervariable regions (**Supplementary Figure 4.S15D-E**, **Table 4.2**).

- In the PRJNA721421 and PRJNA889427 studies, an unexpectedly high relative abundance of the *Blautia* genus was observed, along with a lower-than-expected abundance of typically dominant genera such as *Bacteroides*, *Phocaeicola* and *Prevotella* (**Supplementary Figures 4.S13 and 4.S16**). These patterns suggest technical artifacts, potentially due to inaccurate sample processing during the centrifugation step, where smaller bacterial cells may have been removed.

With the datasets curated and potential technical sources of variability identified, we next proceed to the exploratory analysis of the studies to evaluate their suitability for inclusion in the meta-analysis.

4.4.3. Assessment of dataset heterogeneity for meta-analysis inclusion

We explored the overall taxonomic structure and potential grouping patterns of the nine datasets selected for potential inclusion in the meta-analysis. This step aimed to assess whether the combined datasets exhibited consistent microbiome profiles, and to identify any datasets that might need to be excluded due to substantial divergence from the others.

From the combined abundance matrix comprising 743 samples and 1,059 genera, we quantified the proportion of variance in microbial composition explained by individual variables. All the variables evaluated with univariate models were statistically significant: *dataset* (23.1%), *condition* (1.88%), *BMI* (0.80%), *age* (0,35%) and *sex* (0.34%). The cumulative variance included all these variables, explaining 29,5% of the total with p-value cutoff set at 0.05. *Dataset* was the first variable incorporated in the model, as it accounted for the largest proportion of variability. Notably, the variance explained by the biological variables (*condition*, *BMI*, *age*, *sex*) was lower but not redundant with the *dataset* variable, as they were found significant in the multivariate analysis (**Figure 4.9-A**).

Additionally, the same procedure was applied to the subset of samples from MS individuals, incorporating variables specific to the diseased population: *MS subtype, treatment* and *treatment type.* Results can be consulted in **Figure 4.9-B.** In the univariate analysis, all variables significantly explained a fraction of the microbial composition. Meanwhile, the multivariate model included *dataset*, *treatment type* and *BMI*, which together accounted for 27.9% of the total variance (p-value < 0.05). The variable *treatment* was identified as redundant, likely due to its overlap with *treatment type*. Meanwhile, *MS subtype, age* and *sex* did not contribute to explain additional variance beyond the variables already included in the model.

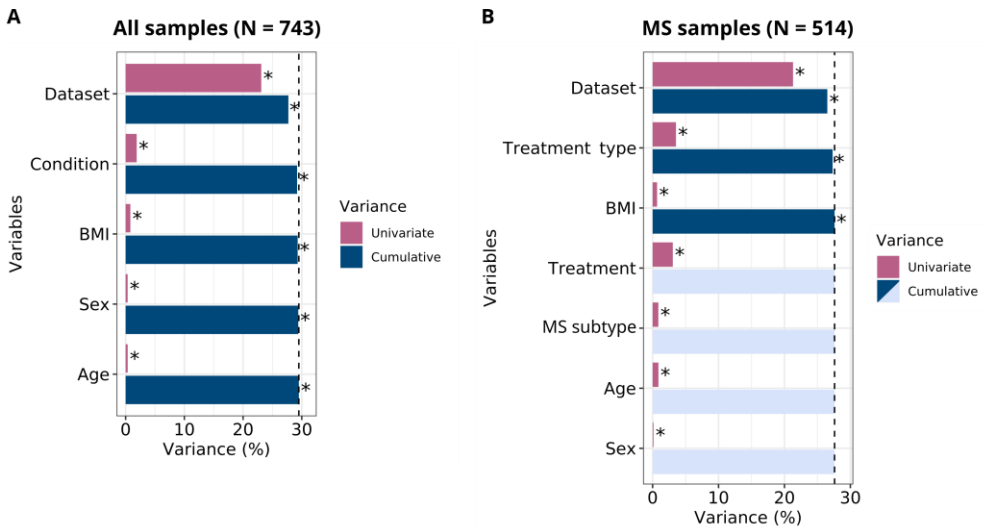


Figure 4.9. Variable contribution to microbiome compositional variation considering (A) all samples and (B) multiple sclerosis samples from the combined dataset conformed by the nine selected studies. Results considering each variable independently (univariate variance) or

in the multivariate model (cumulative variance). The asterisk (*) indicates if the contribution is significant; the black dashed line represents the cut-off for significant non-redundant contribution to the multivariate model considering p-value < 0.05. Sample size (N) is reported from univariate analysis. For multivariate models, samples with missing data were excluded. (A) Number of missing values: age (N = 148), BMI (N = 215). (B) Number of missing values: age (N = 117), BMI (N = 130), MS subtype (N = 3), Treatment type (N = 84). Datasets included in the analysis: PRJNA684124, PRJNA721421, PRJEB99111, PRJNA748865, PRJNA889427, PRJNA732670, PRJEB34168, PRJNA565173, and PRJEB67783. *BMI: body mass index; MS: multiple sclerosis.*

To investigate the existence of distinct microbial community types, we applied the DMM modelling. In cohorts that accurately reflect population variability, this clustering strategy often reveals enterotypes. More details are described in the *Introduction* section of this chapter.

The results revealed, via the Laplace approximation, that the optimal number of clusters was five (**Figure 4.10-A**). Taxa distribution by cluster can be consulted in **Supplementary Figure 4.S25**. When visualizing the distribution of these clusters, we observed that some of them were composed of specific datasets (**Figure 4.10-B,C**). This aligns with the previous results, where the dataset variable explained most of the microbial compositional variation. Specifically, cluster 4 consisted entirely of samples from PRJNA748865 and was considerably distant from the other clusters. Moreover, cluster 1 was composed predominantly of samples from datasets PRJNA721421 and PRJNA889427, in which we previously identified atypical taxonomic profiles (**Supplementary Figures 4.S13 and 4.S16, description in page 158**).

Based on these findings, PRJNA748865, PRJNA721421 and PRJNA889427 datasets were excluded from downstream analysis. Consequently, the meta-analysis included six studies, whose characteristics are summarized in **Table 4.4.**

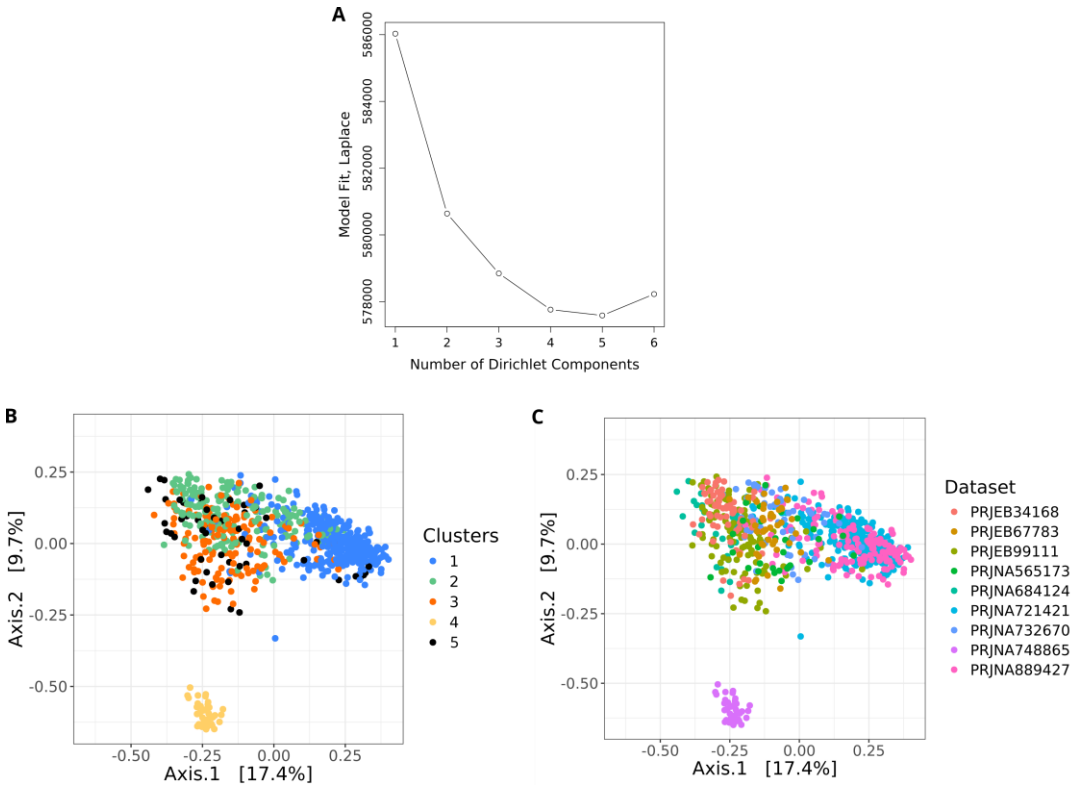


Figure 4.10. Dirichlet multinomial mixture clustering results for the combined dataset conformed by the nine selected studies. (A) Laplace approximation results. (B-C) Dotplot representation based on the first two principal coordinates of Bray–Curtis distances, with samples colored (B) by the cluster assignment derived from the Dirichlet multinomial mixture model and (C) by dataset. Datasets included in the analysis: PRJNA684124, PRJNA721421, PRJEB99111, PRJNA748865, PRJNA889427, PRJNA732670, PRJEB34168, PRJNA565173, and PRJEB67783.

Table 4.4. Summary of the host-associated variables across the six datasets included in the meta-analysis. For categorical variables, the number of samples for each category is shown. For continuous variables (age and BMI), the mean value is reported in parentheses. *BMI: body mass index; MS: multiple sclerosis; PPMS: primary progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis.*

		PRJNA684124	PRJEB99111	PRJNA732670	PRJEB34168	PRJNA565173	PRJEB67783
Total samples		30	113	56	65	30	43
Condition	Control	15	53	34	39	15	12
	MS	15	60	22	26	15	31
Sex	Female	18	64	45	39	15	26
	Male	12	49	11	16	15	17
Age	Available (mean)	30 (60.67)	108 (43.19)	56 (42.91)	65 (43.49)	30 (40.30)	43 (40.73)
	Not available	0	5	0	0	0	0
BMI	Available (mean)	0	0	55 (26.53)	52 (28.19)	30 (23.46)	43 (23.37)
	Not available	30	113	1	13	0	0
Country		Italy	USA	USA	USA	Russia	Italy
MS subtype	Available (N)	RRMS (15)	RRMS (60)	RRMS (22)	RRMS (26)	PPMS (15)	RRMS (23) and PPMS (5)
	Not available	0	0	0	0	0	3
Treatment (N)		Treated (6) Untreated (9)	Untreated (60)	Treated (16) Untreated (6)	Untreated (26)	Untreated (15)	Treated (30) Untreated (1)

We then repeated the exploratory analyses for the six selected studies (**Supplementary Figure 4.S26**). When considering all samples (N = 337), all variables except *age* were significant, although *age* was close to significance (p-value = 0.06) (**Supplementary Figure 4.S26-A**). The *dataset* variable remained the main contributor to variation in microbial composition, explaining 12.4% of the total (p-value < 0.05). In the multivariate model, the variables *dataset*, *condition*, and *BMI* were significant, resulting in a total variance of 17.2% (p-value < 0.05). Among MS patient samples (N = 169), microbiome composition was significantly associated with *dataset* (15.5%), *MS subtype* (3.8%), *BMI* (2.7%), *treatment* (2.3%), *treatment type* (2.1%), and *sex* (0.7%), but not *age* (**Supplementary Figure 4.S26-B**). In this subset, we were unable to construct a multivariate model that explained a greater proportion of variance than that already accounted for the *dataset* variable alone.

When applying the DMM model, the optimal number of clusters identified was three (**Figure 4.11**). In this scenario, the microbial distribution across clusters was more closely to the expected patterns (**Figure 4.11-B,C**). Cluster k3 was characterized by a high relative abundance of *Prevotella*, while k2 was dominated by *Bacteroides* and *Phocaeicola*, which are typically inversely correlated with *Prevotella*. The third cluster k1 displayed greater microbial diversity, including elevated proportions of genera such as *Alistipes* and *Ruminococcus*, potentially representing the *Ruminococcus*-driven enterotype, commonly associated with slower transit time. We did not detect the expected fourth enterotype, also characterized by high relative abundance of *Bacteroides* and *Phocaeicola*. Furthermore, taxa such as *Methanobrevibacter*, *Faecalibacterium*, and *Akkermansia* did not show differential abundance across k1 and k2, despite being expected to be higher in k1 (the *Ruminococcus*-associated enterotype) compared to k2 (the *Bacteroides/Phocaeicola* dominant).

Similarly, we conducted DMM modelling at individual dataset level. Consistent with previous findings, no robust clustering patterns were observed, likely due to the limited sample size within each study (**Supplementary Table 4.S4**). Given these results, along with the persistent influence of dataset-specific variability, we decided not to seek the enterotype assignment, as their identification could not be reliably supported. Overall, **Figure 4.12** illustrates the distribution of samples according to their assigned cluster, source dataset, condition, and sex.

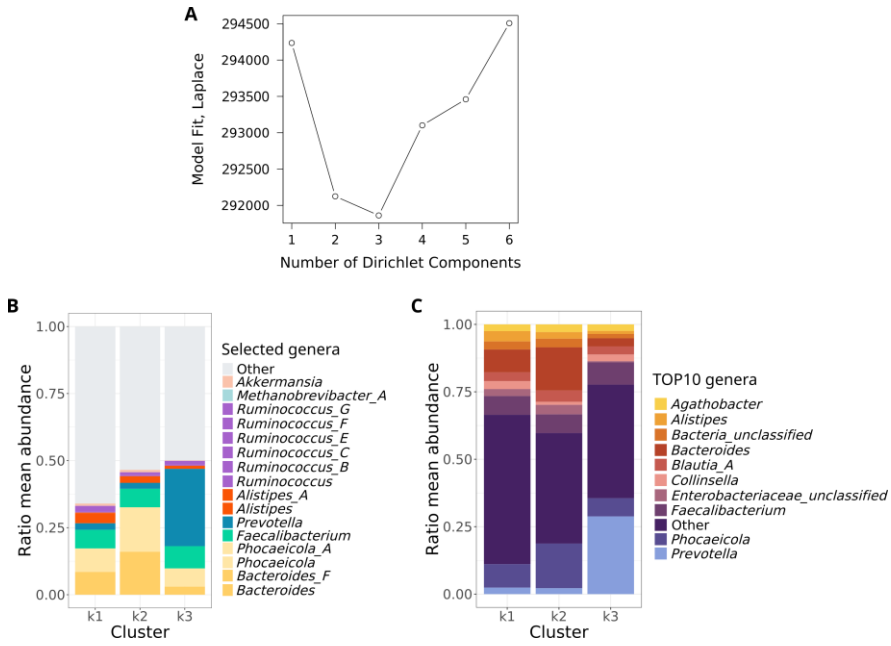


Figure 4.11. Dirichlet multinomial mixture clustering results for the combined dataset conformed by the six datasets included in the meta-analysis. (A) Laplace approximation results. (B-C) Relative genera distribution for each cluster. Datasets included in the analysis: PRJNA684124, PRJEB99111, PRJNA732670, PRJEB34168, PRJNA565173, PRJEB67783.

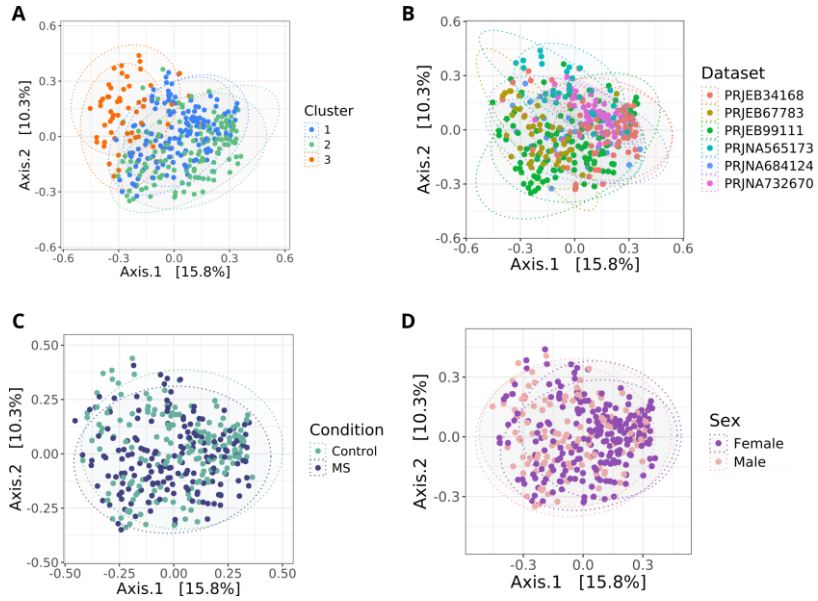


Figure 4.12. PCoA plots showing sample distribution for the combined dataset conformed by the six datasets included in the meta-analysis. Dotplot representation based on the first two principal coordinates of Bray–Curtis distances, with samples colored by (A) cluster

assignment, (B) source dataset, (C) condition, and (D) sex. Datasets included in the analysis: PRJNA684124, PRJEB99111, PRJNA732670, PRJEB34168, PRJNA565173, PRJEB67783. *MS: multiple sclerosis; PCoA: Principal Coordinates Analysis.*

The final metric calculated in the exploratory analysis was alpha diversity, which reflects within-sample microbial richness and evenness. Significant differences were identified across datasets (Kruskal-Wallis test, p-value = $3.55*10^{-5}$) (**Figure 4.13-A**). The datasets PRJEB34168, PRJEB99111, and PRJNA684124 exhibited lower alpha diversity values compared to PRJEB67783, PRJNA565173, and PRJNA732670. A plausible explanation is that PRJEB99111 and PRJNA684124 targeted the V4 hypervariable region, while the three more diverse datasets used the V3–V4 region, providing higher taxonomic resolution. Although PRJNA684124 was also sequenced using the V3–V4 region, participants were instructed to follow a standardized diet for two weeks prior to sampling, which could have reduced interindividual variability and, consequently, their alpha diversity values.

In stratified groups combining condition and sex variables, no significant differences were identified (Kruskal-Wallis test, p-value = 0.21) (**Figure 4.13-B**). However, the evaluation of these groups within individual datasets revealed specific differences in alpha diversity distributions (**Supplementary Figure 4.S27**).

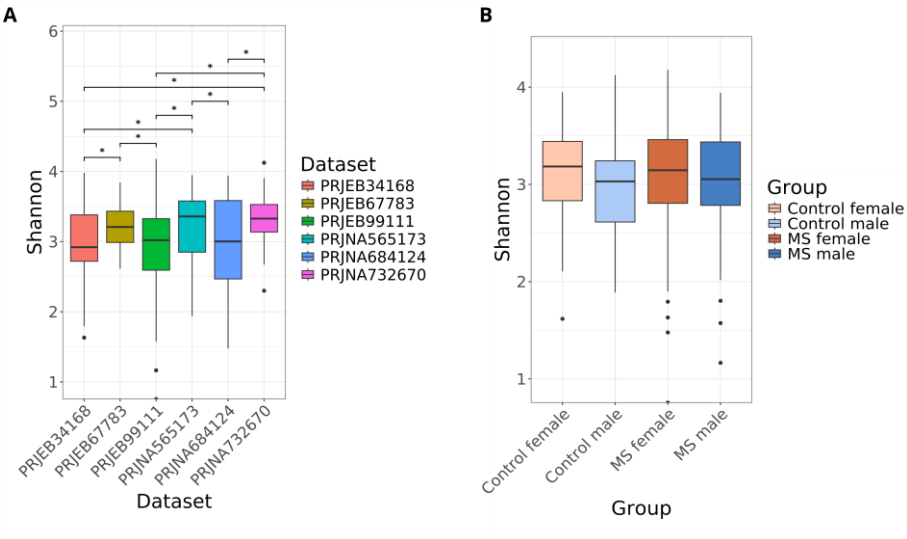


Figure 4.13. Alpha diversity distribution across the six datasets included in the meta-analysis (A) by dataset and (B) by condition and sex. Asterisks (*) indicate pairwise significant differences with adjusted p-values < 0.05 from Dunn's post hoc test (after p-value < 0.05 in Kruskal–Wallis test). *MS: multiple sclerosis.*

4.4.4. INDIVIDUAL ANALYSIS BY DATASET

4.4.4.1. Within-dataset variability characterization

After characterizing the differences and similarities across datasets, we explored the potential sources of variability within each individual dataset. A summary of the results is presented in **Table 4.5.**

Table 4.5. Within-dataset variability determined by distance-based redundancy analysis. For each dataset, the table shows: i) variables identified as significant in the univariate model along with the percentage of variance they explain (in parentheses), and ii) variables tested that were not found to be significant. Significance assigned when p-value < 0.05. *BMI: body mass index; MS: multiple sclerosis.*

Dataset	Samples	Significant variables (% variance)	Not significant variables
PRJNA684124	All samples	Sex (2.27%)	Condition, age
	MS samples	None	Treatment, treatment type
PRJEB99111	All samples	None	Condition, age, sex, sequencing set, center
PRJNA732670	All samples	Condition (3.1%)	BMI, age, sex
	MS samples	None	Treatment, treatment type
PRJEB34168	All samples	Condition (1.22%)	BMI, age, sex
PRJNA565173	All samples	BMI (3.71%), condition (2.38%)	Age, sex
PRJEB67783	All samples	None	Condition, BMI, age, sex
	MS samples	None	Treatment, treatment type, MS subtype

Most of the covariates evaluated did not explain a significant proportion of the variability within the datasets, possibly due to insufficient sample size. However, *condition* (MS or control) emerged as significant in three of the six datasets, accounting for approximately 1% to 3% of the variance. Additionally, in the dataset PRJNA684124, the sex variable was also found to be significant.

Lastly, in the PRJNA565173 dataset, both *BMI* and *condition* were identified as significant variables in the univariate analysis. When included together in a multivariate model, the inclusion of *condition* did not significantly increase the explained variance beyond the one explained by the BMI. This result suggests partial redundancy, potentially due to differences in BMI distribution across experimental groups. Thus, a Wilcoxon rank-sum test comparing BMI distribution between MS and control subjects was performed. As a result, BMI values appeared higher in controls, but this difference did not reach statistical significance ($p = 0.115$) (**Supplementary Figure 4.S28**).

Based on these findings, covariate adjustment was not included in the differential abundance analysis. Instead, we applied the Wilcoxon rank-sum test to assess differential abundance within each dataset individually.

4.4.4.2. Individual differential abundance results

Differential abundance analyses were conducted separately for each dataset, considering the following four pairwise comparisons to explore the effects of sex and condition:

- Impact of sex in the control group (Control comparison): control females *vs.* control males.
- Impact of sex in the MS group (MS comparison): MS females *vs.* MS males.
- Impact of disease in females (Female comparison): MS females *vs.* control females.
- Impact of disease in males (Male comparison): MS males *vs.* control males.

Effect size patterns across studies can be visualized in **Figures 4.14-4.17.** All the outcomes presented high variability at the individual study level.

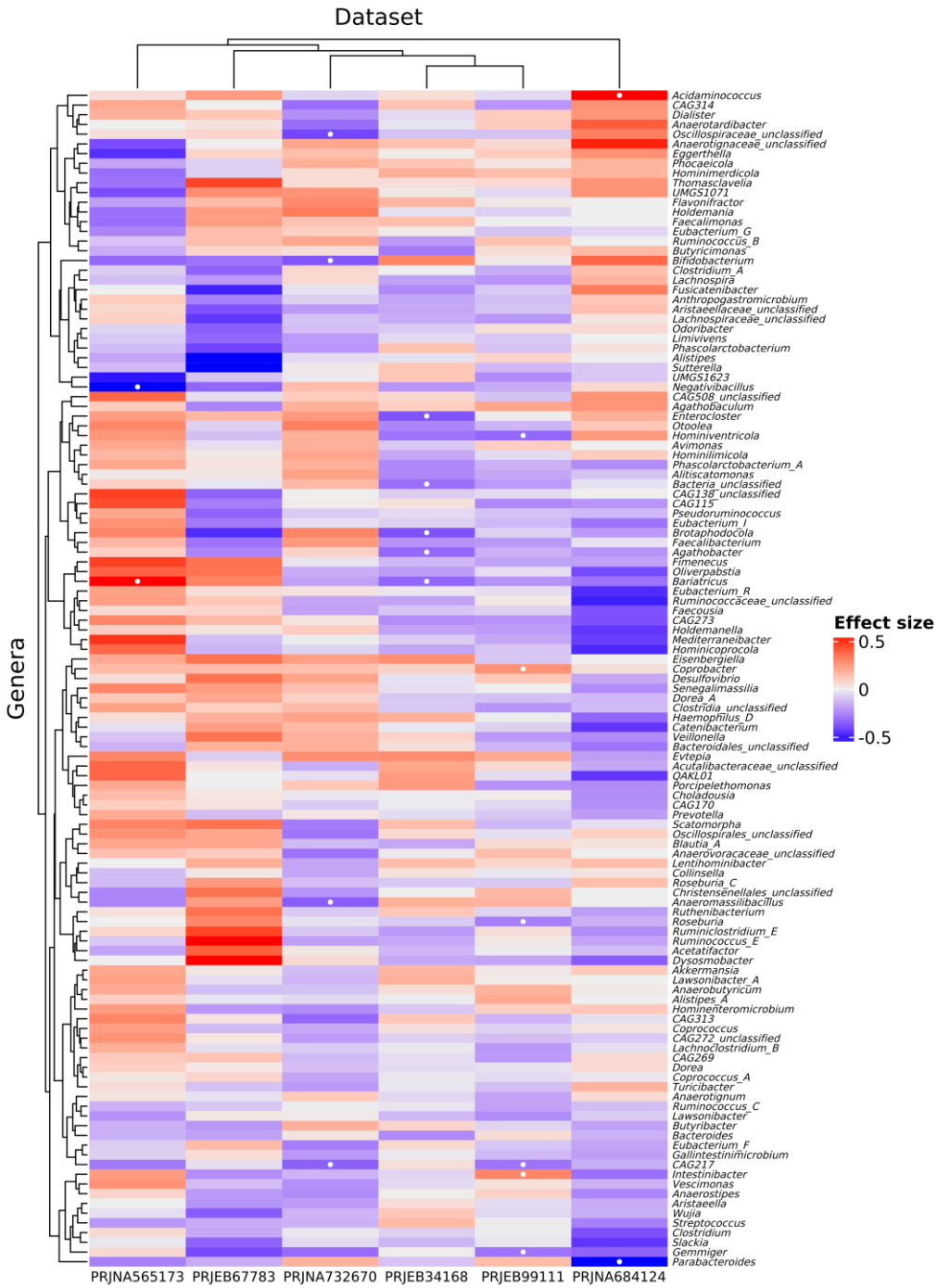


Figure 4.14. Individual differential abundance results when comparing control females *versus* control males. Rows represent genera and columns datasets. The color scale reflects the magnitude and direction of the effect size, where red indicates genera enriched in control females and blue indicates genera enriched in control males. White dots denote p-value < 0.05.

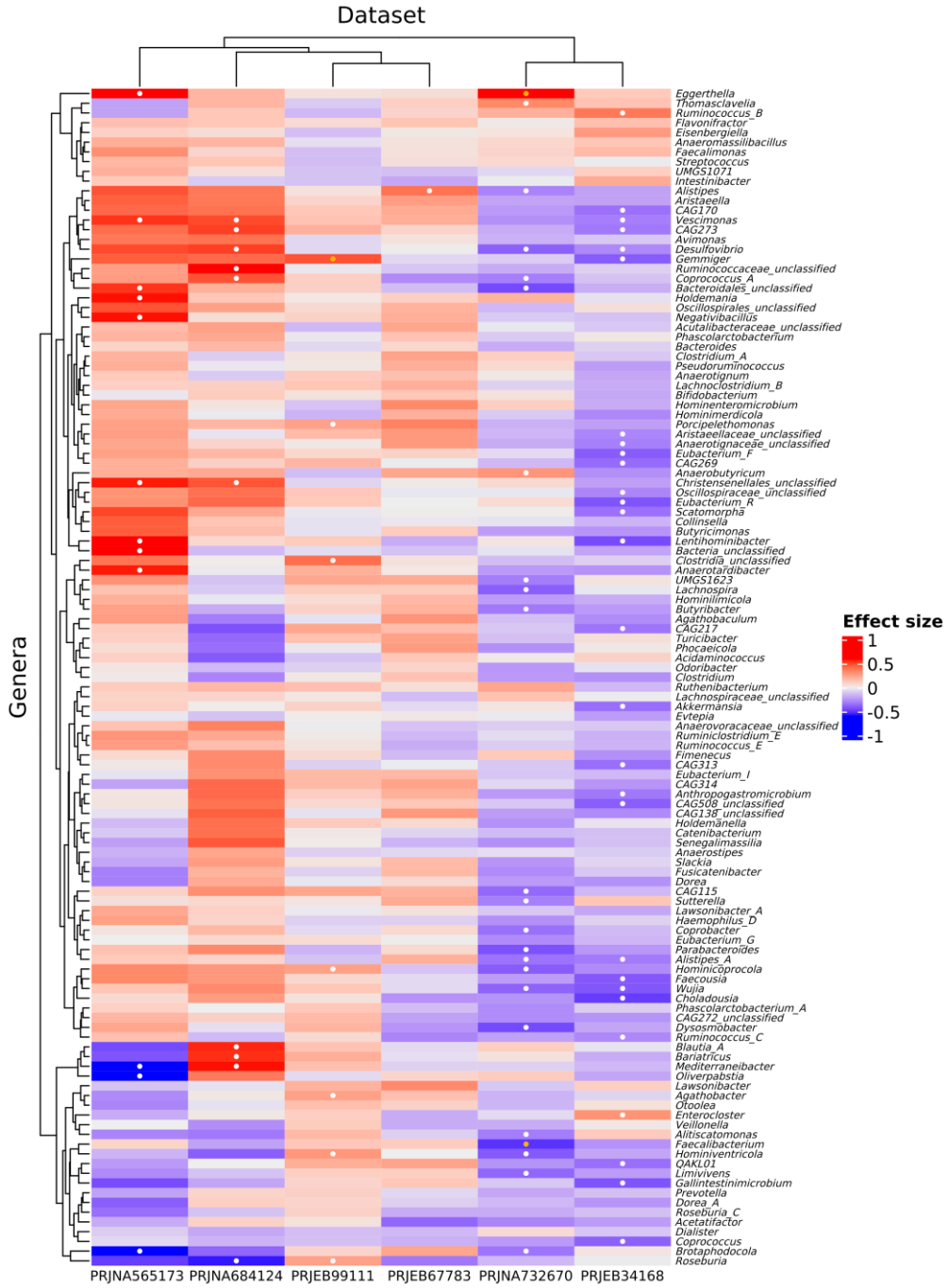


Figure 4.15. Individual differential abundance results when comparing multiple sclerosis females *versus* control females. Rows represent genera and columns datasets. The color scale reflects the magnitude and direction of the effect size, where red indicates genera enriched in multiple sclerosis females and blue indicates genera enriched in control females. White dots denote p-value < 0.05, and yellow dots adjusted p-value < 0.05.

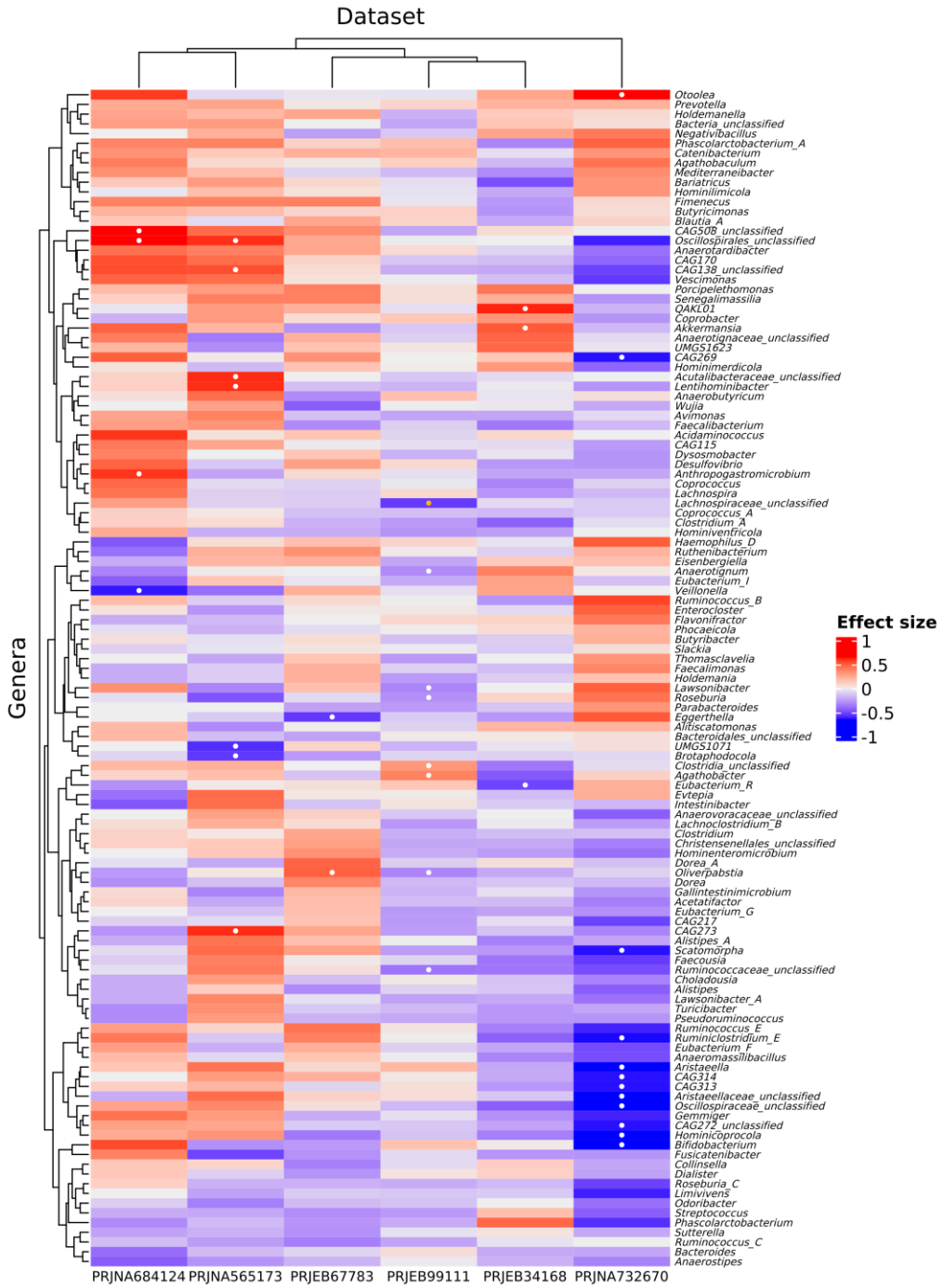


Figure 4.16. Individual differential abundance results when comparing multiple sclerosis males *versus* control males. Rows represent genera and columns datasets. The color scale reflects the magnitude and direction of the effect size, where red indicates genera enriched in multiple sclerosis males and blue indicates genera enriched in control males. White dots denote p-value < 0.05, and yellow dots adjusted p-value < 0.05.

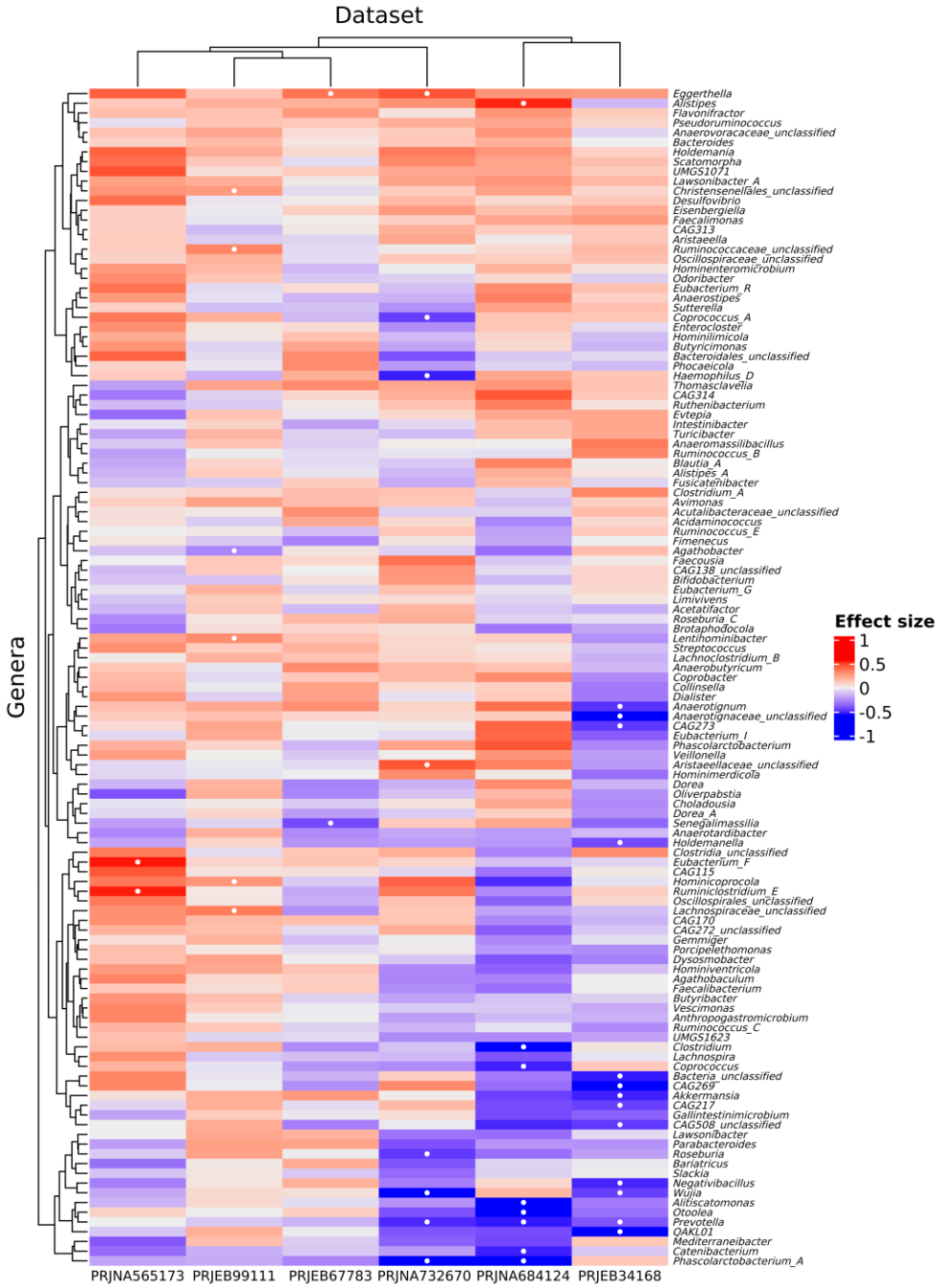


Figure 4.17. Individual differential abundance results for each dataset when comparing multiple sclerosis females *versus* multiple sclerosis males. Rows represent genera and columns datasets. The color scale reflects the magnitude and direction of the effect size, where red indicates genera enriched in multiple sclerosis females and blue indicates genera enriched in multiple sclerosis males. White dots denote p-value < 0.05. MS: multiple sclerosis.

In the Control comparison (**Figure 4.14**), we observed the lowest number of taxa showing a significant trend (raw p-value < 0.05), and generally lower effect size magnitude with respect to the other comparisons. The most divergent results were observed in the dataset PRJNA684124, where participants followed a standardized diet likely affecting results in healthy individuals.

In the Female comparison (**Figure 4.15**), datasets PRJNA732670 and PRJEB34168 exhibited similar differential abundance patterns, which differed from the remaining datasets. These discrepancies were less pronounced in the Male comparison (**Figure 4.16**), although PRJNA732670 was again the most divergent. Notably, MS comparison (**Figure 4.17**) revealed the most consistent pattern across datasets. In particular, the upper section of the heatmap showed taxa that tend to be more abundant in females than in males. Moreover, datasets PRJEB99111 and PRJEB34168 are the most divergent, with a subset of taxa showing significant increased abundance in MS males, and larger effect sizes compared to the other datasets (bottom section of the heatmap).

To further quantify the between-dataset variability observed in the differential abundance results, we evaluated the effect size consistency across six studies in the four comparisons. In detail, we counted, for each taxon, the number of datasets in which the effect size was positive (**Figure 4.18**). Under high consistency, we would expect this number to be either 6 or 0, reflecting a uniform direction of change across all datasets (either all datasets pointing to positive effect sizes in the first group or all in the second group, respectively). A minority of taxa showed complete agreement. Still, 15.7%, 5.7%, 19.8%, and 10.7% in the Control, Female, Male, and MS comparisons, respectively, displayed near-consistent patterns (0–1 or 5–6 positive values), pointing to reproducible signals across cohorts.

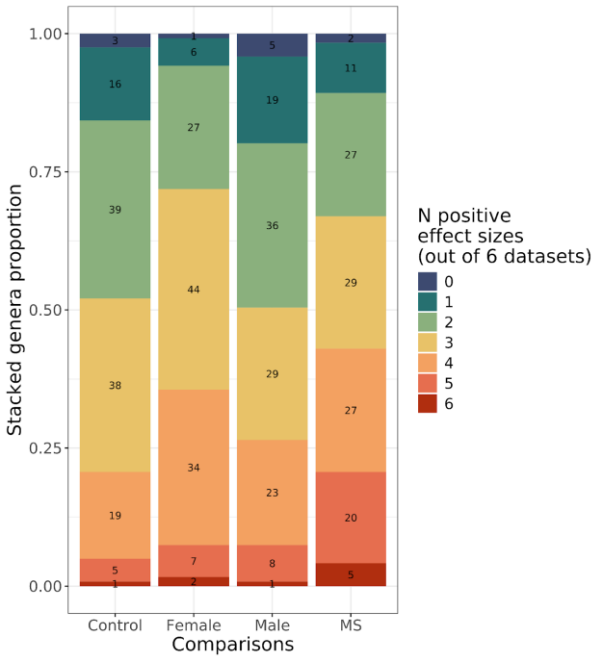


Figure 4.18. Proportion of genera by number of datasets with positive effect sizes across comparisons. Bars are segmented by the number of datasets (from 0 to 6) in which each genus showed a positive effect size (stacked representation). The total count of genera is indicated within the segments. Comparisons: Control (control females *vs.* control males), MS (MS females *vs.* MS males), Female (MS females *vs*. control females), and Male (MS males *vs.* control males). *MS: multiple sclerosis.*

4.4.5. META-ANALYSIS

The individual differential abundance results were integrated through meta-analyses performed separately for each of the four comparisons. This approach allowed the identification of consensus differential abundance signatures across studies in each scenario. All results can be explored through the developed web tool, which is described later in this manuscript.

Specifically, no genera showed significant differential abundance by sex when comparing healthy females to healthy males. When evaluating the effect of disease separately for each sex, we observed a reduced abundance of *Coprococcus* in MS females compared to control females, whereas *Catenibacterium* was significantly more abundant in MS males compared to control males. The highest number of significant taxa was identified in the comparison between MS females and MS males, with nine genera found to be more abundant in females and two in males (**Table 4.6**).

Table 4.6. Significant differential abundant genera identified through meta-analysis integration approach for each comparison. The table indicates the genera found to be significantly more abundant (adjusted p-value < 0.05) in one group compared to the other. Comparisons: Control (control females *vs.* control males), MS (MS females *vs.* MS males), Female (MS females *vs.* control females), and Male (MS males *vs.* control males). *MS: multiple sclerosis.*

Comparison	More abundant in	Genera
Control	Control females	None
	Control males	None
Female	MS females	None
	Control females	*Coprococcus*
Male	MS males	*Catenibacterium*
	Control males	None
MS	MS females	*Desulfovibrio, Eggerthella, Eisenbergiella, Flavonifractor, Holdemania, Lawsonibacter_A, Pseudoruminococcus, Scatomorpha, UMGS1071*
	MS males	*Catenibacterium, Prevotella*

For each taxon included in the meta-analysis, the corresponding statistical metrics were obtained. Results were visualized through three types of plots: forest plots, funnel plots, and influence plots. As an illustrative example, **Figure 4.19** displays the results for two significant genera from the comparison between MS females and MS males: *Flavonifractor* and *Prevotella*.

In the case of *Flavonifractor,* all individual studies reported a positive effect size, indicating consistently higher abundance in MS females than in MS males (**Figure 4.17**). Consequently, the combined effect size is also positive, and its 95% confidence interval (represented by a red diamond in the forest plot) did not cross the zero value, indicating robustness of the result (**Figure 4.19-A**). The funnel plot indicates an absence of anomalous variability in any dataset, as all dots fall within the region defined by the confidence limits. Regarding the influence plot, PRJNA684124 dataset (identified as number 5 in the graphical representation) had the greatest influence on the overall result, as reported in the diagnostic metrics.

Meanwhile, *Prevotella* presented a negative combined effect size, driven by five out of six studies reporting a greater abundance in MS males (**Figure 4.17**), which was confirmed in the outcome of the meta-analysis (**Figure 4.19-B**). The funnel plot revealed slightly more dispersion in this genus, with two studies falling on the boundary of the 95% confidence region, though none outside it. No single study was identified as contributing with anomalous high influence on the influence plots.

Plots for the remaining 10 significant genera are shown in **Supplementary Figures 4.S29-4.S32**. Additionally, the number of individual datasets contributing with positive effect size (and, therefore, the number contributing with negative effect size as the difference from the total) for each significant taxa and comparison is summarized in **Supplementary Figure 4.S33**.

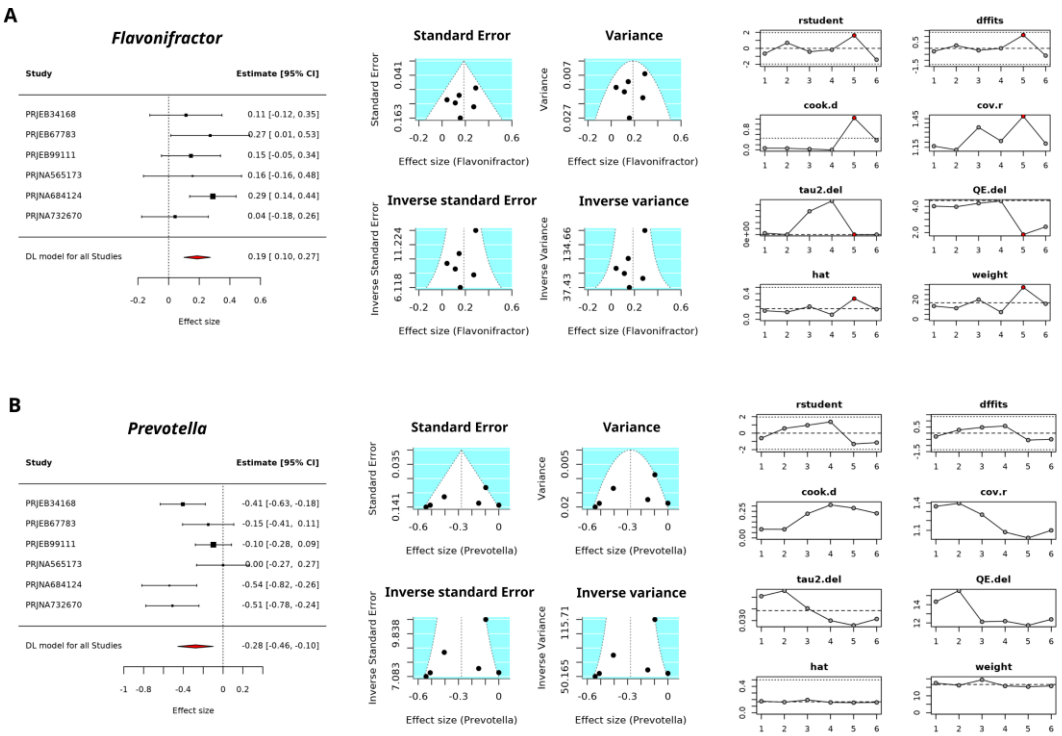


Figure 4.19. Meta-analysis results for two representative genera when comparing MS females *versus* MS males. (A) *Flavonifractor*, (B) *Prevotella*. From left to right: i) Forest plots: each square represents the effect size from an individual dataset, with horizontal lines indicating the 95% confidence intervals. The diamond at the bottom represents the combined effect size and its confidence interval. Positive values indicate higher abundance in MS females; negative values indicate higher abundance in MS males. ii) Funnel plots: effect sizes (X-axis) compared to precision estimates (Y-axis). Each point represents an individual dataset effect size plotted

against its standard error (top left), sample variance (top right), inverse standard error (bottom left), and inverse sampling variance (bottom right). The white triangle shows the 95% confidence region under the null hypothesis of no bias. iii) Influence plots: assessment of each study's influence on the combined effect size using multiple diagnostic metrics. Study ID numbers are ordered based on their position in the forest plot (e.g., PRJEB34168 corresponds to study 1). Detailed descriptions of the metrics are defined in the *Materials and Methods* section. Red dots: influential studies; grey dots: non influential studies. *CI: Confidence Interval; DL: DerSimonian-Laird; ID: Identifier.*

To better characterize the patterns of the significant genera identified in the meta-analysis, the results of the four pairwise comparisons are represented simultaneously in **Figure 4.20**.

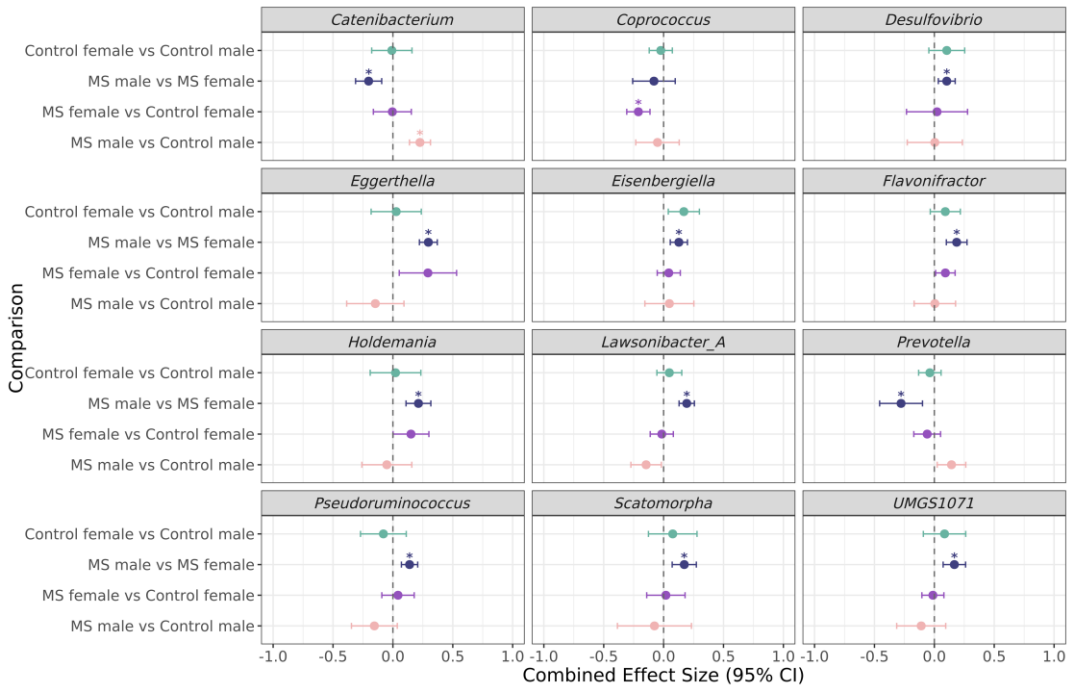


Figure 4.20. Significantly differentially abundant genera identified through meta-analysis in at least one of the comparisons. X-axis indicates the combined effect size, with horizontal lines representing the 95% confidence interval. Rows represent the four pairwise comparisons: Control (control females *vs*. control males), MS (MS females *vs.* MS males), Female (MS females *vs.* control females), and Male (MS males *vs.* control males). Significant effect sizes are marked with an asterisk. A positive effect size indicates higher abundance in the first group; a negative value indicates higher abundance in the second. *MS: multiple sclerosis, CI: confidence interval.*

For *Coprococcus*, which shows significantly lower abundance in MS females compared to control females, we observed minimal differences between sexes in the control group. Interestingly, there is a slight trend toward increased abundance in MS males relative to control males, which may contribute to the absence of significant differences between MS females and MS males. Meanwhile, *Catenibacterium* exhibits no appreciable differences between control males and females, and between control females and MS females. However, its significant increase in MS males compared to control males is reflected in its significant difference observed in the disease comparison, with significant higher abundance in MS males compared to MS females.

Regarding the genera that showed significantly higher abundance in MS females compared to MS males, we identified the following patterns:

- *Desulfovibrio* and *Eisenbergiella*. The sex-related differences observed in MS patients appeared to originate from pre-existing differences between males and females in the control group. No changes were detected when comparing MS to control within each sex.

- *Holdemania*. This genus showed a sex-specific increase trend, with higher abundance in MS females compared to control females. No significant differences were observed in Control and Male comparisons.

- *Eggerthella*. This genus exhibited opposing patterns depending on sex. In males, its abundance tends to decrease in MS compared to controls, while in females, it increased in MS relative to controls.

- *Flavonibactor*, *Lawsonibacter*, *Pseudoruminococcus*, *Scatomorpha*, and *UMGS1071*. These genera present differential abundance patterns that suggest a combined influence of both sex and disease status. Specifically:

    - *Flavonifractor*. Its tendencies pointed toward higher abundance in control females compared to control males, which is further accentuated in MS females relative to control females. No change is observed between MS males and control males.

    - *Lawsonibacter*, *Scatomorpha*, and *UMGS1071*. These taxa tended to be more abundant in control females than in control males, and decreased in MS males compared to control males. Meanwhile, no differences are detected in Female comparison.

- *Pseudoruminococcus*. This genus showed a more complex interaction pattern. It tends to be more abundant in control males compared to control females. However, in MS, its abundance slightly increases in MS females compared to control females, while it decreases in MS males relative to control males, indicating potential sex-specific response to MS.

Regarding the genera that were found to be more abundant in MS males compared to MS females:

- *Prevotella.* It showed an opposite pattern depending on sex. In males, its abundance tends to increase in MS compared to controls, while in females, it decreases in MS compared to controls. Additionally, there was a slight tendency to be more abundant in male controls than in female controls, suggesting a sex-disease interaction.

- *Catenibacterium.* It exhibited a sex-specific disease effect. No differences are observed between female controls and MS females, while in males, we identified a significant increase in MS compared to controls. This sex-specific response was reflected in its higher abundance in MS males compared to MS females.

For further exploration, we identified genera with consistent differential abundance patterns across the six individual studies (**Table 4.7**) and were compared with the meta-analysis results. Genera showing positive effect sizes in all six studies were also significant in the meta-analysis for all comparisons. Similarly, taxa showing negative effect sizes in all six studies were significant in the meta-analyses of Control and Female comparisons.

However, this was not observed in taxa showing negative effect sizes for Control and Male comparisons. All taxa except *Prevotella* showed a trend with a significant raw p-value in the meta-analyses results, but not reached significance when correcting by multiple testing. For the Control comparison the raw p-values were: *Limivivens* (p-value = 0.03), *Ruminococcus_C* (p-value = 0.12), and *Slackia* (p-value = 0.02). For the Male comparison: *Anaerostipes* (p-value = 0.009), *Brotaphodocola* (p-value = 0.02), *Limivivens* (p-value = 0.01), *Odoribacter* (p-value = 0.01), and *Ruminococcus_C* (p-value = 0.01).

Table 4.7. Taxa with consistent differential abundant patterns across individual studies. Taxa with positive effect sizes indicate higher abundance in the first group of the comparison, while those with negative effect sizes indicate higher abundance in the second group. Only taxa with consistent directionality across all six studies are included. Comparisons: Control (control females *vs.* control males), MS (MS females *vs.* MS males), Female (MS females *vs.* control females), and Male (MS males *vs.* control males). *MS: multiple sclerosis.*

Comparison	Consistent negative effect size	Consistent positive effect size
Control	*Limivivens, Ruminococcus_C, Slackia*	*Coprobacter*
MS	*Catenibacterium, Prevotella*	*Eggerthella, Flavonifractor, Holdemania*
Male	*Anaerostipes, Brotaphodocola, Limivivens, Odoribacter, Ruminococcus_C*	*Prevotella*
Female	*Coprococcus*	*Eggerthella, Flavonifractor*

4.4.6. COMPUTATIONAL VALIDATION

4.4.6.1. Variability characterization

To validate the findings from the meta-analysis, we analyzed PRJEB32762 dataset. This study was chosen due to its large sample size (N = 1,152) and the inclusion of participants from multiple geographical regions. The technical and biological characteristics of this dataset are summarized in **Table 4.8**.

Table 4.8. Summary of demographic, clinical, and technical characteristics of the PRJEB32762 validation dataset. (*Next page*) For categorical variables, the number of samples in each condition is indicated. For numerical variables, their mean value is determined. *BMI: body mass index; CNS: central nervous system; EDSS: Expanded Disability Status Scale; MS: multiple sclerosis; PPMS: primary progressive MS; RRMS: relapsing-remitting MS; SPMS: secondary progressive MS.*

Variable		Distribution	Variable		Distribution
Total samples		1,152	**Cohort**	1	128
Condition	Control	576		2	448
	MS	576	**Location**	San Francisco: 164	
Sex	Female	601		Boston: 42	
	Male	551		New York: 59	
Age (mean)		49.76		Pittsburgh: 12	
BMI (mean)		26.11		Buenos Aires: 129	
MS subtype	RRMS	437		Edinburgh: 131	
	PPMS	71		San Sebastián: 39	
	SPMS	68	**Treatment**	B-cell depleting: 28	
MS duration years (mean)		14.21		Classic immunomodulator: 241	
EDDS (mean)		2		CNS adhesion inhibitor: 27	
				Lymphocyte retention: 71	
				Untreated: 209	

Prior to performing the differential abundance analysis, we assessed the proportion of variance in the taxonomic structure that could be explained by technical or host-related biological variables, following the same approach used for the datasets incorporated in the meta-analysis. When considering all samples (**Figure 4.21-A**), all variables explained a significant fraction of the variance (p-value < 0.05). The two variables with the highest values were the technical factors *cohort* and *location* (sample collection site), accounting for 4.32% and 3.88% of the variance, respectively. Consequently, both were included in the multivariate model, as that they captured non-redundant variability. All host-related biological variables —*condition*, *BMI*, *age*, and *sex*— were also included in the multivariate model, together explaining 9.19% of the variance. The distribution of

samples based on the first principal coordinates colored by *cohort*, *location*, *condition* and *sex* is shown in **Figure 4.22**.

For the MS sample subset (**Figure 4.21-B**), all variables assessed were significantly associated with variance in the taxonomic structure (p-value < 0.05). The largest contributions were also *cohort* and *location* variables. In the cumulative multivariate model, these two variables, together with disease severity (*EDSS*), *sex*, and *BMI*, explained a total of 9.21% of the variance. In contrast, *treatment*, *age*, and *disease duration* did not contribute significantly.

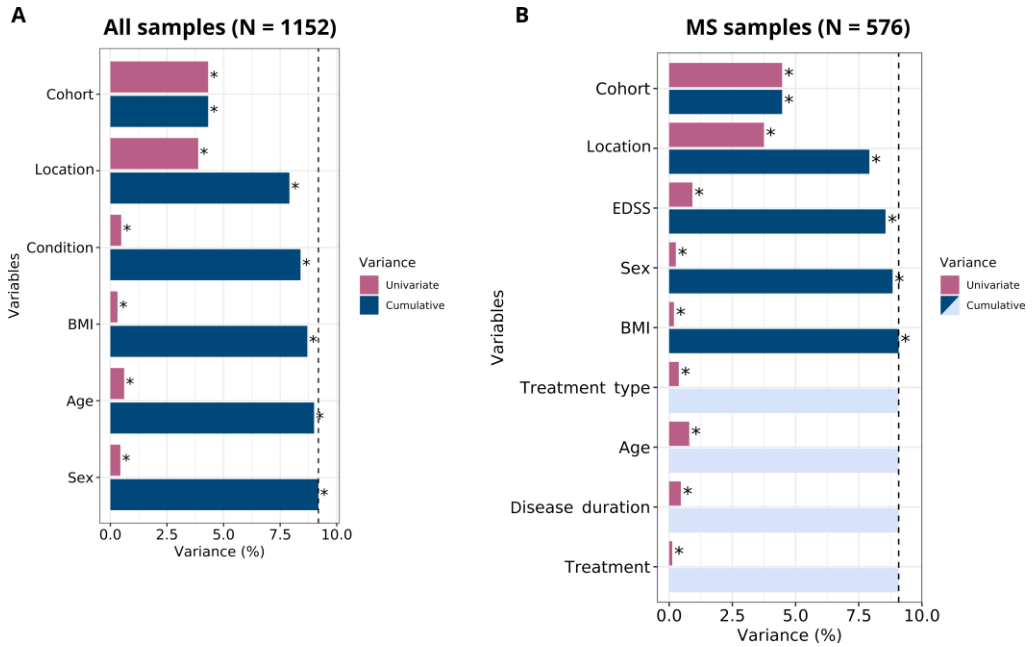


Figure 4.21. Variable contribution to microbiome compositional variation considering (A) all samples and (B) multiple sclerosis samples from the validation dataset. Results considering each variable independently (univariate variance) or in the multivariate model (cumulative variance). The asterisk (*) indicates if the contribution is significant; the black dashed line represents the cut-off for significant non-redundant contribution to the multivariate model considering p-value < 0.05. *BMI: body mass index; EDSS: Expanded Disability Status Scale; MS: multiple sclerosis.*

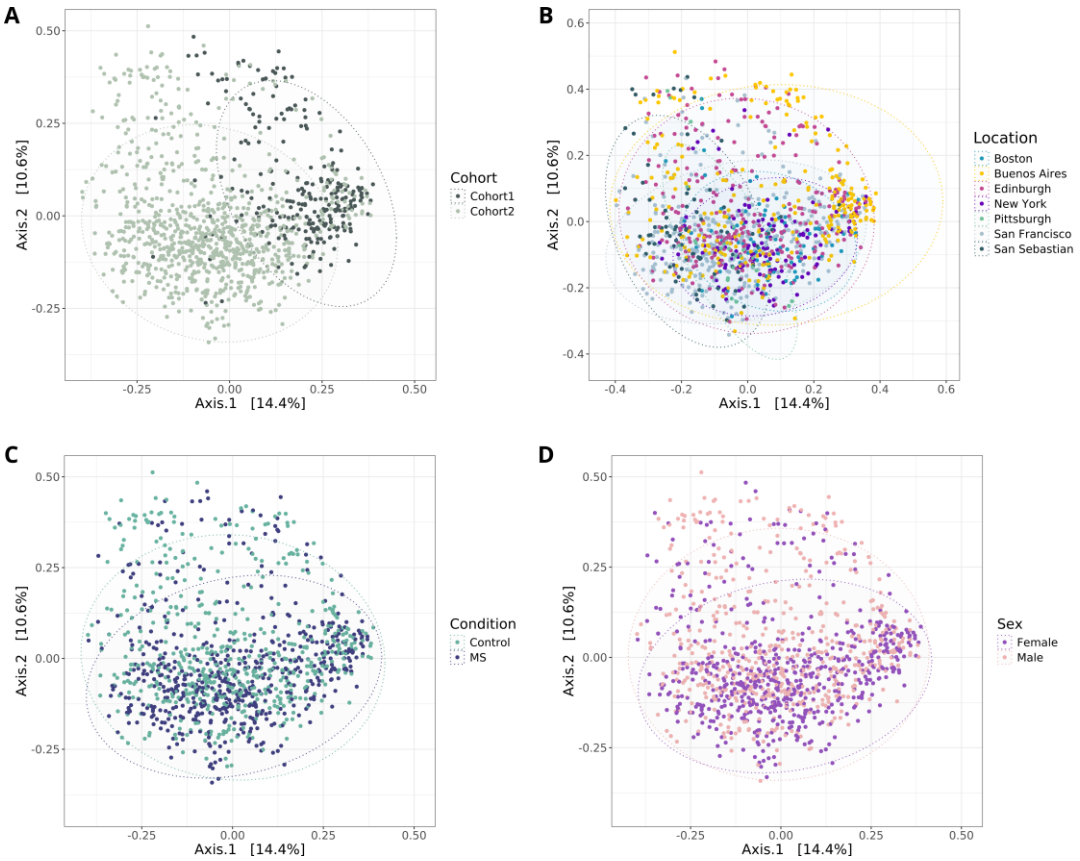


Figure 4.22. PCoA plots showing sample distribution for the validation dataset. Dotplot representation based on the first two principal coordinates of Bray-Curtis distances, with samples colored by (A) cohort (B) sample collection site (C) condition, and (D) sex. *MS: multiple sclerosis; PCoA: Principal Coordinates Analysis.*

This dataset was processed with the same workflow as for the meta-analysis datasets (including DADA2 pipeline, VST normalization, etc.). Considering the high proportion of variance explained by the technical factors cohort and location of sampling collection, we applied a blocked Wilcoxon rank-sum test for the differential abundance analysis, controlling for these variables to validate the meta-analysis significant results. Further details are provided in the *Materials and Methods* section.

4.4.6.2. Validated taxa

The effect size patterns for the significant taxa identified in the meta-analysis, together with the corresponding validation results across the four comparisons, can be visualized in **Figure 4.23**. Overall, we observed a consistent tendency for the direction of change, particularly in the Control, Female, and MS comparisons, with lower concordance in the Male comparison. In terms of statistical significance, 5 out of 12 genera (42%) were validated: *Coprococcus* in the Female comparison (more abundant in female controls than in MS females); *Eggerthella*, *Eisenbergiella*, and *Flavonifractor* in the MS comparison (more abundant in MS females than in MS males); and *Prevotella* in the MS comparison (more abundant in MS males than in MS females).

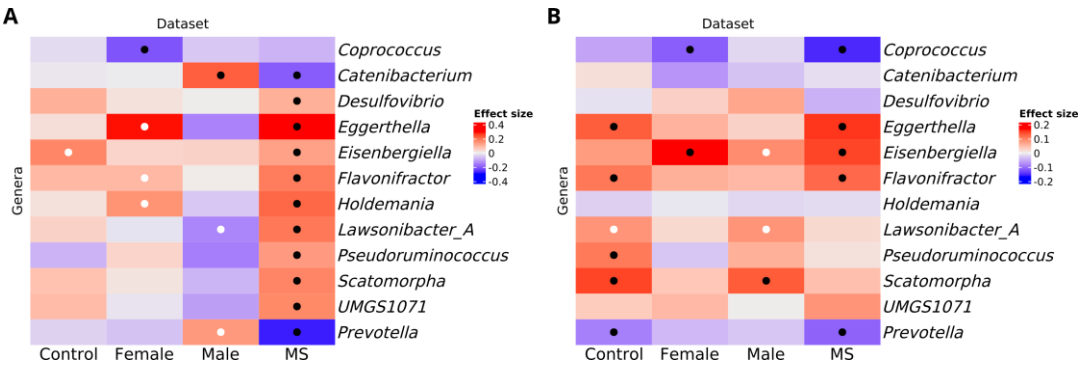


Figure 4.23. Effect sizes for the 12 genera with significant differential abundance in (A) the meta-analysis, compared with results from (B) the validation dataset, across the four comparisons. Rows represent the genera and columns represent comparisons. Color intensity indicates the magnitude of the effect size, and the color type its direction (red: higher in the first group of the comparison; blue: higher in the second). White dots mark genera with raw p-value < 0.05 and black dots mark genera with adjusted p-value < 0.05. Comparisons: Control (control females *vs.* control males), MS (MS females *vs.* MS males), Female (MS females *vs.* control females), and Male (MS males *vs.* control males). *MS: multiple sclerosis.*

4.4.7. Association with MS features

Finally, we explored whether the validated genera were associated with clinical characteristics of MS. Specifically, within the MS subset of the validation cohort (N = 576), we assessed potential associations between their normalized abundance and disease severity (EDSS), disease duration, and MS subtype (**Figures 4.24 and 4.25**).

As a result, we identified that *Eggerthella* was more abundant in the SPMS subtype, which corresponds to the advanced stage of MS (**Figure 4.24**). Additionally, this genus shows a positive association tendency with disease duration (**Figure 4.25-A**). We confirmed that this positive trend also correlates with the age of the individual ($\rho = 0.10$, p-value = 0.01), and that disease duration and age are moderately positively correlated in MS samples ($\rho = 0.57$, **Supplementary Figure 4.S34**). Interestingly, this positive association with age was not observed in the healthy control subset ($\rho = 0.04$, p-value = 0.33), suggesting that *Eggerthella* could be a potential marker to evaluate MS disease progression. Furthermore, *Eggerthella* association with disease duration is stronger in MS females (**Figure 4.25-B**) and not significant in MS males (**Figure 4.25-C**), pointing to a potential sex-specific modulation.

Conversely, the genus *Eisenbergiella* was found to be more abundant in progressive forms of multiple sclerosis (PPMS and SPMS) compared to RRMS (**Figure 4.24**). Additionally, *Eisenbergiella* abundance showed a positive association with both disease severity and disease duration (**Figure 4.25-A**). Notably, these associations persisted in MS females but were not significant in MS males, suggesting another potential sex-specific effect (**Figure 4.25-B-C**). A positive association with age was also observed ($\rho = 0.13$, p-value = 1.4×10^{-3}); however, unlike *Eggerthella*, this association was also present within the control population ($\rho = 0.14$, p-value = 8.3×10^{-4}), suggesting that age-related changes in *Eisenbergiella* abundance may not be exclusively linked to disease status.

As the abundance of *Eggerthella* and *Eisenbergiella* both tend to increase with disease duration, we confirmed their abundance also show a positive correlation ($\rho = 0.32$, p-value = 5.9×10^{-15}).

Finally, the analysis stratified by sex revealed a negative association tendency between the genus *Flavonifractor* and both disease severity and disease duration in MS males (**Figure 4.25-C**). This association was not observed when analyzing the entire MS cohort (**Figure 4.25-A**) or when examining MS females independently (**Figure 4.25-B**), suggesting that it may be specific to MS males.

None of the five validated genera patterns exhibited significant differences based on treatment status (treated versus untreated; **Supplementary Figure 4.S35**) or type of treatment (**Supplementary Figure 4.S36**); indicating that the observed associations are unlikely to be confounded by treatment effects.

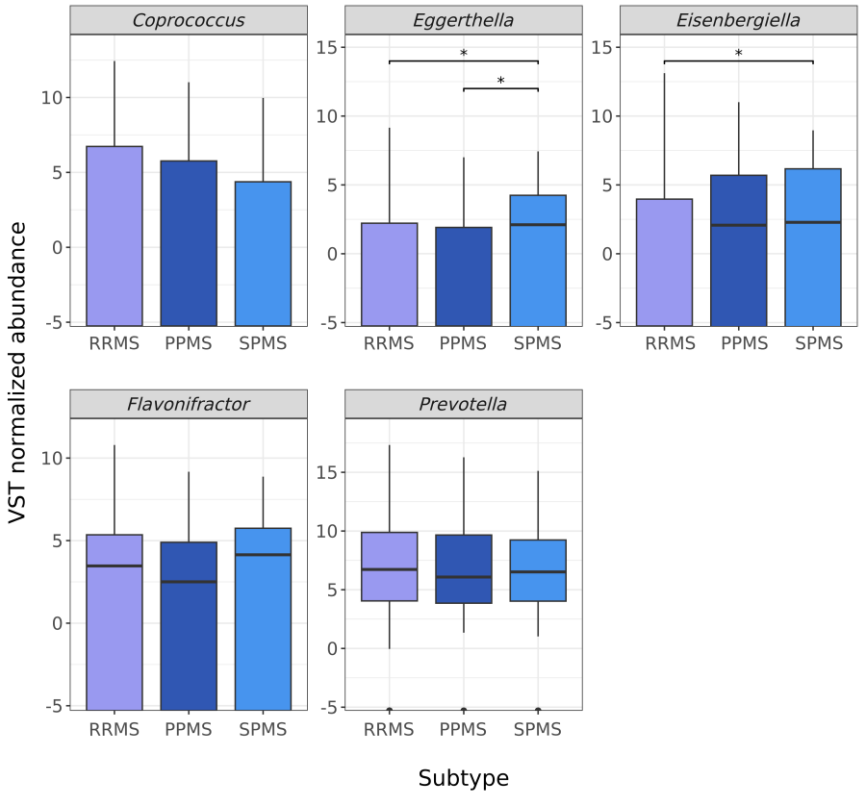


Figure 4.24. Normalized abundance of the validated genera across multiple sclerosis subtypes within the diseased cohort of the validation dataset. Asterisks (*) indicate pairwise significant differences with adjusted p-values < 0.05 from Dunn's post hoc test (after p-value < 0.05 in Kruskal–Wallis test). *PPMS: primary progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis; SPMS: secondary progressive multiple sclerosis; VST: variance stabilizing transformation.*

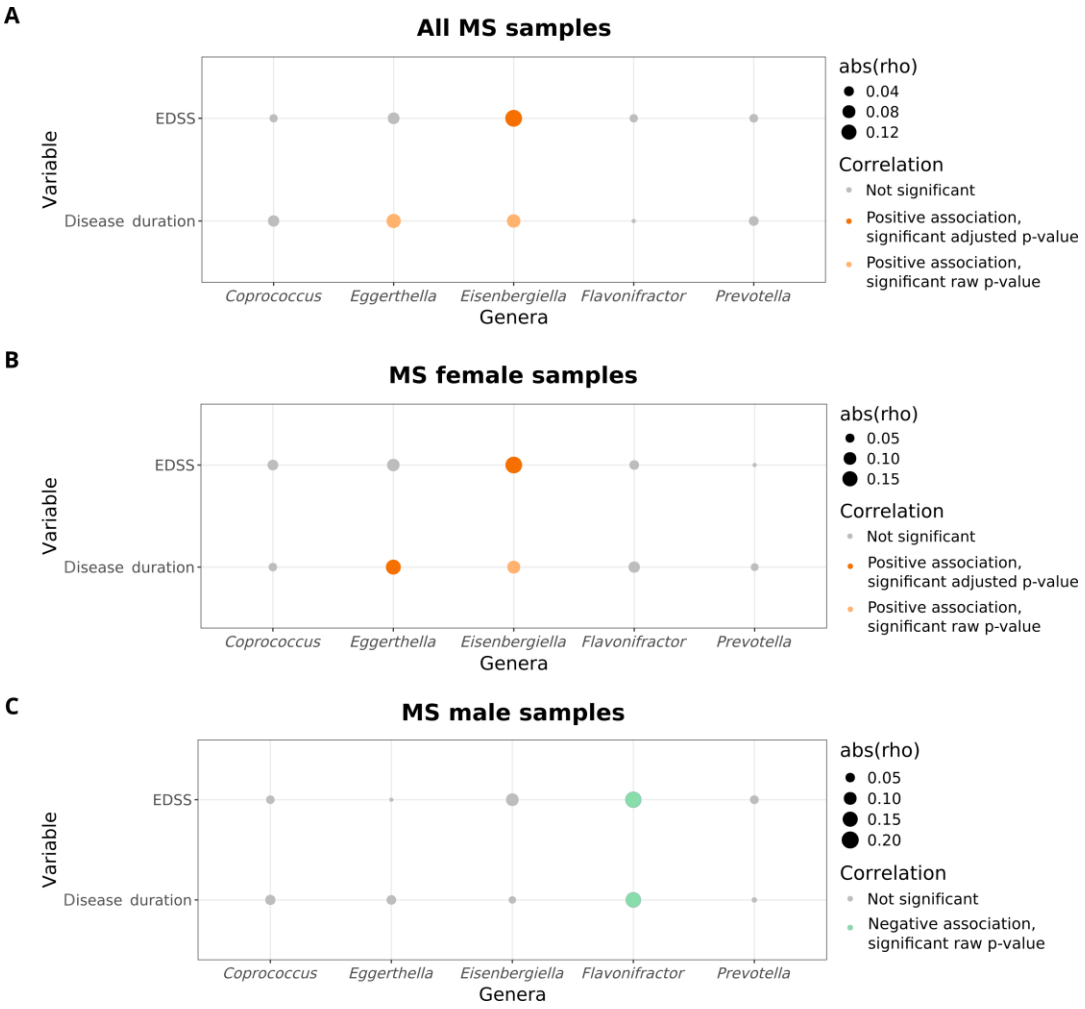


Figure 4.25. Associations between the abundance of validated genera and clinical variables in the multiple sclerosis cohort of the validation datasets. Dot colors indicate the statistical significance and the direction of change. Point size corresponds to the absolute value of ρ, reflecting the strength of the correlation. Panels represent (A) all MS samples, (B) female MS patients, and (C) male MS patients. Disease duration is measured in years. *EDSS: Expanded Disability Status Scale; MS: multiple sclerosis.*.

4.4.8. WEB PLATFORM

The web resource (https://irsoler.shinyapps.io/metaanalisis_16s_ms/) (**Figure 4.26**) includes detailed results of integrative analysis, offering free access to users to the complete findings. The interactive interface allows users to explore the results generated in the individual differential abundance analysis for each dataset and comparison. In this tab, the user can explore the resulting metrics for the genera of interest, as well as establishing filters based on desired cut-off points. The user can also explore the meta-analysis results. Selecting the genera and the comparison of interest, the forest, funnel, and influence plots are obtained, as well as a summary table with the corresponding results.

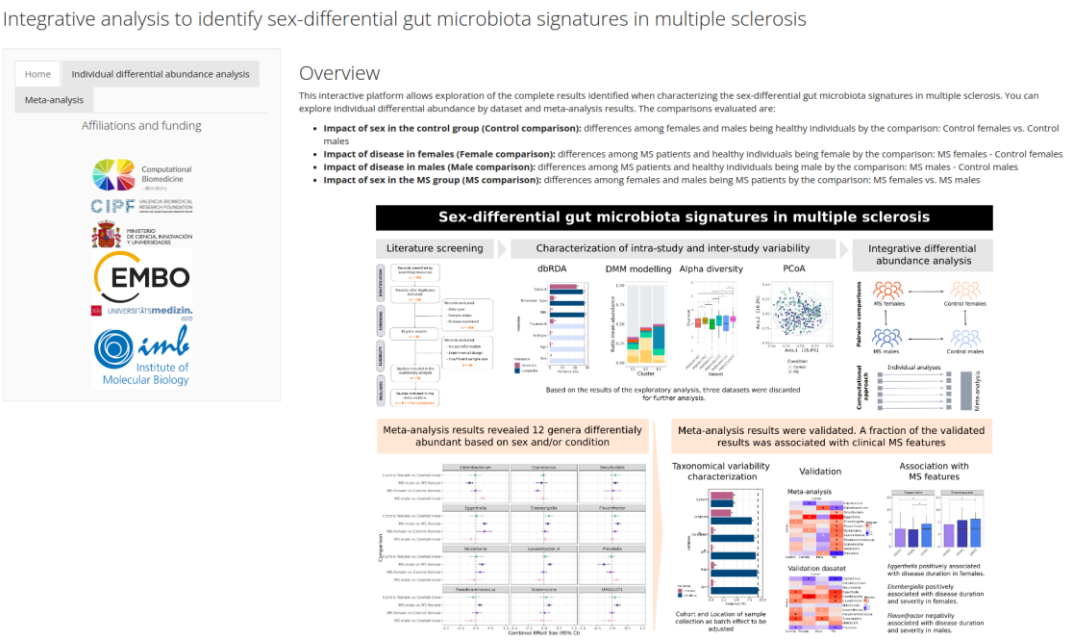


Figure 4.26. Home page of the interactive web tool. The left navigation bar provides access to the different modules of the platform: Home, Individual differential abundance analysis and Meta-analysis. The central section displays the framework of the tool. *MS: multiple sclerosis.*

4.5. DISCUSSION

The involvement of the gut microbiota in MS pathogenesis is well established, as the disease does not develop in model organisms that lack this microbial community[401]. Likewise, different studies analyzing human fecal samples have associated specific microbial taxa with MS, as well as with distinct clinical features like disease progression[402,403]. However, microbial composition is influenced not only by disease status but also by host-related factors. Among these, sex is a crucial variable to explore given its potential in modulating microbial communities, regulating immune system homeostasis, and influencing the progression of MS[368-369]. Taking these considerations into account, this thesis aimed to characterize sex-related differences in the taxonomic composition of the gut microbiota in MS.

Discordant findings are frequently reported across individual metagenomic studies, a scenario that was also evident in our work. From the nine studies selected through the systematic review, we observed that more than 20% of the variability in taxonomic composition was attributable to the dataset of origin, each of which included both MS and healthy individuals. Indeed, three studies were excluded from the analysis due to potentially large technical discrepancies that could substantially bias the microbial profiles. Several reports documented the high impact of the selected methodological approaches on microbiome research, which reduce the reproducibility and comparability potential across studies[398,474,475]. Beyond technical factors, this dataset-driven variability may also reflect the biological heterogeneity of individuals and their environments in each dataset. The Milieu Intérieur Consortium evaluated 1,000 healthy individuals across 110 demographic, clinical, and environmental variables. These variables all together explain a total of 16.4% of non-redundant variance in microbial composition. Interestingly, sex emerging was one of the strongest factors (each accounting for approximately 1% of the variance)[476]. Diet is another major driver of taxonomical variability, accounting for 5–20% of the variance depending on the cohort analyzed[329,477], partly through its impact on gut transit time and stool consistency[316,478]. Unfortunately, such associations could not be evaluated in our study because the available datasets lacked to report metadata variables such as stool moisture, cell counts, or host-related variables beyond condition, sex, BMI, treatment, and age. These observations highlight the need to retrieve as many variables as possible, including technical parameters, to better identify sources of heterogeneity in integrative microbiome analyses.

Considering the six studies ultimately integrated, we found that, in addition to the dataset of origin, disease status—but not sex—contributed significantly to overall variability in the taxonomical composition. However, when exploring the MS cohort, sex emerged as a significant factor, suggesting that distinct taxa may be differentially associated with MS depending on sex. Differences in alpha diversity were also primarily attributable to the dataset, with lower values observed in datasets sequenced with V4 and higher values with V3–V4, as expected given the higher resolution of the latter[479]. Meanwhile, no significant differences in alpha diversity were identified when exploring the four groups defined by condition (control or MS) and sex (female or male), in agreement with previous reports that did not report intraindividual differences between MS patients and controls[420,421].

Intra-study variability was reflected in the individual differential abundance analyses, where most taxa displayed discordant effect size trends across studies. Nevertheless, we were able to identify consistent microbial signatures across the different datasets. Specifically, the meta-analysis revealed two sex-specific associations: *Coprococcus*, whose abundance was reduced in MS females compared with female controls, and *Catenibacterium*, whose abundance was increased in MS males compared with male controls. Beyond these findings, the remaining significant associations were sex-differential patterns in MS, with nine genera more abundant in females and two genera more abundant in males. These results suggest that sex differences may be accentuated in the context of disease, since no significant results were observed in the comparison between female and male controls. However, several other taxa —primarily in the Control and Male comparisons— exhibited consistent trends across studies, reaching significance in the meta-analysis when not correcting the p-value for multiple testing. This may be due to the small effect sizes of the associations, which will require large sample sizes to be detected as statistically significant.

The dataset published by the iMSMS consortium was employed as an external validation cohort[409]. It encompasses geographically diverse populations and represents the dataset with the largest sample size included in the systematic review. The major sources of variability within this study were attributable to the cohort structure defined by the consortium and to the sample collection and processing location. Thus, these factors were accounted for in the differential abundance analyses. Overall, the majority of effect size patterns identified with the meta-analysis were reproducible in the validation cohort. Notably, 42% of the taxa considering all comparisons were validated after multiple-testing correction, including *Coprococcus* in the MS female versus female control comparison, and *Eggerthella*, *Eisenbergiella*, *Flavonifractor*, and *Prevotella* in the MS females *versus* MS males comparison.

Specifically, *Coprococcus* was identified with lower abundance in MS females compared with female controls. This taxon, a member of the *Lachnospiraceae* family within the *Firmicutes* phylum, is an important producer of the SCFA butyrate. Interestingly, previous studies in MS have also reported reduced abundance of *Coprococcus* compared to controls, without considering the sex of the individuals[418]. Beyond MS, decreased *Coprococcus* abundance has been linked to depressive symptoms[480], Parkinson's disease[481], systemic inflammation, and pancreatic cancer[482]. Overall, *Coprococcus* is a potential taxon for future research of sex differences in MS. It may be a potential taxa to explore why lower SCFA concentrations have been detected in MS females compared to MS males[427], and its depletion could also be involved in the exacerbated inflammatory patterns in MS females[225].

The three taxa validated with higher abundance in MS females than in MS males (*Eggerthella*, *Eisenbergiella* and *Flavonifractor*) were associated with clinical characteristics of MS, but each with a different pattern. None of these associations differed according to the treatment received by patients in the validation cohort. These genera may contribute to autoimmune processes that are exacerbated in females, as described in the following paragraphs, while their influence on the gut-brain axis remains undefined.

Eggerthella is a genus within the family *Eggerthellaceae* and the phylum *Actinobacteria*, and it is generally considered an opportunistic bacterium associated with proinflammatory environments. In our work, its abundance presents opposing trends in the sex-specific comparisons: increasing in MS females compared to control males while decreasing in MS males compared to control males. We also found that *Eggerthella* was also positively associated with disease duration and with the SPMS subtype, the progressive and advanced subtype of MS. This genus has been implicated in the production of proinflammatory compounds[483] and in autoimmune disorders for being involved in the activation of proinflammatory T lymphocytes[484,485].

By contrast, *Eisenbergiella* is a member of the family *Lachnospiraceae* within the phylum *Firmicutes*. The sex-related differences observed in MS patients appear to arise from pre-existing differences already present between males and females in the control group. *Eisenbergiella* was positively associated with disease duration and severity, as well as with progressive MS subtypes (PPMS and SPMS), suggesting its potential role in disease progression. Isolates of this genus have been reported to produce the SCFAs butyrate, lactate, acetate, and succinate[486]. Compared with *Eggerthella*, this taxon is less well described, but it has been positively associated with asthma[487] while showing a negative association with the inflammatory bowel disease[488]. Although further research is needed, in the context of MS our findings suggest that *Eisenbergiella* may contribute

to mechanisms that counteract inflammation and mitigate progression to more severe disease stages, including higher severity scores.

Interestingly, *Flavonifractor* displayed differential abundance patterns that appear to reflect the combined influence of both sex and disease status. Overall, it was not associated with MS clinical characteristics. However, when stratifying by sex, a negative trend emerged in MS males, where *Flavonifractor* abundance was inversely associated with disease severity and duration. Taxonomically, this genus belongs to the family *Oscillospiraceae* within the phylum *Firmicutes*. Functionally, *Flavonifractor* is involved in the metabolism of dietary flavonoids, compounds derived from fruits and vegetables[489]. Experimental studies in mice further suggests a protective role, as it has been shown to suppress autoimmune responses and promote anti-inflammatory activity[490,491]. Remarkably, its abundance has also been associated with improved responses to CAR-T therapy and with a lower incidence of severe adverse effects[492], reinforcing its potential contribution to immune regulation.

Finally, *Prevotella* was validated as the taxon with more abundance in MS males compared with MS females. This genus has been consistently linked to diets rich in fiber and complex carbohydrates[493], and has been identified as more prevalent in males than females[368]. Notably, *Prevotella* has also been identified as the most important sex-differential taxon in inflammatory bowel disease after adjusting for host variables, with higher abundance in males than in females[494]. In the meta-analysis of Qingqi Lin *et al.* *Prevotella* was reported to decrease in MS patients compared with controls in all studies, with statistically significant differences in more than half individual studies[420]. Importantly, this taxon has been identified to suppress Th17-mediated autoimmune responses and to reduce disease severity in the mouse model of MS[495]. However, it has been also associated with chronic inflammation[496] suggesting that its immunomodulatory role is context-dependent[496]. Although we did not observe associations with MS phenotypic characteristics in our work, *Prevotella* is a major producer of SCFAs[497]. Given its high prevalence in the gut microbiota, it could also contribute to the increased levels of SCFAs identified in males compared to females in fecal samples[427].

These four bacteria have also been associated with MS in individual studies, although without accounting for sex[406,409,413,498]. Thus, our study emphasized the need to explore sex as a biological factor in microbial studies. Moreover, it is equally important to consider the same representation of both sexes. An illustrative example are the associations with disease severity and duration of the validated taxa. They presented the same patterns when considering all samples together as when exploring MS females. However, results differed when exploring MS males. This is likely explained by the

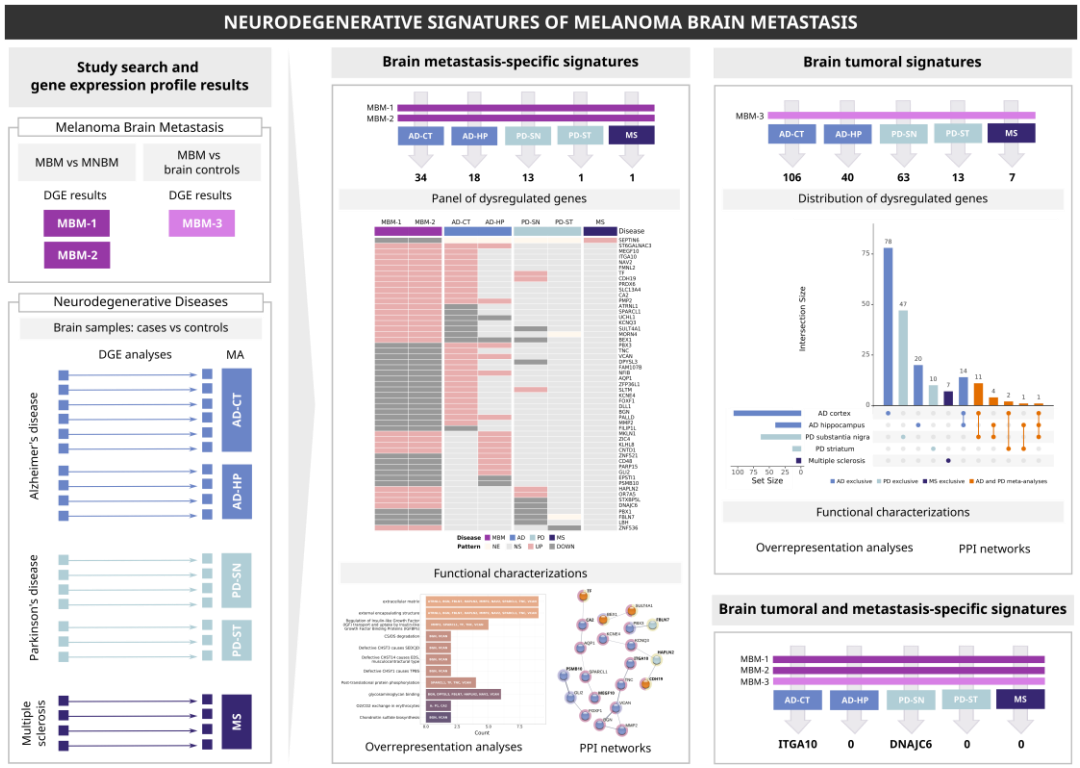
imbalance groups, which included 400 females and 176 males with MS. Consequently, the male subgroup was underrepresented, and some features that appeared to be shared by both sexes may in fact have been driven primarily by the female subcohort.

To the best of our knowledge, this is the first study performed in humans that explicitly evaluates sex and MS in the context of gut microbiome. Importantly, this analysis was not conducted in isolation but instead integrated multiple independent studies, rather than relying on findings from a single cohort. Furthermore, the main results were partially validated in an independent dataset, strengthening their robustness. Our work focused on associations that remained significant after adjustment for multiple testing in both the meta-analysis and the validation cohort. However, consistent trends across studies that did not reach statistical significance may also be of interest to be further explored. We also associated different taxa with clinical characteristics of MS (e.g., disease subtype, disability), contributing to a better understanding of how sex-specific microbiota signatures may relate to the disease course. Future studies are needed to confirm that this associations are not merely a dataset-driven effect.

The limitations of the work should be acknowledged. The high degree of heterogeneity across studies led us to exclude some datasets from the integrative analyses. Differences in sample size, sequencing platforms and targeted 16S regions, among others, are sources of variability that may restrict the comparability of results. In addition, the limited availability of metadata hindered the interpretation of this heterogeneity. Although unlikely, the presence of unreported comorbidities could also have confounded the observed associations. This study was focused on taxonomic composition, but future work using whole-genome shotgun metagenomics and metabolomics will be of great interest to identify functional pathways and mechanistic links between sex, microbiota, and MS.

Overall, despite the inherent challenges of microbiome research, this study provides a novel and integrative perspective of the interconnection between sex and MS in the gut microbiota. We identified accentuated sex-differential patterns in MS, potentially linked to autoimmune processes that may contribute to the exacerbated immune response observed in females compared with males. Within our results, taxa such as *Eggerthella*, more abundant in females, may promote inflammation. Meanwhile *Prevotella*, more abundant in males, could present protective effects. Ultimately, the sex-differential microbial contributions to MS pathogenesis appear to be multifaceted and context-dependent, and further research is needed to elucidate their implication in modulating the disease.

5. STUDY III: Unveiling common transcriptomic features between melanoma brain metastases and neurodegenerative diseases



This chapter was adapted from the manuscript *Unveiling common transcriptomic features between melanoma brain metastases and neurodegenerative diseases* by Irene Soler-Sáez, Alcida Karz, Marta R. Hidalgo, Borja Gómez-Cabañes, Adolfo López-Cerdán, José F. Català-Senent, Kylie Prutisto-Chang, Nicole M Eskow, Benjamin Izar, Torben Redmer, Swaminathan Kumar, Michael A Davies, María de la Iglesia-Vayá, Eva Hernando, and Francisco García-García. **Journal of Investigative Dermatology**, 2025. Volume 145, Issue 5, Pages 1135-1146. DOI: https://doi.org/10.1016/j.jid.2024.09.005.

5.1. INTRODUCTION

This introduction provides the foundational background to facilitate the understanding of the present study. First, the section *Neurodegenerative diseases share molecular mechanisms* summarizes the presence of common molecular patterns across neurodegenerative disorders, with a particular focus on MS, AD, and PD. Next, the section *The brain as a metastatic niche* focuses on cancer, with particular emphasis on brain metastasis. Finally, the section *Molecular features of melanoma brain metastasis* outlines the brain metastatic potential of melanoma, the cancer type under study in this doctoral thesis.

5.1.1. NEURODEGENERATIVE DISEASES SHARE MOLECULAR MECHANISMS

Neurodegeneration refers to the progressive loss of structure and/or function of neurons, ultimately leading to cell death. Consequently, neurodegeneration disrupts the integrity of neuronal circuits and compromises the proper function of both the brain and the spinal cord. This process is involved in a wide range of CNS disorders, being particularly detrimental given the limited regenerative capacity of the nervous tissue[499].

Neurodegeneration represents the principal hallmark of neurodegenerative diseases, which are defined by the gradual and irreversible decline of neuronal function. Despite their clinical and pathological heterogeneity, these conditions share several pathological features. David Wilson *et al.* 2023[91] recently summarized them across 14 different neurodegenerative diseases (**Figure 5.1**). They include:

- Pathological protein aggregation. The most well-known examples are Aβ plaques and NFTs in AD, and α-synuclein accumulation in PD.
- Synaptic and neuronal network dysfunction, including dendritic spine loss, impaired neurotransmission, and reduced synaptic plasticity.
- Aberrant proteostasis resulting from impaired autophagy-lysosomal and ubiquitin-proteasome pathways that fail to remove toxic aggregates and damaged organelles.
- Cytoskeletal abnormalities, leading to disrupted axonal transport and impaired neurotransmission.

- Altered energy homeostasis, reflecting the high metabolic demands of neurons, which are particularly vulnerable to mitochondrial dysfunction.
- DNA and RNA related defects, including mutations, impaired DNA replication, and altered RNA.
- Chronic neuroinflammation, driven by sustained activation of microglia and astrocytes, which release pro-inflammatory cytokines and reactive oxygen species that exacerbate neuronal injury.

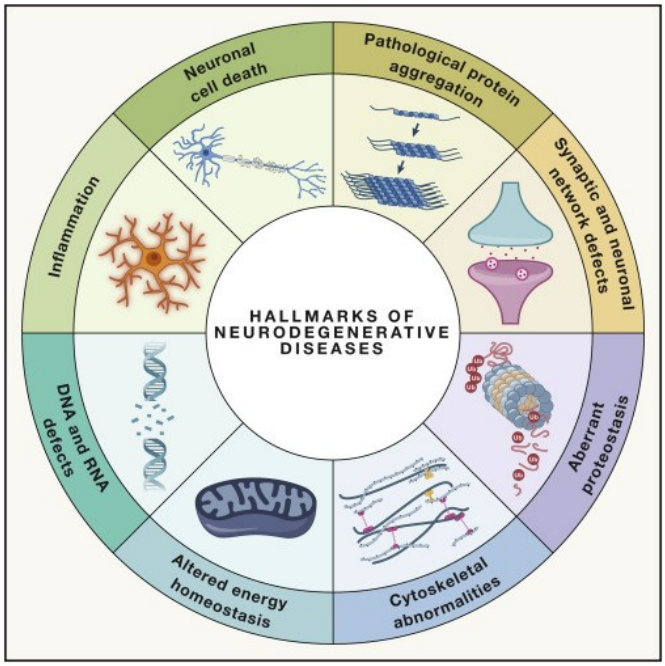


Figure 5.1. Hallmarks of neurodegenerative diseases. The categories illustrated summarize key pathological mechanisms recapitulated by David M. Wilson *et al.* 2023[91]. The color assignment and the ordination are arbitrary, established solely for visualization. Figure from David M. Wilson *et al.* 2023[91].

These hallmarks rarely act in isolation. Instead, they interact in self-reinforcing feedback loops. For example, mitochondrial dysfunction promotes oxidative stress, which exacerbates protein misfolding, while chronic neuroinflammation promotes both synaptic loss and neuronal death. Moreover, the contribution of these processes is not uniform across neurodegenerative disorders, as some mechanisms predominate more in some diseases than in others[91].

Despite the predominance of certain molecular mechanisms, the manifestation of distinct neurodegenerative diseases is partly explained by the concept of selective vulnerability, where specific neuronal populations are differentially susceptible to damage in each disorder. This selective vulnerability arises from multiple factors, including neuronal subtype, connectivity, metabolic demands, and the influence of the local microenvironment[500,501]. In this chapter, we focus on AD, PD, and MS as representative models of neurodegenerative pathology. Each disease is characterized by the principal affected regions: hippocampal and cortical neurons in AD[101], dopaminergic neurons of the substantia nigra in PD[105], and MS demyelinated regions in which the neuronal damage is reflected in the anatomical distribution of the lesions[21].

In AD, selective vulnerability is present in the sequential progression of affected brain regions. The earliest alterations emerge in the limbic system, primarily the hippocampus, with prominent involvement of pyramidal neurons. This early vulnerability reflects why episodic memory deficits represent the first clinical hallmark of the disease. As the pathology advances, changes extend to cortical regions, leading to impairments in language, executive functions, and visuospatial processing[101,502]. Meanwhile, the most affected neuronal population in PD is the dopaminergic neurons of the substantia nigra pars compacta. This neuronal loss leads to a reduction of dopamine release in the striatum, severely disrupting the nigrostriatal circuit that is involved in the motor symptoms characteristic of PD such as bradykinesia, rigidity, and tremor[503,504].

In MS the selective vulnerability takes a different form, primarily targeting myelinated axons and oligodendrocytes rather than specific neuronal subtypes. The clinical manifestations reflect the anatomical distribution of demyelinating lesions, with functional impairment depending on which region are the lesions located[21].

In line with the neurodegenerative hallmarks described above (**Figure 5.1**), several of these mechanisms can be illustrated with specific examples across AD, PD, and MS[91]. Synaptic dysfunction, is partly mediated by glutamate-driven neuronal hyperexcitability, as has been documented in AD[505], PD[506], and MS[507]. Oxidative stress represents another shared mechanism, as the accumulation of reactive oxygen species induces cellular damage[508]. Within this context, mitochondrial dysfunction exacerbates the production of reactive oxygen species while simultaneously altering the neuronal energy metabolism[508]. In addition, resident CNS glial cells such as astrocytes and microglia display dysregulated responses across the three disorders, including reduced metabolic neuronal support, non-adaptative stress responses, and chronic neuroinflammation[509,510].

Clinical evidence also suggests that neurodegenerative diseases frequently exhibit overlapping features, likely as a consequence of their shared molecular mechanisms[508].

While most patients are diagnosed with a unique neurodegenerative disorder, it is frequent to observe that they also present pathological changes that are representative of other conditions[511]. In severe scenarios, two neurodegenerative diseases may coexist within the same patient. Reported examples include the concomitant diagnosis of MS and PD[512], MS and AD[513], and AD and PD[514].

5.1.2. THE BRAIN AS A METASTATIC NICHE

As discussed in the *General Introduction* chapter, CNS disorders not only include neurodegenerative diseases but also other pathological conditions like cancer. Cancer comprises a group of diseases defined by uncontrolled cell proliferation, resistance to cell death, metabolic reprogramming, and the ability to invade surrounding tissues and disseminate to distant organs[515]. This dissemination capacity, known as metastasis, allows tumor cells to detach from the primary tumor, enter the bloodstream or the lymphatic vessels, and colonize anatomically distant tissues. Importantly, metastasis represents one of the leading causes of cancer-related mortality[516].

Among metastatic sites, brain metastases stand out due to their severity and high mortality. It is estimated that up to 40% of cancer patients in severe stages will eventually develop brain metastasis. This risk is strongly influenced by the tumor type, being lung, breast, and melanoma the cancers most prone to colonize the brain[517]. Brain metastasis typically manifest with neurological dysfunction, including motor, cognitive, and language impairments. Moreover, the prognosis remains extremely poor, with a median survival of less than six months after diagnosis[518].

To successfully metastasize the brain, tumoral cells must go through sequential events, as illustrated in **Figure 5.2**. Independent of their origin site, primary tumor cells acquire molecular and phenotypic changes that alter their cell-cell adhesions and extracellular matrix interactions, a process defined as the epithelial-to-mesenchymal transition. This transition enables tumor cells to detach from the primary tumor and migrate towards the blood and lymphatic vessels. First, they enter the circulation (i.e., intravasation). Once in the circulation, tumoral cells establish transient interactions with the vascular endothelium and, after successful adhesion to the vasculature, they exit the circulation (i.e., extravasation). In the specific case of brain metastases, the tumoral cells cross the blood-brain barrier. At this stage, tumoral cells find a microenvironment that differs substantially from the primary tumor. Therefore, the majority of cells die during the vascular circulation or shortly after colonizing the new tissue. However, a subset may survive by entering a dormancy state, characterized by a quiescent, rounded morphology, or alternatively, by initiating proliferation. In the latter scenario, additional

genetic and epigenetic adaptations allow these cells to colonize the brain microenvironment. As they grow and divide, the process culminates with the establishment of clinically detectable metastatic lesions[519].

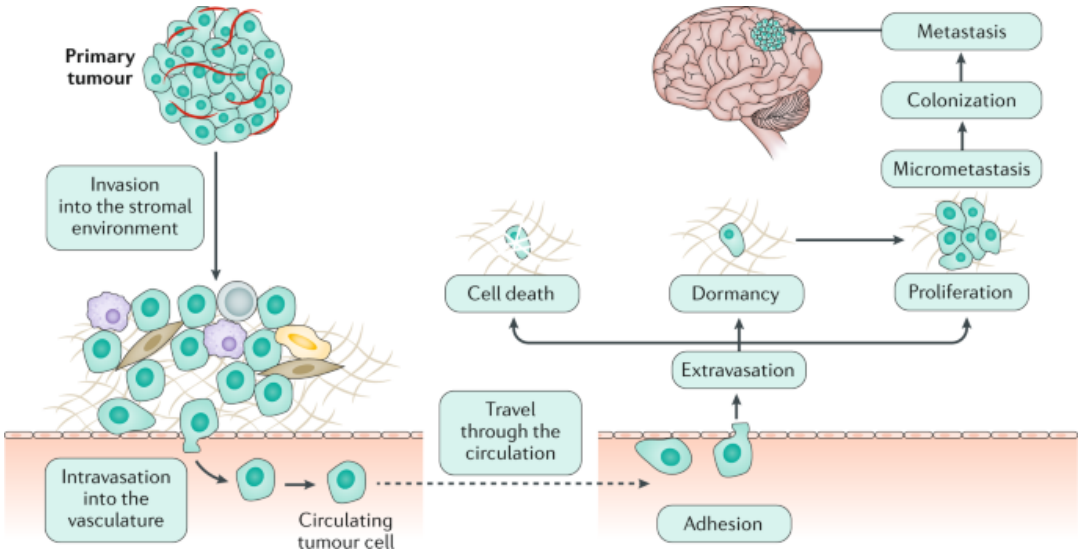


Figure 5.2. Illustration of the sequential steps for metastatic brain tumor formation. Tumor cells detach from neighboring cells and the extracellular matrix, intravasate into blood and lymphatic vessels, and circulate as individual tumoral cells. To establish brain metastases, cells must adhere to and cross the brain vasculature. Most cells die, some persist in a dormant (quiescent) state, and a minority proliferate to generate stable metastases. Turquoise cells represent tumor cells. Figure from Achal Singh Achrol *et al.* 2019[519].

Brain metastasis must be adapted to complex and context-specific interactions that differ from those observed in other metastatic locations. Different studies have identified specific molecular profiles of brain metastases compared to their matched primary tumors or to extracranial metastases. Giannoudis *et al.* 2024[520] performed genomic analyses on a large breast cancer cohort, reporting that brain metastatic cells harbor a higher prevalence of genetic alterations compared to both primary tumors and extracranial metastases. Similarly, Yoo *et al.* 2025[521] explored the transcriptomic differences between primary breast tumors and brain metastases. Among other findings, the study revealed the presence of neuronal-like gene expression patterns in brain metastases suggesting active tumor-neuron interactions[521].

In general terms, it is defined that even before metastatic colonization occurs in the brain, the systemic state of a patient with a primary tumor can influence the CNS environment. Primary tumors release exosomes, cytokines, growth factors, and other signaling molecules that alter the brain and may promote the metastatic establishment of the tumoral cells[522]. Once tumoral cells survive extravasation and reach the brain parenchyma, they activate processes to support their survival and growth, creating a supportive niche that offers physical anchorage, metabolic resources, and proliferative signals, while evading immune detection[523]. To achieve it, tumoral cells present high plasticity by their reprogramming capacity via genomic and epigenomic mechanisms, among others, that modulates both the tumoral cells themselves and the brain microenvironment[524].

Specifically, brain metastases promote vascular permeability, which ensures nutrient and oxygen supply to the tumoral cells[525]. Moreover, the extracellular matrix is altered. Changes in its stiffness and composition facilitate adhesion of metastasizing cancer cells, regulate their activation state, and contribute to immune evasion. However, the precise molecular mechanisms underlying these processes remain incompletely understood[526]. Brain metastatic cells also reprogram their metabolic profile. They increase the activity of pathways to metabolize glucose, fatty acids, and amino acids to survive, proliferate and to influence neighboring brain-resident cells in favor of tumor growth. Additionally, they also process neuronal metabolites, such as GABA and glutamate, which can be metabolized as alternative energy sources[527].

Notably, brain metastatic cells must interact with resident CNS cell types that are absent in other anatomical sites[528,529]. One strategy that tumor cells can adopt is the acquisition of neuronal-like patterns. Indeed, exposure to neurons has been shown to induce the expression of CNS-specific genes in cancer cells, a process considered crucial for their adaptation within the brain microenvironment[530]. Supporting this concept, the recent single-cell landscape presented by Xudong Xing *et al.* 2025[531] compared brain metastases with primary tumors from multiple anatomical origins and identified neuronal gene expression programs as a hallmark of brain metastases.

However, these interactions are highly context-dependent, partly by interactions with glial cells. Astrocytes are the cell type most extensively studied. Initially, this cell type presents anti-tumoral effects by releasing molecules that trigger apoptosis of invading tumoral cells. However, they are progressively reprogrammed by tumor-derived cytokines and adhesion molecules to acquire tumor-supportive roles. In this state, astrocytes can physically surround metastatic tumor, promote its survival, and suppress adaptive immune responses, which has led to their consideration as promising therapeutic targets[532–535]. The role of microglia is less well characterized, but current

evidence suggests that metastatic cells similarly reprogram microglia. Thus, this cell type shifts from pro-inflammatory and cytotoxic states towards immunosuppressive phenotypes that may limit the efficacy of immunotherapies[532,536,537]. Notably, peripheral immune cells are also recruited to the metastatic niche, where they also contribute to an immunosuppressive state[538].

Overall, the development of brain metastases is a multistep process shaped by tumor cell dissemination, adaptation, and survival within the highly specialized CNS microenvironment. These processes are particularly relevant in melanoma, a tumor type with high propensity for colonizing the brain with clinically aggressive metastatic behavior.

5.1.3. MOLECULAR FEATURES OF MELANOMA BRAIN METASTASIS

Melanoma is one of the most aggressive and lethal tumor types. Its aggressiveness is mostly driven by its high metastatic potential, which accounts for approximately 90% of melanoma-related deaths. Among the most frequent metastatic sites we found the lymph nodes and visceral organs, particularly the lung, liver, and brain[539,540].

Brain metastasis represents one of the most frequent and severe complications of melanoma, corresponding for nearly 10% of all metastatic brain tumors[541]. When melanoma is diagnosed, nearly 20% of patients already present brain metastases, and more than 50% will eventually develop them during the course of the disease, as confirmed by *postmortem* brain analyses[542,543]. The median overall survival of patients with melanoma brain metastases (MBM) was established approximately four months after diagnosis, although this has currently improved with advances in diagnostic imaging and immunotherapy[544].

Understanding how melanoma metastasizes the brain remains a major challenge for the scientific community. Melanoma is characterized by a high mutational burden, probably driven by chronic exposure to ultraviolet radiation, which promotes the acquisition of metastatic traits. Among the most recurrent alterations are mutations in BRAF and NRAS, as well as loss of function of PTEN. However, these mutations alone are insufficient for the development of MBM, reflecting the complexity of the biological and microenvironmental factors that drive this process[545,546].

To better understand the heterogeneity of melanoma, recent studies have undertaken the molecular characterization of MBM[547–554]. Collectively, these studies revealed that MBM is specifically characterized by its metabolic adaptations, immune evasion strategies, and the acquisition of neuronal-like transcriptional programs, among others.

One of the most consistent features identified across independent studies of MBM is the upregulation of oxidative phosphorylation. Tumoral cells typically depend on aerobic glycolysis, known as the Warburg effect. Through this process, they proliferate by fermenting glucose into lactate even in the presence of oxygen. Although this pathway produces less ATP per glucose unit compared to the oxidative phosphorylation, it accelerates the glucose uptake and enables the generation of glycolytic intermediates that sustain the biosynthetic demands of rapidly dividing cancer cells. Strikingly, MBM appear to enhance oxidative phosphorylation compared to extracranial metastases[548,550,551,555], although this feature may be restricted to specific subpopulations of MBM[556]. The upregulation of the oxidative phosphorylation have been associated with improved metastatic spread[550] and the resistance to therapies[557].

Another aspect of MBM is the ability of tumor cells to establish an immunosuppressive microenvironment that prevents their elimination by the host immune system. Kumar *et al.* 2025[547] have shown that MBM with higher levels of immune infiltration tends to respond more favorably to immunotherapy and exhibit reduced therapeutic resistance. However, MBM cells can evade the adaptive immune surveillance through several mechanisms, including the overexpression of coinhibitory immune molecules and the recruitment of regulatory T cells, which suppress cytotoxic T lymphocyte activity[549]. These findings highlight the relevance of immunotherapy strategies aimed at reactivating antitumor immune responses, enabling effective recognition and destruction of cancer cells within the brain[553].

Beyond immune evasion, MBM cells must be adapted to the specific characteristics of the brain microenvironment. This capacity may be partly explained by the developmental origin of melanocytes, which arise from multipotent neural crest cells during neural tube formation[558]. MBM employs neuronal mimicry strategies to adopt transcriptional programs that resemble those of neurons. Unlike extracranial metastases, MBM upregulates the expression of different genes involved in neuronal development, differentiation, and synaptic function[548]. Among these, the nerve growth factor receptor NGFR is one of the most studied. This gene, primarily involved in neuronal development, contributes to melanoma tumoral traits such as survival, migration, stemness, and therapy resistance, while promoting metastasis and cellular plasticity[559,560]. Indeed, recent studies have reported that NGFR-expressing subpopulations in MBM are associated with an invasive, stem-like, and drug-resistant phenotype[560]. Despite these advances, the precise mechanisms by which MBM adopt neuronal-like programs remain poorly understood, and further research is needed to clarify their role in tumor progression and therapeutic resistance.

5.2. CONTEXTUALIZATION, MOTIVATION AND OBJECTIVES

Cancer and neurodegenerative diseases are often perceived as antagonistic biological processes: tumoral cells emerge from the evasion of cell death mechanisms to sustain uncontrolled survival, whereas neurodegeneration is characterized by increased neuronal loss. When we explore their underlying molecular pathways, they often converge on similar cellular processes as DNA damage response, oxidative stress regulation, and inflammation; all of which can be dysregulated in the same or opposite directions[561]. Different studies have identified inverse correlation between neurodegenerative diseases and cancer, supporting the idea that one condition may protect against the other[562,563]. However, recent findings reveal a more complex relationship, where certain cancers exhibit co-occurrence with neurodegenerative disorders. For instance, AD has been linked to an increased risk of colon cancer, possibly due to the accumulation of Aβ peptides in enteric neurons, which could compromise the intestinal barrier integrity[564].

Melanoma stands out as a cancer type with potential molecular associations with neurodegenerative diseases, due to its high adaptative capacity to the brain environment[539]. In the following paragraphs, we summarize the antecedents linking MBM to the neurodegenerative disorders MS, AD and PD.

MS and melanoma. A population-based cohort study conducted in the Netherlands[565] reported a higher incidence of melanoma among individuals with MS compared to a reference population without the disease. Regardless, the most characterized association between MS and melanoma has been established through treatments. Some DMTs prescribed to MS patients are immunosuppressive and may promote tumor development[566]. Indeed, DMTs may facilitate melanoma progression within its microenvironment[567] and have been associated with the development of metastatic melanoma[568]. Conversely, melanoma treatments such as BRAF/MEK inhibitors can trigger autoimmune responses, potentially exacerbating MS symptomatology[569].

AD and melanoma. In the study conducted by Kevin Kleffman *et al.* 2022[143], the authors identified that brain melanoma metastatic cells, compared to those from extracranial metastases, express proteins implicated in neurodegenerative pathologies. One example of affected processes is oxidative phosphorylation. The results also showed that MBM secretes the AD-associated peptide Aβ to suppress local inflammation in the brain microenvironment. This secretion promotes interaction with astrocytes, favoring tumoral cell survival.

PD and melanoma. A positive correlation between melanoma incidence and PD has been repeatedly observed[570–572], with both melanoma being associated with an increased risk of PD and vice versa[573]. Recently, the work from Arnold *et al.* 2025[574] demonstrated that loss of function of the PD-associated protein α-synuclein in melanoma cells leads to the impairment of DNA repair process, favoring tumor development. However, the direction of this association remains unclear, as other studies have proposed that melanoma cells in mice overexpressing α-synuclein exhibit increased metastatic potential[575]. In humans, α-synuclein is positively correlated with tumor growth and poorer survival in melanoma patients[576].

Collectively, altered molecular mechanisms in MS have been directly implicated in other neurodegenerative diseases like AD and PD, and may be potential targets to explore the adaptation of metastatic melanoma to the brain microenvironment[577,578].

To the best of our knowledge, despite the previously described associations, an explicit, unbiased comparison between melanoma gene signatures and those of neurodegenerative diseases has not yet been conducted. Thus, the major objective of this chapter is to identify common transcriptomic patterns of metastatic melanoma adaptation to the brain by evaluating the relationship between MBM gene signatures and those of AD, PD and MS; which may expand our understanding of the biology of these tumors. To achieve it, the following specific objectives were established:

1. To identify peer-reviewed manuscripts analyzing the transcriptomic profile of MBM, both in comparison with extracranial melanoma metastases and with non-tumor-bearing brain controls. To select studies of interest according to defined inclusion and exclusion criteria.

2. To perform individual differential expression analyses comparing neurodegenerative disease cases against controls, based on the processed transcriptomic data previously published for MS[79], AD[579], and PD[580]. Then, to integrate the individual results into meta-analyses stratified by disease and affected brain region, obtaining consensus expression signatures.

3. To identify and functionally characterize transcriptomic patterns shared between MBM when it is compared with extracranial melanoma metastases, in relation to the consensus profiles of each neurodegenerative disease.

4. To identify and functionally characterize transcriptomic patterns shared between MBM when it is compared with non-tumor-bearing brain controls, in relation to the consensus profiles of each neurodegenerative disease.

5. To develop an open-access, user-friendly web resource for dissemination of the complete results.

5.3. Materials and Methods

5.3.1. Workflow description

We conducted a systematic screening of transcriptomic profiles already published in the literature, both comparing MBM with extracranial metastases and normal brain samples. From the selected publications, we retrieved the individual results from the differential gene expression analysis. In parallel, we performed meta-analyses of transcriptomic datasets from MS, AD, and PD, previously processed in our laboratory. Specifically, we compared diseased *versus* healthy samples individually for each dataset, and then generated consensus molecular signatures for each neurodegenerative disease and brain region. The resulting profiles were compared with those of MBM to identify shared altered genes, as defined in **Figure 5.3.** They were subsequently functionally characterized to elucidate the potential processes underlying both neurodegeneration and MBM.

All bioinformatics analyses were conducted with version 4.1.2. of the R programming language (R Core Team 2021). The packages used and their corresponding versions can be consulted in the **Supplementary Table 5.S1**. The code generated in this work is publicly available in a GitHub repository https://github.com/IrSoler/cbl-neurodegeneratives-mbm.

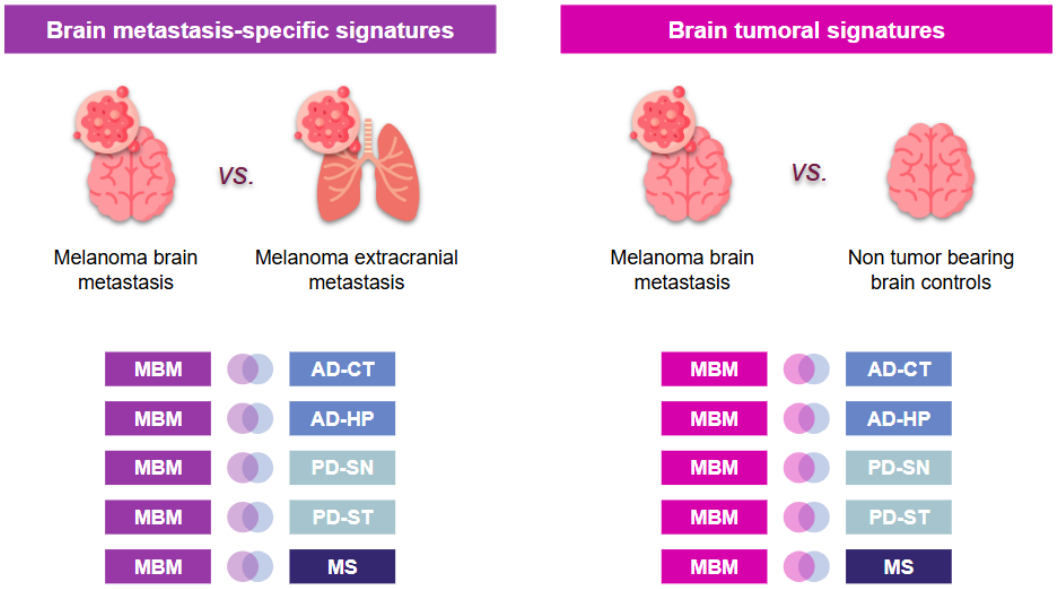


Figure 5.3. Illustration of design of this research. Left panel shows the melanoma brain-specific tumor signature, obtained by comparing MBM with extracranial melanoma metastases. Right panel represents the brain tumor signature, derived from the comparison of MBM with non–tumor-bearing brain controls. For each MBM signature, intersections with the consensus profiles of the five neurodegenerative disease meta-analyses are illustrated: AD-CT (Alzheimer's disease, cortex), AD-HP (Alzheimer's disease, hippocampus), PD-SN (Parkinson's disease, substantia nigra), PD-ST (Parkinson's disease, striatum), and MS (multiple sclerosis). *MBM: melanoma brain metastasis.* Icons downloaded from *Flaction.*

5.3.2. Data collection

The literature screening was conducted up to December 2022. Regardless the disease addressed, all selected studies met the following inclusion criteria: i) type of data: transcriptomic data, ii) organism: human (*Homo sapiens*), iii) type of biological sample processing: tissue sections (not cell lines). Additional inclusion specifications were applied depending on the disease type.

- MBM retrieved data comprise differential gene expression results from MBM versus melanoma extracranial metastasis, or comparing MBM with non-tumor bearing brain controls.

- For neurodegenerative diseases, the systematic review and dataset selection have already been published: MS in Català-Senent *et al.* 2023[79], AD in López-Cerdán *et al.* 2024[579] and PD in López-Cerdán *et al.* 2022[580]. In these works, different microarray and RNA-seq datasets were processed. Exploratory analysis included PCA, clustering, and gene expression distribution by sample. Samples where the presence of an anomalous pattern was confirmed were excluded from further analyses.

To explore whether the observed transcriptomic signals are exclusive to melanoma brain metastases or instead represent common features of brain metastases regardless of tumor origin, we also included two publicly available datasets of breast cancer brain metastases (BBM). Specifically, BBM-1 results were obtained by performing a differential expression analysis on the RNA-seq dataset reported in Cosgrove *et al.* 2022[581], while BBM-2 corresponded to the differential expression results published in Varešlija *et al.* 2019[582]. Both datasets contained brain metastases derived from breast cancer and extracranial breast cancer metastases, thereby enabling direct evaluation of brain-specific metastatic signatures reported in the *Results* section of this doctoral thesis.

5.3.3. DIFFERENTIAL GENE EXPRESSION AND META-ANALYSES IN NEURODEGENERATIVE DISEASES STUDIES

Differential expression analyses were conducted independently for each selected study of the three neurodegenerative diseases. First, the individual raw count matrices were normalized. For microarray datasets, normalization followed the manufacturer's recommended protocols for the specific commercial platform. For RNA-seq datasets, normalization was performed using the trimmed mean of M values method[583].

For the identification of differentially expressed genes, case *versus* control comparisons were conducted using the linear regression model implemented in the limma R package[584]. To ensure comparability between data types, the *voom* transformation was applied to RNA-seq datasets. This step stabilizes the mean-variance relationship and approximates the data distribution to the microarray structure. Finally, the standard limma pipeline was applied to both the transformed RNA-seq data and the microarray datasets[584].

Potential biological and technical confounders (such as age or batch effects) were incorporated as covariates into the regression models to minimize their impact on the results. The specific covariates considered varied depending on the availability of metadata reported in each study, and were therefore dependent on the information provided in the original datasets. To correct for multiple testing, p-values were adjusted using the BH method[250], with a FDR < 0.05 considered statistically significant. For each gene, the logFC was calculated to determine both the direction and magnitude of the expression change. In this context, positive logFC values correspond to genes upregulated in cases (and downregulated in controls), whereas negative logFC values correspond to genes upregulated in controls (and downregulated in cases).

Differential expression results from individual studies were integrated into five independent meta-analyses, according to the neurodegenerative disease and the brain region examined: Alzheimer's disease – cortex, Alzheimer's disease – hippocampus, Parkinson's disease – substantia nigra, Parkinson's disease – striatum, and multiple sclerosis – whole brain. To account for variability across studies and sample sources, we applied the DL random-effects model[464]. This approach allowed us to capture both the within-study and between-study heterogeneity, thus providing more robust estimates of effect size. Further details can be consulted in the *Study II* (*chapter 4, section 4.3.14*) of this doctoral thesis. Meta-analysis statistics were computed for each gene, and the resulting p-values were adjusted using the BH method[250], considering FDR < 0.05 as the threshold for statistical significance.

5.3.4. INTERSECTION ANALYSIS

For both MBM and each neurodegenerative disease meta-analysis, significant genes were defined by applying a threshold of FDR < 0.05 to the resulting adjusted p-value. Genes were then classified according to the direction of change in their respective comparisons: upregulated in MBM (logFC > 0) or downregulated in MBM (logFC < 0). The same classification was established for the neurodegenerative diseases: upregulated in disease (logFC > 0) or downregulated in disease (logFC < 0).

To identify potential shared genes, we performed intersection analyses between the sets of significant genes from MBM and those obtained for each neurodegenerative disease. Then, the direction of change in both types of diseases was evaluated. To visualize the results, we elaborated bar plots with the ggplot2[585] R package, upset plots with the UpSetR[586] R package, and heatmaps with the ComplexHeatmap[587] R package.

In the specific case of MBM, two independent studies compared MBM with melanoma extracranial metastases (named MBM-1 and MBM-2 in the *Results* and *Discussion* sections of this chapter). To avoid redundant or inconsistent signals, only those genes that displayed a significant concordant regulation pattern in both MBM studies (i.e., upregulated in both or downregulated in both) were retained for further analysis.

5.3.5. RESAMPLING

To evaluate whether the overlaps detected in the intersection analysis reflected true biological signals rather than random coincidences, we performed a resampling-based statistical approach. For each pairwise comparison between MBM and a given neurodegenerative disease, we first selected the gen sets assessed in both conditions. Then, we generated 10.000 independent iterations. In each iteration:

1. A random subset of genes was selected from each study, with the subset size matching the number of genes identified in the actual intersection analysis.
2. Within this randomly selected gene subset, we identify how many genes were significant in the intersection results.

Then, the distribution of simulated overlaps was obtained, and the median value across the 10,000 simulations was calculated.

5.3.6. FUNCTIONAL SIGNATURES OF THE COMMON TRANSCRIPTOMIC FEATURES

To gain biological insight into the shared transcriptomic alterations identified between MBM and neurodegenerative diseases, we performed a functional characterization of the resulting gene sets. The methodological principles are the same as those outlined in *Study I* of this doctoral thesis: ORA and PPI networks analyses (*chapter 3 sections 3.3.11.3, 3.3.11.4, respectively*). The details of the methodological strategies can be consulted on that chapter, while the particularities of this study are described below.

ORA analyses were conducted with the clusterProfiler R package[588]. To explore the biological relationships, functional annotations were obtained from multiple curated databases accessible through the AnnotationDbi R package[589]: i) Reactome pathways, ii) KEGG pathways, iii) GO Biological Processes, iv) GO Molecular Functions, and v) GO Cellular Components (the last three being from the *Gene Ontology* database). For each functional enrichment test, gene sets were filtered to include only those ranging

from 10 to 500 genes, excluding results that were either too specific or too general. The resulting p-values were adjusted by the BH method[250], considering statistical significance when FDR < 0.05.

Meanwhile, PPI analyses were conducted using the STRING R package and the STRING web tool[195]. Interaction networks were generated using default parameters, where we evaluated both functional and physical protein associations. Significant networks were considered when PPI enrichment p-value < 0.05. For visualization, isolated nodes were excluded, and the confidence score was used to represent the strength of the protein-protein interactions.

5.3.7. WEB TOOL

To enhance accessibility and promote reproducibility, we developed the web tool https://bioinfo.cipf.es/metafun-mbm/ as an open and interactive resource for the scientific community. The platform was developed with the Quarto system, providing a user-friendly and interactive environment to navigate through the web. The tool is organized into seven sections: i) overview of the pipeline, ii) MBM expression profiles, iii) neurodegenerative diseases expression profiles, iv) neurodegenerative signature and functional profiling of brain metastasis-specific signatures (MBM-1-2 results), v) neurodegenerative signature and functional profiling of MBM tumor profile (MBM-3 results), and vi) methods description.

5.4. RESULTS

This section outlines the findings that enable the characterization of the relationship between neurodegenerative gene signatures and those characteristic of MBM. To address this objective, we performed an *in silico* strategy. First, we identified previously published transcriptomic profiles of MBM, and brain expression datasets from patients with AD, PD, and MS. We then performed differential expression analyses for individual datasets, and meta-analyses for each neurodegenerative disease and its corresponding affected brain region. In these analyses we compared diseased individuals to their respective controls to obtain disease-specific transcriptomic profiles. From this, we identified genes dysregulated in both neurodegenerative diseases and MBM, and the corresponding pathways and biological functions enriched. All results are available as an open resource on the web tool (https://bioinfo.cipf.es/metafun-mbm/).

5.4.1. DATA COLLECTION AND DIFFERENTIAL GENE EXPRESSION ANALYSES BY NEURODEGENERATIVE DISEASE

Literature screening resulted in the identification of publicly available transcriptomic studies for the diseases of interest. After applying the defined inclusion and exclusion criteria we selected 3 MBM datasets, hereafter referred to as MBM-1, MBM-2, and MBM-3 (**Figure 5.4, left panel**). From them we retrieved the differential gene expression results, whose complete statistics can be explored in the *Melanoma Brain Metastasis* section of the web tool. The first two studies compared the transcriptomic profiles of MBM and extracranial metastases, named in this work as MBM-1 and MBM-2. In contrast, MBM-3 compared gene expression profiles of MBM versus normal brain tissue. Further details are presented in **Table 5.1**.

For ease of reading, throughout this work we will use the term melanoma brain-specific metastatic signature to refer to the comparison of MBM versus extracranial melanoma metastases (MBM-1 and MBM-2). Conversely, the term melanoma brain tumoral signature denotes the MBM-3 comparison, i.e., MBM versus normal brain tissue.

Next, we analyzed data from our previously published meta-analyses of 10 datasets for AD, 6 datasets for PD, and 4 datasets for MS (AD in López-Cerdán *et al.* 2024[579], PD in López-Cerdán *et al.* 2022[580] and MS in Català-Senent *et al.* 2023[79]). A summary of the three systematic reviews is provided in **Figure 5.4, right panel**. As a first step, differential gene expression analyses were performed for each individual dataset, consistently comparing cases (AD, PD, or MS samples) *versus* controls (healthy samples). The number of significantly differentially expressed genes for each study is provided in **Supplementary Table 5.S2**. Then, the individual results were integrated according to the disease and the brain region under study, resulting in five meta-analyses: AD cortex (AD-CT), AD hippocampus (AD-HP), PD substantia nigra (PD-SN), PD striatum (PD-ST), and demyelinated lesions from MS (MS). These meta-analyses contain the transcriptomic alterations consistently observed in each brain region by integrating results across studies. **Table 5.2** summarizes the technical characteristics and the results of each meta-analysis.

Statistical metrics by gene and dataset can be consulted in the neurodegenerative diseases sections of the web tool, subsections *Individual analysis of selected studies*. For the meta-analyses, the web tool subsection *Meta-analysis results* reports the statistical metrics obtained for each gene, i.e. p-value, adjusted p-value, logFC with upper and lower confidence intervals, QE, QEp, SE, tau2, I2, H2 and the number of studies where the gene was evaluated.

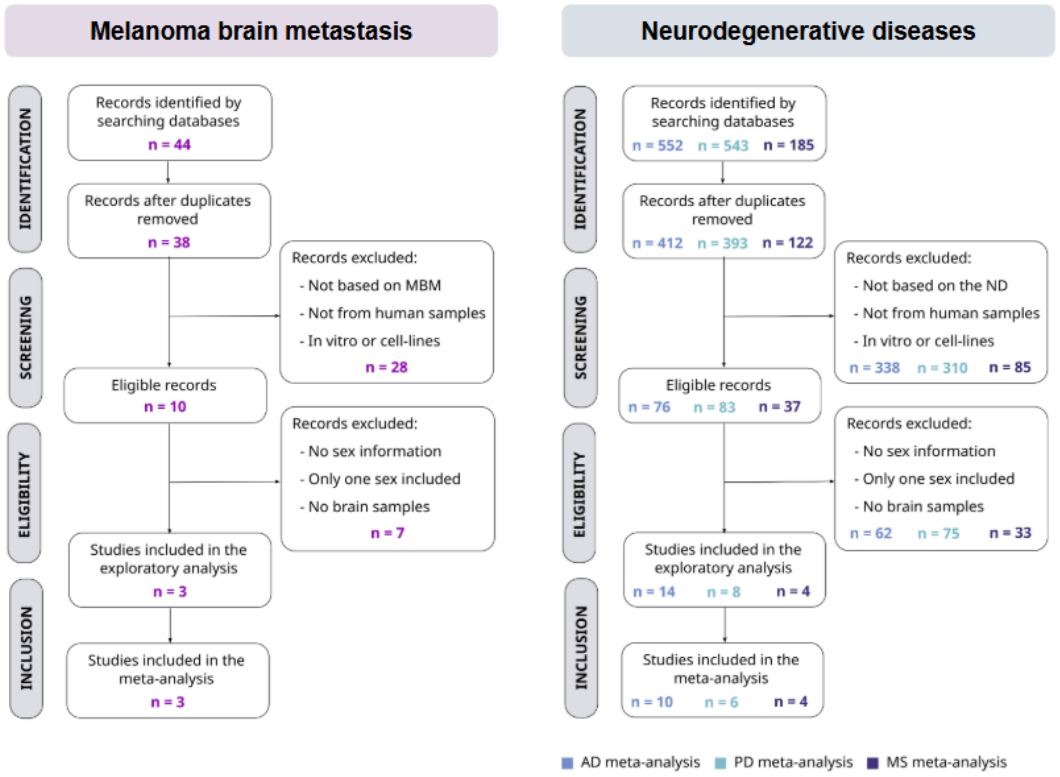


Figure 5.4. Systematic review conducted for melanoma brain metastasis (left) and neurodegenerative diseases (right). Multistep systematic review according to PRISMA guidelines.[135] The number of retained or discarded studies are indicated at each phase (n) for each disease (defined by color). *AD: Alzheimer's disease; MS: multiple sclerosis; ND: neurodegenerative disease; PD: Parkinson's disease; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.*

Table 5.1. Descriptive characteristics of the selected MBM studies. Upregulated genes refer to the number of significant genes with higher expression in the first term of the comparison (logFC > 0), and downregulated genes to those more highly expressed in the second term of the comparison (logFC < 0). *FFPE: Formalin-Fixed Paraffin-Embedded; MBM: melanoma brain metastasis; MNBM: melanoma non-brain metastasis; PMID: PubMed identifier; snRNA-seq: single nucleus RNA-seq.*

	MBM-1	**MBM-2**	**MBM-3**
Organism	Human	Human	Human
Data type	RNA-seq	snRNA-seq	RNA-seq
Samples	Microdissected FFPE tumors	Fresh frozen surgical resections	Microdissected FFPE tumors
Comparison	MBM vs MNBM	MBM vs MNBM	MBM vs non tumor-bearing brain controls
Results availability	All genes	All genes	Significant genes
Upregulated genes	418	7201	685
Downregulated genes	568	9377	431
PMID	30787016	35803246	36435874

Table 5.2. Meta-analysis results for the neurodegenerative diseases. (*Next page*) Upregulated genes refer to the number of significant genes more expressed in the first term of the comparison (logFC > 0), and downregulated genes to those more expressed in the second term of the comparison (logFC < 0). *MS brain region varies according to the localization of the lesion. For a detailed description refer to *Supplementary Table 1*. *AD: Alzheimer's disease; CT: cortex; HP: hippocampus; MS: multiple sclerosis; ND: neurodegenerative disease; PD: Parkinson's disease; SN: substantia nigra; ST: striatum.*

	AD-CT	AD-HP	PD-SN	PD-ST	MS
ND	Alzheimer's disease	Alzheimer's disease	Parkinson's disease	Parkinson's disease	Multiple sclerosis
Brain region	Cortex	Hippocampus	Substantia nigra	Striatum	Different*
Organism	Human	Human	Human	Human	Human
Data type	RNA-seq and microarray	RNA-seq and microarray	RNA-seq and microarray	RNA-seq and microarray	RNA-seq and microarray
Datasets in the meta-analysis	GSE118553, GSE125583, GSE132903, GSE15222, GSE37263, GSE48350, GSE5281, GSE84422	GSE1297, GSE29378, GSE48350, GSE5281, GSE84422	E-MEXP-1416, GSE20295, GSE7621, GSE8397	GSE20146, GSE20295, GSE28894	GSE108000, GSE111972, GSE131281, GSE135511
Comparison	AD *vs.* control	AD *vs.* control	PD *vs.* control	PD *vs.* control	MS *vs.* control
Upregulated genes	1505	1203	559	64	32
Downregulated genes	1702	296	782	159	50

5.4.2. MELANOMA BRAIN METASTASIS-SPECIFIC GENES ARE OFTEN FOUND DYSREGULATED IN MULTIPLE NEURODEGENERATIVE DISEASES

Genes differentially expressed by both MBM (compared to extracranial metastasis) and neurodegenerative diseases (compared to controls) may reveal cellular processes that are shared within the CNS microenvironment. Therefore, we searched for differentially expressed genes in MBM-1 and MBM-2 that were also found significantly altered in neurodegenerative diseases, as indicated by our five meta-analyses (AD-CT, AD-HP, PD-SN, PD-ST and MS) (**Supplementary Figure 5.S1-A**). We organized genes into four categories, which we termed as "patterns". Specifically, we identified genes with the four possible behaviors (**Figure 5.5-A**):

- Pattern 1: genes upregulated in both MBM and the corresponding neurodegenerative disease.
- Pattern 2: genes downregulated in both MBM and the corresponding neurodegenerative disease.
- Pattern 3: genes upregulated in MBM and downregulated in the corresponding neurodegenerative disease.
- Pattern 4: genes downregulated in MBM and upregulated in the corresponding neurodegenerative disease.

For each pattern, we further distinguished whether the genes were specific to a single neurodegenerative disease/region or shared across multiple conditions (**Supplementary Figure 5.S1-B**). As a result, most of the genes shared between MBM and neurodegenerative diseases were specific to a given disorder and brain region. Among the cross-disease overlaps, the strongest intersections were observed between MBM and AD-CT together with PD-SN, while very few shared genes were found between AD or PD with MS.

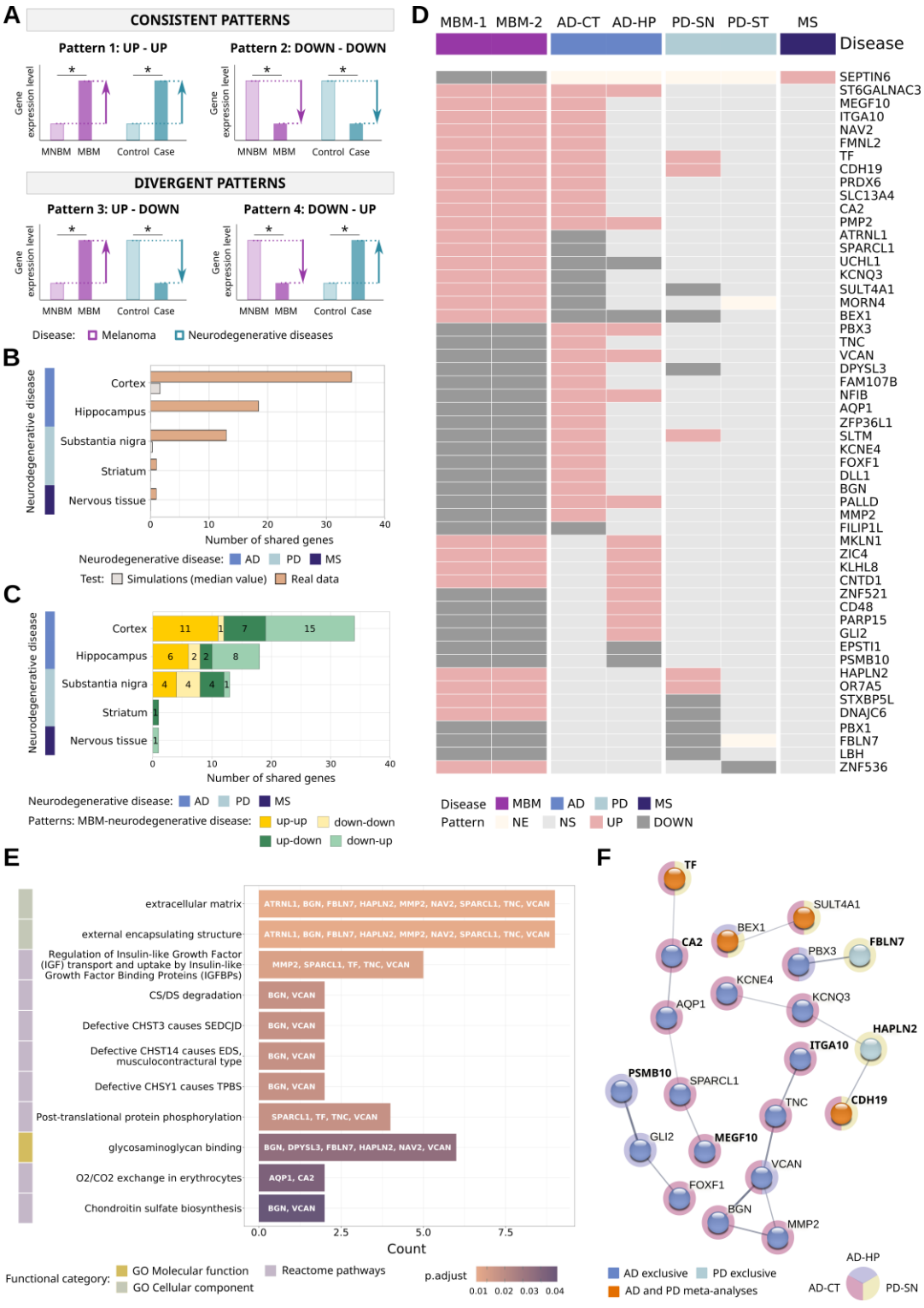
To assess the robustness of the detected signals, we focused on the subset of genes consistently significant in both MBM-1 and MBM-2, hereafter referred to as MBM-1-2 genes. We then examined which of these MBM-1-2 genes were also dysregulated across the three neurodegenerative diseases.

A resampling analysis, in which intersections were simulated using all evaluated genes, confirmed that in every scenario the observed number of shared significant genes was higher than expected by random chance (**Figure 5.5-B**). This finding supports the existence of a shared biological signal between MBM and neurodegenerative diseases. Next, we analyzed the distribution of MBM-1-2 genes by neurodegenerative disease and by expression pattern (**Figure 5.5-C**). Significant genes with all four patterns were identified in AD-CT, AD-HP, and PD-SN, while only one divergent pattern was detected in PD-ST and MS, respectively.

By combining all genes, we obtained the *MBM-1-2 neurodegenerative* signature, composed by a panel of 53 genes (**Figure 5.5-D**). Functional enrichment analyses of this signature revealed associations with 11 GO and Reactome terms (**Figure 5.5-E**). Notably, molecular function and cellular component categories pointed to extracellular matrix alterations, while most enriched Reactome pathways were driven by the proteoglycans VCAN (versican) and BGN (biglycan), which may also modulate cell–matrix interactions. We also generated the PPI network of the signature (**Figure 5.5-F**), which presented significantly more interactions than expected (PPI enrichment p-value:

0.00111). The resulting network included 19 genes dysregulated in AD, 6 in PD, and none in MS. Interestingly, genes with both consistent and divergent patterns were connected, without forming separate clusters, suggesting that these genes are functionally or physically related regardless of their expression patterns. Results for the genes and functions dysregulated by pattern and neurodegenerative disease can be found in the web tool (*Melanoma brain-specific metastasis signature* section).

Figure 5.5. Neurodegenerative signature of melanoma brain-specific metastasis. (*Next page*) (A) Qualitative representation of the patterns exhibited by significant genes. (B) Number of significant genes resulting from resampling analysis. (C) Distribution and (D) gene dysregulation profile by study/meta-analysis assessed. (E) Enriched functional terms. Number of genes from the neurodegenerative signature (X-axis) included in each significant function (Y-axis). Genes names displayed within the bars. (F) Protein-protein interaction network. Nodes represent genes, colored according to the neurodegenerative disease where it is significant. Normal text: consistent patterns; bold text: divergent patterns. Edge thickness: structural and functional confidence of the interaction. Non-connected nodes are not shown. *AD: Alzheimer's disease; CT: cortex; GO: gene ontology; HP: hippocampus; MBM: melanoma brain metastasis; MNBM: melanoma non-brain metastasis; MS: multiple sclerosis; NE: not evaluated; NS: not significant; PD: Parkinson's disease; SN: substantia nigra; ST: striatum.*


A
CONSISTENT PATTERNS
Pattern 1: UP - UP
Pattern 2: DOWN - DOWN
DIVERGENT PATTERNS
Pattern 3: UP - DOWN
Pattern 4: DOWN - UP
Gene expression level
MNBM MBM Control Case
Disease: Melanoma Neurodegenerative diseases
B
Neurodegenerative disease
Cortex
Hippocampus
Substantia nigra
Striatum
Nervous tissue
Number of shared genes
Neurodegenerative disease: AD PD MS
Test: Simulations (median value) Real data
C
Neurodegenerative disease
Number of shared genes
Neurodegenerative disease: AD PD MS
Patterns: MBM-neurodegenerative disease: up-up down-down up-down down-up
D
MBM-1 MBM-2 AD-CT AD-HP PD-SN PD-ST MS
Disease
SEPTIN6
ST6GALNAC3
MEGF10
ITGA10
NAV2
FMNL2
TF
CDH19
PRDX6
SLC13A4
CA2
PMP2
ATRNL1
SPARCL1
UCHL1
KCNQ3
SULT4A1
MORN4
BEX1
PBX3
TNC
VCAN
DPYSL3
FAM107B
NFIB
AQP1
ZFP36L1
SLTM
KCNE4
FOXF1
DLL1
BGN
PALLD
MMP2
FILIP1L
MKLN1
ZIC4
KLHL8
CNTD1
ZNF521
CD48
PARP15
GLI2
EPSTI1
PSMB10
HAPLN2
OR7A5
STXBP5L
DNAJC6
PBX1
FBLN7
LBH
ZNF536
Disease MBM AD PD MS
Pattern NE NS UP DOWN
E
extracellular matrix
external encapsulating structure
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
CS/DS degradation
Defective CHST3 causes SEDCJD
Defective CHST14 causes EDS, musculocontractural type
Defective CHSY1 causes TPBS
Post-translational protein phosphorylation
glycosaminoglycan binding
O2/CO2 exchange in erythrocytes
Chondroitin sulfate biosynthesis
Count
Functional category: GO Molecular function GO Cellular component Reactome pathways
p.adjust
F
TF
SULT4A1
BEX1
CA2
PBX3
FBLN7
KCNE4
AQP1
KCNQ3
HAPLN2
ITGA10
PSMB10
SPARCL1
TNC
CDH19
GLI2
MEGF10
VCAN
FOXF1
BGN
MMP2
AD exclusive PD exclusive
AD and PD meta-analyses
AD-HP
AD-CT PD-SN

To evaluate whether the detected signal was exclusive to MBM or represented a broader phenomenon of brain metastasis, we extended this characterization to breast cancer brain metastases by comparing them to extracranial brain metastasis (**Figure 5.6**). Of the 53 MBM-1-2 genes, 26 were also dysregulated in breast cancer brain metastases, all displaying the same expression patterns observed in MBM. This observation suggests that, although MBM may display unique features that enhance adaptation to the brain microenvironment, different brain metastatic cancers may converge on common molecular mechanisms.

Figure 5.6. Neurodegenerative profile of breast and melanoma brain metastases. Pattern of change of genes significantly dysregulated in MBM-1-2 and at least one neurodegenerative disease. Results are explored in brain breast metastasis (BBM), melanoma brain metastasis (MBM), Alzheimer's disease cortex region (AD-CT), Alzheimer's disease hippocampus region (AD-HP), Parkinson's disease substantia nigra region (PD-SN), Parkinson's disease striatum region (PD-ST) and multiple sclerosis (MS). Pattern: NE: not evaluated, NS: not significant, UP: upregulated, DOWN: downregulated.

We next investigated the functional profile of the MBM-1-2 neurodegenerative signature in AD, PD, and MS separately, classifying genes by their corresponding pattern. For clarity, in the following sections we will refer to the neurodegenerative disease, although all the described genes are included in the MBM-1-2 signature.

In AD Pattern 1 (genes upregulated in both MBM-1-2 and AD, either AD-CT or AD-HP), we identified functions related to cell adhesion molecule binding (CDH19, FMNL2, ITGA10, PRDX6), glycosylation (ST6GALNAC3), and lipid metabolism (PMP2). By contrast, AD Pattern 4 (genes downregulated in MBM-1-2 but upregulated in AD) encompasses the major functional implications. These genes were enriched in processes linked to neurogenesis, developmental programs, and extracellular matrix signaling (**Figure 5.7**). Within neurogenesis-related functions, we found genes specific to the hippocampus (ZNF521, GLI2) and others shared by both brain regions (VCAN, PBX3, PALLD, NFIB). Developmental functions were predominantly represented in the cortex, with AD-CT–specific genes such as ZFP36L1, KCNE4, FOXF1, and AQP1. Finally, extracellular matrix-related functions involved a smaller subset of genes detected in both cortical and hippocampal regions.

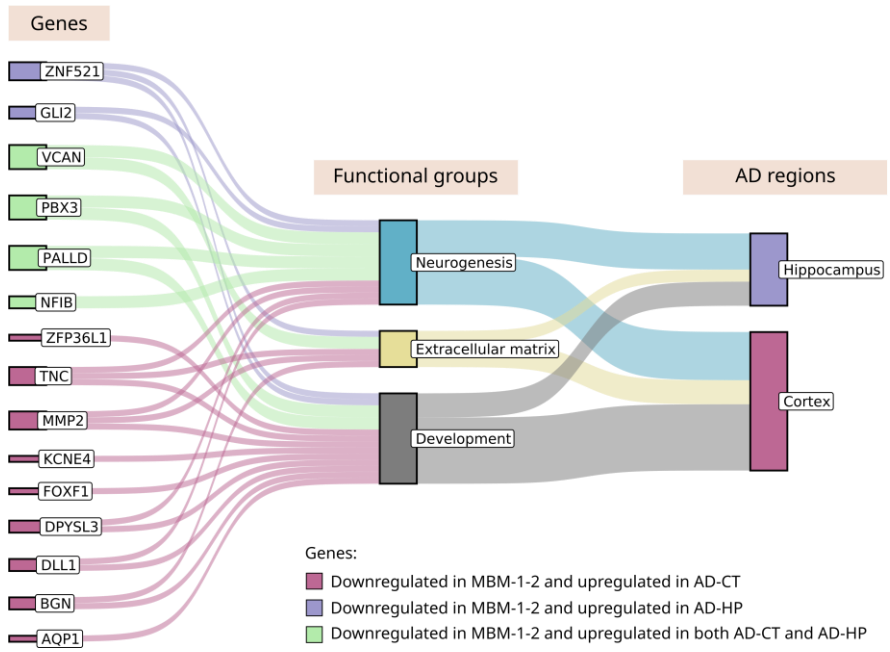


Figure 5.7. Functional classification of genes from MBM-1-2 and AD belonging to Pattern 4. Functional groups attributed to the genes downregulated in MBM-1-2 and upregulated in AD-CT and/or AD-HP. The size of the bars in *Functional groups* and *AD regions* is proportional to the total number of genes found in each feature. *AD: Alzheimer's disease; CT: cortex; HP: hippocampus; MBM: melanoma brain metastasis.*

Meanwhile, the functional significance of MBM-1-2 genes shared with PD are those related to the substantia nigra region. Among upregulated genes in both diseases, we identified CDH19, HAPLN2, and OR7A5, linked to adhesion and signal transduction. Conversely, glycosaminoglycan-binding genes such as DPYSL3 and FBLN7 were consistently downregulated in both conditions. Notably, all genes classified under Pattern 3 (upregulated in MBM-1-2 but downregulated in PD-SN) displayed neuronal-specific features, including BEX1 (a potential glioblastoma biomarker), STXBP5L (neurotransmitter release), DNAJC6 (neuronal endocytosis), and SULT4A1 (neurotransmitter metabolism). Finally, within the neurodegenerative signature, only one gene overlapped with MS: SEPTIN6, a GTPase involved in cytokinesis, which was downregulated in MBM-1-2 and upregulated in MS.

5.4.3. COMMON GENETIC FEATURES BETWEEN MELANOMA BRAIN METASTASIS TUMORAL SIGNATURE AND NEURODEGENERATIVE DISEASES PROFILES

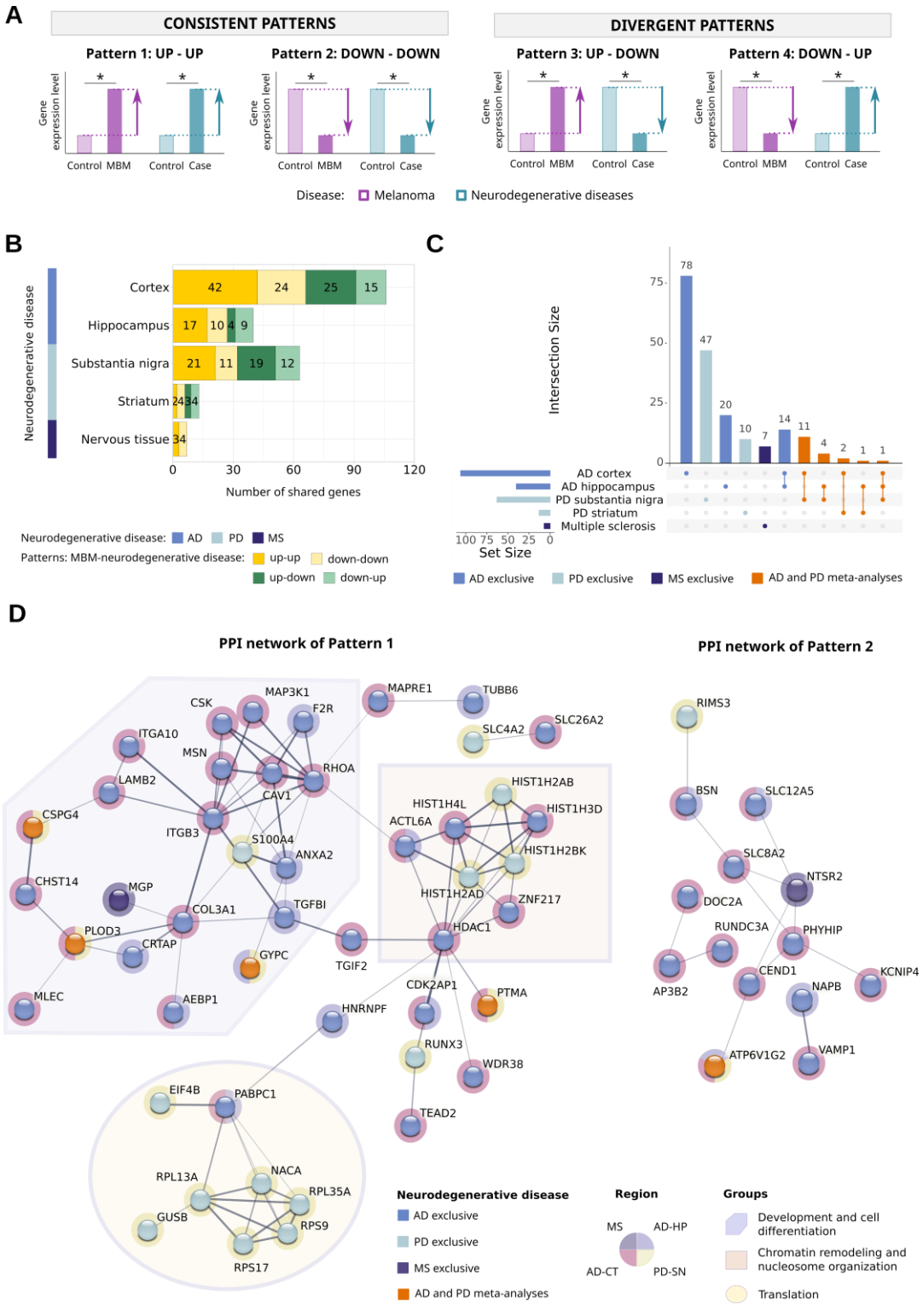
As an orthogonal approach to analyze studies comparing brain and extracranial melanoma metastasis, we employed data from MBM-3, which compares the gene expression profiles of melanoma-bearing brain samples and healthy brain regions. We compared the differential gene expression results obtained in MBM-3 with the results of the five meta-analyses of neurodegenerative diseases (AD-CT, AD-HP, PD-SN, PD-ST and MS). In total, 195 significant genes were identified across all four patterns (**Figure 5.8-A**), with the majority of genes belonging to consistent Patterns 1 and 2 (**Figure 5.8-B**). The highest number of shared genes are reported in the intersections of MBM-3 with AD-CT, AD-HP and PD-SN: 112 were AD-specific, 57 PD-specific, 7 MS-specific and 19 were dysregulated in both AD and PD (**Figure 5.8-C**). None of the MS associated genes were dysregulated in the other neurodegenerative diseases.

Among the total genes, those belonging to consistent patterns stand out. Genes upregulated in both MBM-3 and the neurodegenerative diseases constitute a significant PPI network (PPI enrichment p-value: 2.79e-08) organized into three major clusters (**Figure 5.8-D, Pattern 1**). The first cluster includes translation-related genes predominantly linked to PD (RPL13A, NACA, RPL35A, RPS9, RPS17, GUSB, PABPC1, EIF4B). The second cluster groups development and cell differentiation genes, mostly associated with AD (MAP3K1, F2R, RHOA, CAV1, MSN, ITGB3, S100A4, ANXA2, TGFBI, GYPC, COL3A1, AEBP1, CRTAP, MGP, PLOD3, MLEC, CHST14, CSPG4, LAMB2, ITGA10). The third cluster involves chromatin remodeling and nucleosome organization genes, with a balanced distribution across AD and PD

(HIST1H2AB, HIST1H3D, HIST1H2BK, HIST1H2AD, HIST1H4L, ACTL6A, ZNF217, HDAC1). Conversely, Pattern 2 composed of genes downregulated in both MBM-3 and neurodegenerative diseases, is largely dominated by AD-related genes. They also generate a significant network (PPI enrichment p-value: 0.00015) (**Figure 5.8-D, Pattern 2**), composed of genes involved in the synaptic process, particularly those contributing to vesicle formation.

The PPI network of the two divergent patterns combined was also significant (PPI enrichment p-value: 0.00564). Most of the signal was driven by Pattern 3 (PPI enrichment p-value: 8.63e-07), which includes genes upregulated in MBM-3 but downregulated in the corresponding neurodegenerative disease. The corresponding genes were functionally linked to cytoskeleton organization, metabolism, and proteasome activity. By contrast, Pattern 4 did not generate a significant network. The neurodegenerative signature, together with the functional characterization and PPI networks of these genes for each disease, can be further explored in our web tool (*Melanoma brain metastasis tumoral signature* section, *Functional profiling* subsection).

Figure 5.8. Neurodegenerative signature for MBM tumoral profile. (*Next page*) (A) Qualitative representation of the patterns exhibited by significant genes in MBM-3 and neurodegenerative diseases meta-analyses. (B) Number of significant genes of the signature separated by neurodegenerative disease and pattern. (C) Upset plot of the signature. Representation of the total number of significant genes by neurodegenerative disease (horizontal bars), and the fraction present in the intersection of the designated groups (vertical bars), identified by colored dots beneath. (D) Protein-protein networks for consistent genes between MBM-3 and neurodegenerative diseases: Pattern 1 (upregulated genes), Pattern 2 (downregulated genes). Nodes correspond to genes, color-coded established based on their significance in the neurodegenerative diseases. Edge thickness indicates the structural and functional confidence of the interaction. Non-connected nodes have been discarded. *AD: Alzheimer's disease; MBM: melanoma brain metastasis; MS: multiple sclerosis; PD: Parkinson's disease*.


A
CONSISTENT PATTERNS
DIVERGENT PATTERNS
Pattern 1: UP - UP
Pattern 2: DOWN - DOWN
Pattern 3: UP - DOWN
Pattern 4: DOWN - UP
Gene expression level
Control MBM
Control Case
Disease: Melanoma Neurodegenerative diseases
B
Neurodegenerative disease
Cortex 42 24 25 15
Hippocampus 17 10 4 9
Substantia nigra 21 11 19 12
Striatum 2 4 3 4
Nervous tissue 3 4
0 30 60 90 120
Number of shared genes
Neurodegenerative disease: AD PD MS
Patterns: MBM-neurodegenerative disease: up-up down-down up-down down-up
C
Intersection Size
78 47 20 10 7 14 11 4 2 1 1
AD cortex
AD hippocampus
PD substantia nigra
PD striatum
Multiple sclerosis
100 75 50 25 0
Set Size
AD exclusive
PD exclusive
MS exclusive
AD and PD meta-analyses
D
PPI network of Pattern 1
PPI network of Pattern 2
MAP3K1
CSK
F2R
ITGA10
MSN
RHOA
LAMB2
CAV1
CSPG4
ITGB3
S100A4
ANXA2
CHST14
MGP
COL3A1
TGFBI
PLOD3
CRTAP
GYPC
MLEC
AEBP1
MAPRE1
TUBB6
SLC4A2
SLC26A2
HIST1H2AB
HIST1H4L
HIST1H3D
ACTL6A
HIST1H2BK
HIST1H2AD
ZNF217
HDAC1
TGIF2
CDK2AP1
PTMA
HNRNPF
RUNX3
WDR38
TEAD2
EIF4B
PABPC1
RPL13A
NACA
RPL35A
GUSB
RPS9
RPS17
RIMS3
BSN
SLC12A5
SLC8A2
NTSR2
DOC2A
RUNDC3A
PHYHIP
AP3B2
CEND1
KCNIP4
NAPB
ATP6V1G2
VAMP1
Neurodegenerative disease
AD exclusive
PD exclusive
MS exclusive
AD and PD meta-analyses
Region
MS
AD-HP
AD-CT
PD-SN
Groups
Development and cell differentiation
Chromatin remodeling and nucleosome organization
Translation

5.4.4. DNAJC6 AND ITGA10 AS POTENTIAL KEY GENES IN THE NEUROBIOLOGY OF MELANOMA BRAIN METASTASES

To identify genes that were not only differentially expressed in MBM compared to control tissue (MBM-3 genes) but also specific to brain metastasis (MBM-1-2 genes), we performed an intersection analysis to identify significant genes in all three MBM differential gene expression results. We then determined which of these genes were also dysregulated in neurodegenerative diseases. Two genes met these gene expression patterns: ITGA10, which is upregulated in MBM-1, MBM-2, MBM-3 and AD-CT; and DNAJC6, which is upregulated in MBM-1 and MBM-2, whilst downregulated in MBM-3 and PD-SN. The consistent dysregulation of these genes across MBM and neurodegenerative contexts suggests they may define conserved molecular adaptations to the brain microenvironment.

5.4.5. WEB TOOL

The web resource https://bioinfo.cipf.es/metafun-mbm/ (**Figure 5.9**) includes the detailed results of each section developed in this manuscript, offering free access to users for confirmation of the described findings. The interactive interface allows users to explore the intermediate results generated to obtain the dysregulated profiles of MBM and the neurodegenerative diseases (i.e., systematic review, exploratory analysis, differential expression and meta-analysis). Users will also be able to explore the neurodegenerative signature of MBM-1-2 and MBM-3 together with the functional profiles and PPI networks presented in this manuscript.

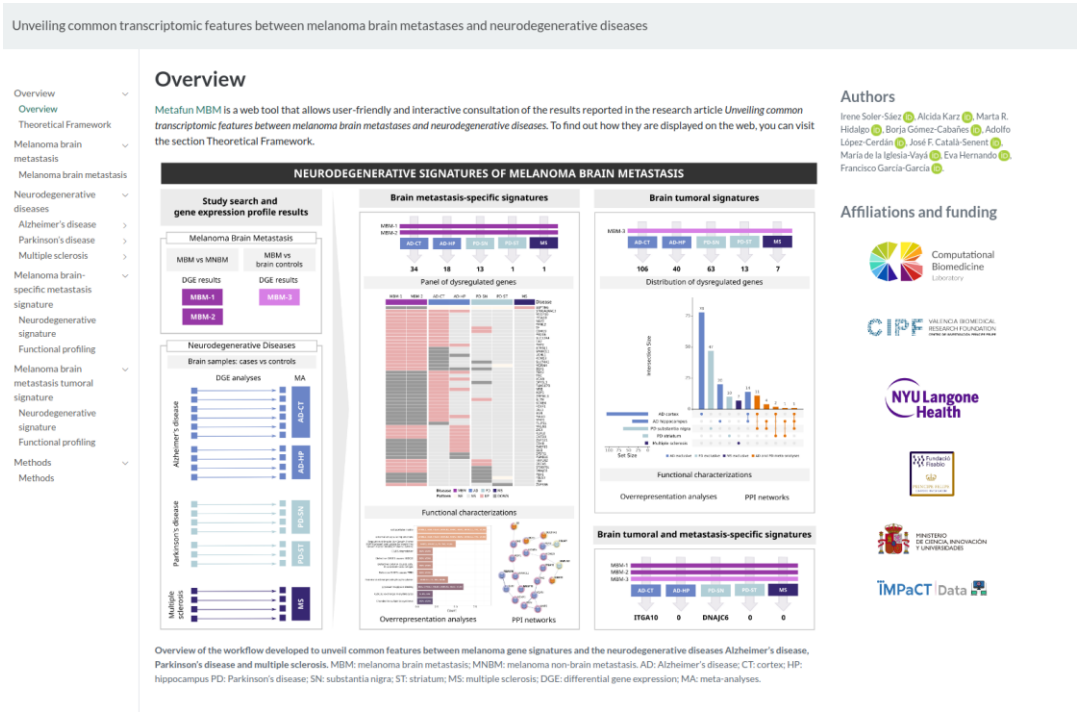


Figure 5.9. Home page of the web tool. The side navigation bar provides access to the different modules of the platform: Overview, individual results for melanoma brain metastasis, individual results for neurodegenerative diseases (Alzheimer's disease, Parkinson's disease, and Multiple sclerosis), the neurodegenerative signatures for melanoma brain-specific metastasis and tumoral melanoma metastasis, and the description of the computational methods. The central section displays the framework of the tool, while the right panel includes author information, institutional affiliations, and funding acknowledgements. *MBM: melanoma brain metastasis.*

5.5. DISCUSSION

Cancer neuroscience has emerged as a multidisciplinary field that seeks to understand how tumor cells interact with the nervous system to promote invasion and colonization of distant tissues[590,591]. Among the research covering this theme, different studies have characterized the crosstalk between tumor cells and CNS cell types, including neurons, astrocytes, and microglia, revealing their capacity to shape a tumor-permissive microenvironment[590,591]. This line of research has also translated into clinical applications, where neuroactive drugs are being tested in combination with conventional anticancer therapies, with the goal of reducing or disrupting tumor–nervous system

interactions to improve patient outcomes[592]. In this study, we expand on this perspective by highlighting the molecular mechanisms shared between MBM and neurodegenerative diseases. This perspective suggests a bidirectional opportunity: while cancer neuroscience could gain from integrating concepts of neurodegeneration, the study of tumor biology may provide novel perspectives on unresolved mechanisms in neurodegenerative disorders.

The gene expression profiling of MBM compared to extracranial metastases (MBM-1-2) revealed a significant overlap with neurodegenerative pathologies. Throughout our analysis, genes associated with the extracellular matrix emerged as a point of congruence between neurodegenerative diseases and MBM-1-2. While the extracellular matrix constitutes a major structural component of the brain, occupying approximately 20% of its volume, its involvement in tumor progression is still poorly characterized[593]. In our work, different genes related to the extracellular matrix emerged with potential relevance for both MBM and neurodegeneration.

One such gen is DPYSL3, which was downregulated in MBM-1-2 and PD-SN, but upregulated in AD-CT. DPYSL3 was identified by the Clinical Proteomic Tumor Analysis Consortium as a specific marker of claudin-low triple-negative breast cancers, and its knockdown slowed proliferation in breast cancer cell lines while increasing motility and markers of the epithelial-to-mesenchymal transition[594]. Thus, DPYSL3 may integrate signals that modulate both proliferative dynamics and invasive traits in different environments. The dysregulation of glycosylation-related genes is also of great relevance, as glycosylation is a process with unique features in brain cancers[595]. Among the dysregulated genes we identified that BGN, being downregulated in MBM-1-2 and upregulated in AD-CT and AD-HP. Strikingly, BGN knockout mice exhibit impaired memory and metabolic dysfunction, linking its expression to brain physiology[596]. Moreover, BGN was also shown to be upregulated in dormant breast cancer brain metastasis models, suggesting a potential role in long-term tumor cell survival within the brain niche[597]. Likewise, the glycosylation-related enzyme ST6GALNAC3, upregulated in MBM-1-2, AD-CT and AD-PD, may play a significant role in metastasis as it influences the development or progression of prostate cancer[598].

Another neuronal-like feature of metastatic brain cells is neurogenesis, mostly represented by genes downregulated in MBM-1-2 and upregulated in AD-CT and/or AD-HP. Among the potential targets we found the RNA polymerase II regulators ZNF521, PBX3 and NFIB. These transcriptional regulators modulate the expression of mRNAs and lncRNAs involved in neuron fate commitment and migration, which may trigger alterations in neurogenesis previously characterized in AD[599]. The dysregulation

of RNA polymerase II itself could also be a central mechanism linking tumor progression with neuronal dysfunction[600].

Interestingly, the MBM associated signature with PD-SN revealed the opposite pattern, with genes upregulated in MBM-1-2 but downregulated in PD-SN. The genes involved also exhibit neuronal-specific attributes. Among them: BEX1 is a promoter of glioblastoma progression[601]; STXBP5L is a syntaxin-binding protein involved in neurotransmitter release and highly expressed in glioma tissues[602]; DNAJC6 is a heat shock protein associated with neuronal endocytosis which favors epithelial-mesenchymal transition in hepatocellular carcinoma[603]; and SULT4A1 is a brain-specific sulfotransferase involved in neurotransmitter metabolism that promotes growth under low oxygen concentration[604]. Specifically, SULT4A1 may be relevant to protect against the oxidative stress generated by enhanced oxidative phosphorylation in MBM[605], while simultaneously modulating dopamine toxicity by interacting with SULT1A3[606]. This dual functionality of SULT4A1 may help explain its opposite regulation in MBM and PD-SN: upregulation in MBM could represent an adaptive response to mitigate oxidative stress, whereas downregulation in PD-SN may exacerbate dopamine toxicity and increase neuronal vulnerability.

In addition to MBM, we extended our analyses to breast cancer brain metastases, comparing them against extracranial breast metastases. Interestingly, approximately half of the MBM-1-2 signature genes were also dysregulated in BBM, and consistently displayed the same expression patterns. This overlap suggests that certain molecular programs may be conserved across distinct tumor types that metastasize to the brain, reflecting common adaptations to the neural microenvironment. However, the other half of the genes were not significant in BBM, supporting the idea that specific mechanisms to the different tumor types also play an important role in brain colonization. Previous work has shown that while brain metastases from different primary tumors share features such as neuronal mimicry and metabolic adaptation, they also retain signatures linked to the biology of their tissue of origin[519,607]. Nevertheless, given the developmental origin of melanocytes in the neural crest, melanoma may share deeper molecular gene expression patterns with neurodegeneration than other cancer types. Further research is needed to test whether these mechanisms are similar in other forms of brain metastasis.

We extended our comparison of MBM and extracranial melanoma metastases by analyzing the neurodegenerative profile of MBM relative to healthy brain regions (MBM-3 results). Although MBM and neurodegenerative diseases have distinct outcomes, our results reveal overlapping molecular programs. Genes upregulated in both MBM and neurodegenerative diseases conformed a significant PPI network, in which three clusters emerge related to development, translation, and chromatin remodeling.

Development-related genes were particularly upregulated in MBM-3 and AD. Again, the cluster mainly involved extracellular matrix associated genes such as integrins (e.g., ITGB3, ITGA10), collagen-related (e.g., TGFBI, COL3A1) and glycosylation-related (e.g., GYPC, MLEC, LAMB2) genes. As discussed above, the extracellular matrix plays a pivotal role in both contexts. Specifically, MBM requires adaptation to the specialized brain microenvironment, while in neurodegeneration the interactions between the extracellular matrix and neuronal receptors may regulate the maintenance of processes such as cell migration, axon guidance, and synapse formation[608].

Meanwhile, translation-related genes were predominantly upregulated in MBM-3 and PD-SN. These results are consistent with Bowley *et al.* 2022[609], who reported that the expression signature of large and small ribosomal subunits in melanoma circulating tumor cells is associated with MBM onset and progression, in both animal models and patient samples. In PD, Garcia-Esparcia *et al.* 2015[610] documented the alteration of rRNA synthesis and altered translation initiation/elongation factors in the substantia nigra and cortex, and Martin *et al.* 2015[611] also highlighted the role of dysregulated mRNA translation in PD pathogenesis.

The third cluster consisted of genes upregulated in MBM-3, AD, and PD, which were primarily related to chromatin remodeling, including multiple histone subunits and epigenetic regulators. Among them, histone deacetylase HDAC1 is of particular interest, as its knockdown has been shown to promote autophagy in HeLa cells[612]. Therefore, its upregulation in MBM and neurodegenerative diseases may instead contribute to reduced autophagy. Impaired autophagy has different but equally detrimental consequences: in neurodegeneration, it fosters the accumulation of toxic aggregates, while in tumors it may allow cells to adapt themselves through alternative metabolic routes. The importance of autophagy has led to the development of therapies targeting it, with potential applications in both cancer (including melanoma) and neurodegenerative diseases[613,614].

Lastly, ITGA10 and DNAJC6 are highlighted as potential key biomarkers for being significant in all MBM profiles and at least one neurodegenerative disease. ITGA10, an integrin with a crucial role in cell adhesion, is upregulated in MBM-1, MBM-2, MBM-3 and AD-CT. This gene has previously been associated with metastasis, and evidence has shown higher expression in cutaneous melanoma compared to normal skin[615]. DNAJC6 presents chaperone activity and is upregulated in MBM-1 and MBM-2, while downregulated in MBM-3 and PD-SN. Mutations in DNAJC6 have been widely reported to be associated with juvenile onset of PD[616,617]. While uncharacterized in the context of melanoma to the best of our knowledge, this gene is reported to be involved

in the hepatocellular carcinoma epithelial-mesenchymal transition via activation of the TGF-β pathway[603].

We conducted a meta-analysis strategy whenever possible, integrating data from multiple datasets to increase the robustness and reproducibility of our findings. Nevertheless, variability in technical platforms, clinical annotations, patient phenotypes, and prior treatments constrained the inclusion of some studies within the meta-analyses. Despite these limitations, sufficient data were available to conduct meta-analyses for AD-CT, AD-HP, PD-SN, PD-ST, and MS. In contrast, for MBM the limited number of publicly accessible transcriptomic datasets did not allow us to reach the minimum threshold of studies required for a formal meta-analysis. Consequently, the MBM transcriptomic profiles were characterized through independent analyses of each dataset.

Another important factor to consider in our analysis is the purposeful *apples-to-oranges* approach: we have compared gene expression signatures from tumors composed of melanoma cells to changes in brain tissue in neurodegenerative pathology. This strategy was taken in order to further explore the neuronal-mimicry capacity of MBM, identified by previous works in the field[143,548]. Despite the intrinsic biological differences between the tissues analyzed, we identified common molecular changes. Future research should extend these comparisons to the altered brain microenvironments in MBM and neurodegeneration, ideally using single-cell approaches, since the brain is a complex ecosystem where each cell type plays a specific role in disease progression.

Overall, our findings identify novel intersections between MBM and neurodegenerative disease. The unveiled dysregulated genes, together with their associated biological functions and signaling pathways, represent potential biomarkers of future study to further advance our understanding of MBM towards developing new strategies to improve clinical outcomes for the patients.

6. General discussion

This thesis aimed to advance the molecular characterization of diseases affecting the CNS, through a holistic approach based on omics data analyses. In the three studies presented in this manuscript, we relied on publicly available datasets to address new research questions beyond the scope of the original studies. In *Study I*, we investigated sex differences in MS using single-cell transcriptomics from both CNS and peripheral blood samples. In *Study II*, we expanded this perspective to the gut microbiome, identifying sex-differential abundant taxa and linking them to clinical characteristics of MS. Finally, in *Study III*, we broadened the research to cancer neuroscience, exploring molecular mechanisms shared between melanoma brain metastases and neurodegenerative diseases. In this section, we summarize the findings of each study, while also reflecting on transversal aspects concerning all three works.

In *Study I,* we identified the potential molecular mechanisms underlying sex differences in MS. Sex plays a critical role in the disease: females are two to three times more likely to develop MS and often display more inflammatory patterns, whereas males tend to experience faster neurodegeneration with greater severity[618]. To gain insight into these processes, we analyzed single-cell transcriptomic data from both CNS tissue and PBMCs, the two most affected tissues in MS. We tried to cover MS spectrum based on the available datasets, generating cell-type specific landscapes for SPMS in the CNS, and for RRMS and PPMS in PBMCs. Neurons, astrocytes, microglia, oligodendrocytes and OPCs were identified for CNS samples; and CD4+ T cells, CD8+ T cells, NK cells and monocytes for PBMCs. For each MS subtype and cell type, evaluated the gene signatures from differentially expressed genes, functional profiling, pathway activation, and cell-cell communication networks for females, males, and their sex-differential profiles.

Our results suggest that female neurons in SPMS may exhibit protective mechanisms against neurodegeneration, potentially counteracting excitotoxicity. In addition, females may experience slower disease progression by being more resistant to the demyelination process. We identified that female oligodendrocytes expressed more neuronal-like genes, and relied on cell-cell interactions that may cope with myelin damage compared to males.

Regarding PBMCs, in the inflammatory form RRMS, females presented increased expression of AP-1 subunits (FOS, JUN, JUNB, JUND) across all four evaluated immune cell types, which may contribute to the stronger immune activation. By contrast, the adaptive immune system of RRMS males exhibited higher expression of genes related to the mitochondrial electron transport chain, which may be linked to greater mitochondrial impairment and oxidative stress. Conversely, in PPMS, which is primarily characterized by neurodegeneration, we observed the largest sex-related differences in

CD8+ T cells. Males exhibited cytolytic profiles that may accelerate neurodegeneration when crossing the blood brain barrier, while females presented higher expression in genes related to regulatory and maintenance functions. When comparing the PBMCs results between the two subtypes, we identified 67 genes with strong sex-dependent differences. They were involved in functions associated with reactive oxygen species, cytokines, lipids, and leukocyte differentiation. Taken together, our findings indicate that sex differences in MS are not uniform but rather influenced by the disease subtype. These results also align with previous reports that distinguish different inflammatory dynamics in RRMS and progressive forms of MS[158,619].

In *Study II*, we extended the analysis of sex differences in MS to the gut microbiome. This is of particular relevance, since MS pathology does not develop in germ-free mouse models[400,401]. To the best of our knowledge, this constitutes the first integrative meta-analysis explicitly focused on sex and MS. Compared with *Study I*, the broader availability of public datasets enabled us to perform an integrative approach. Our analysis revealed the large variability attributable to the dataset of origin, which explained over 20% of microbial composition variance. Such variability likely reflects both technical and biological factors. Indeed, three studies had to be excluded due to excessive discrepancies in their taxonomic profile.

By integrating six independent 16S datasets, we identified 12 taxa associated with MS in a sex-related manner. These included taxa with tendencies that reflect basal sex differences that were amplified in the disease, others that were sex-specific, and others were both sex and disease status contributed. Of them, five taxa were subsequently validated in the cohort evaluated by the iMSMS consortium, the largest human MS cohort published to date[409]. In *Coprococcus* we identified a significantly reduced abundance in MS females *versus* female controls; *Eggerthella*, *Eisenbergiella*, and *Flavonifractor* with significantly higher abundance in MS females compared to MS males; and *Prevotella* with significantly higher abundance in MS males compared to MS females. These taxa have been previously linked to immunomodulatory functions and SCFA production[406,409,413]. In our work, we were also able to associate some of them with clinical features of MS. Specifically, *Eggerthella* was positively associated with disease duration and SPMS subtype, while *Eisenbergiella* was positively linked with disease duration, severity and progressive MS subtypes. *Flavonifractor* showed a negative tendency associated with disease severity in MS males, suggesting a potential sex-specific immunomodulatory role linked to flavonoid metabolism. Finally, although *Prevotella* was not directly associated with clinical features of MS, its higher abundance in males suggests that it may exert protective effects by modulating inflammatory responses and enhancing SCFA production[497]. Notably, our findings are based on

taxonomic profiling, which inherently limits the depth of biological interpretation. Future studies using whole-genome shotgun metagenomics or metabolomics will be required to characterize more precisely the sex-differential functional alterations.

Lastly, the *Study III* broadened the focus of this thesis to include other CNS-related disorders. Melanoma represents a major clinical challenge due to its aggressive nature and poor prognosis. This malignancy shows a remarkable capacity to adapt to the brain microenvironment, yet the molecular mechanisms underlying this adaptation remain poorly understood. Recent evidence suggests that MBM may share biological processes that overlap with those observed in neurodegenerative diseases[143]. To further explore this connection, we compared the transcriptional profiles of MBM with the consensus profiles of AD, PD, and MS. Two complementary scenarios were considered: the melanoma brain-specific metastatic signature, obtained by studies that compare MBM with extracranial melanoma metastases; and the melanoma brain metastasis tumoral signature, derived from studies that compare MBM with non–tumor-bearing brain tissue.

Specifically, we identified that the melanoma brain-specific metastatic signature shared 53 genes that were also altered in at least one of the neurodegenerative diseases. These genes were enriched in 11 functional categories, primarily related to the extracellular matrix and developmental processes. In detail, the genes shared between MBM and AD included potential transcriptional regulators of RNA polymerase II, such as ZNF521, PBX3, and NFIB, which are implicated in the neuronal fate commitment[600]. By contrast, the genes overlapping with PD exhibited neuronal-like attributes, with some of them previously associated with gliomas[601-604]. Meanwhile, only SEPTIN6 was commonly dysregulated in MBM and MS, which is a GTPase involved in cytokinesis. When MBM was compared to non–tumor-bearing brain controls, 195 genes were found to be shared with the neurodegenerative diseases. These genes were primarily associated with development and cell differentiation, translation, and chromatin remodeling. Specifically, extracellular matrix and development-related genes were altered in MBM and AD, translation-related genes in MBM and PD, and chromatin remodeling genes in MBM, AD and PD.

Across the two MBM scenarios, two genes were dysregulated and shared with the neurodegenerative diseases: ITGA10 (an integrin involved in cell adhesion) and DNAJC6 (a chaperone associated with juvenile onset of PD and tumoral epithelial-mesenchymal transition)[603]. They represent potential molecular markers at the interface between neurodegeneration and brain metastasis biology.

This doctoral thesis has been made possible entirely through the analysis of publicly available datasets, originally generated by other groups to address their own research

questions. By exploration of these data, we have been able to formulate new hypotheses of interest. In particular, this thesis has explored the dimension of sex in MS, a factor not explicitly considered in the original studies. It has also identified shared dysregulated genes across different diseases by integrating evidence from datasets that were initially focused on a single pathology.

The analysis of previously published data offers multiple advantages but also presents different challenges. Among its benefits, this approach allows us to maximize the use of existing resources, avoiding the need to recruit new participants. By combining heterogeneous datasets, we can increase statistical power and robustness, as larger sample sizes can be achieved. Working with publicly available data also fosters transparency and reproducibility, as it enables cross-validation and reinforces confidence in previously published results. In addition, access to diverse cohorts allows researchers to study different disease subtypes, tissues, or experimental conditions that would otherwise be difficult to consider in a single study[620–623].

However, one of the recurrent limitations is the lack of metadata standardization. The current process of metadata documentation is dependent on the authors who generate the data. Thus, relevant metadata frequently are either missing or inconsistently reported, since their inclusion depends on the willingness and awareness of the original researchers[624,625]. Even when metadata files are provided, we normally found inconsistencies in metadata annotation: attribute names often differ across studies, information is frequently incomplete, and there is a general lack of controlled vocabulary (for example, the same condition may be reported as “MS” or “multiple sclerosis”)[626,627]. An illustrative example from our work was the standardization of treatment variables in *Study II*. In some datasets, patients were annotated as having received treatment, without specifying which drug. Meanwhile, others presented a mix of generic names and commercial brands. These were often reported with inconsistent use of capitalization, hyphens, or typographical errors.

These problems are further accentuated when sample identifiers in the metadata do not match those in the corresponding omics datasets, making unfeasible their association. A further limitation is that data can be available at different stages of the analytical workflow, ranging from raw sequencing reads to pre-processed normalized count matrices[128], which limits the computational strategies that can be applied across studies.

Therefore, beyond encouraging that research data are made publicly available, it is also important to promote the FAIR principles[123]. These guidelines, developed by scientists across diverse disciplines, aim to standardize data deposition practices, ensuring that datasets are stored in repositories in a homogeneous and structured manner. By

following these principles, other researchers can more easily locate, access, and ultimately reuse the data, thereby maximizing their scientific value.

Whenever possible, we conducted a meta-analysis strategy. In many cases, different studies share a similar experimental design, even if they were originally defined to answer different research questions. Notably, results derived from the same omics technology under the same comparisons can often be contradictory between studies. In this context, meta-analysis provides a powerful approach, allowing the integration of individual results to extract consensus profiles, as outlined in the introduction of this thesis[628,629]. Specifically, we applied the meta-analysis strategy in *Study II* to integrate 16S microbiome datasets, and in *Study III* to identify consensus patterns in AD, PD, and MS by comparing cases versus controls samples. Conversely, individual analyses were performed in *Study I* for scRNA-seq data and in *Study III* for the MBM transcriptomic profiles.

Considering the above, one limitation of our work is the imbalance and small size of the available datasets. For instance, in the scRNA-seq analyses, results rely on a single dataset per tissue and disease subtype. Additional datasets would be highly valuable to confirm our findings, either by increasing sample size, enabling new meta-analyses, or providing independent validation across the three studies of this thesis. Moreover, it is important to note that we cannot accurately determine whether the mechanisms described are deregulated to exacerbate disease conditions or, conversely, to modulate disease processes. Likewise, we cannot fully resolve apparently contradictory responses, which may instead reflect the multifaceted nature of the conditions evaluated. Future studies, particularly those incorporating complementary omics technologies and experimental validation of specific mechanisms, will be of interest to extend the characterization of our results.

To further promote transparency and reproducibility, the code generated in this thesis has been deposited in public GitHub repositories:

- *Study I*: https://github.com/IrSoler/cbl-atlas-ms
- *Study II*: https://github.com/IrSoler/cbl-metaanalysis-16S-MS
- *Study III*: https://github.com/IrSoler/cbl-neurodegeneratives-mbm

In addition, we developed interactive web tools using Quarto and Shiny, enabling researchers to navigate the full set of results and to examine specific aspects of interest that were not detailed in this manuscript:

- *Study I*: https://bioinfo.cipf.es/cbl-atlas-ms/
- *Study II*: https://irsoler.shinyapps.io/metaanalisis_16S_MS/

- *Study III*: https://bioinfo.cipf.es/metafun-mbm/

Overall, we hope that the results of this doctoral thesis help to improve the understanding of CNS-associated diseases, from sex-related differences in MS to the neurodegenerative features of MBM. We also expect these findings to be useful for the wider scientific community, providing a basis to identify potential targets, biomarkers, and new hypotheses. In the long term, this knowledge may contribute to a better characterization of these diseases, leading to advances in their diagnosis and treatment.

7. Conclusions

The principal conclusions derived from this work are summarized as follows:

1. This thesis was conducted entirely analyzing publicly available datasets, applying meta-analysis strategies whenever possible to obtain consensus patterns from independent studies.

2. Single-cell transcriptomic analysis revealed sex-differential molecular mechanisms across multiple sclerosis subtypes. In the secondary progressive form, female brain cell types may promote protective responses against neurodegeneration. In the relapsing-remitting form, female immune cells exhibited an inflammatory gene core, whereas male immune cells may be more prone to mitochondrial impairment. In the primary progressive form, male CD8+ T cells exhibited cytolytic profiles, while females showed higher expression of genes that support homeostatic processes.

3. The integrative analysis of 16S rRNA sequencing datasets identified microbial taxa differentially associated with sex in multiple sclerosis patients. Notably, *Eggerthella*, *Eisenbergiella*, and *Flavonifractor* were more abundant in females, whereas *Prevotella* was more abundant in males. These genera, implicated in immune modulation and short-chain fatty acid production, may modulate sex-differential disease progression.

4. Neurodegenerative transcriptomic signatures were identified for melanoma brain metastases. The altered genes mostly overlapped with changes identified in the consensus profiles of cortex and hippocampus from Alzheimer's disease and the substantia nigra from Parkinson's disease. These genes were involved in processes such as extracellular matrix organization, neuronal development, translation, and chromatin remodeling. Among them, ITGA10 and DNAJC6 emerged as consistent markers, highlighting their potential role at the intersection between neurodegeneration and tumor progression.

5. Interactive and open-access web platforms were developed to disseminate the complete results, enabling their exploration by the scientific community.

6. The findings of this thesis contribute to a better understanding of sex-differential mechanisms in multiple sclerosis and the neurobiology of melanoma brain metastases. They may provide a foundation for future research to expand this knowledge and improve the study of central nervous system disorders.

8. Scientific contributions

Scientific contributions associated with the period of this doctoral dissertation:

Journal publications

1. **Irene Soler-Sáez,** Alcida Karz, Marta R Hidalgo, Borja Gómez-Cabañes, Adolfo López-Cerdán, José F Català-Senent, Kylie Prutisto-Chang, Nicole M Eskow, Benjamin Izar, Torben Redmer, Swaminathan Kumar, Michael A Davies, María de la Iglesia-Vayá, Eva Hernando and Francisco García-García. *Unveiling Common Transcriptomic Features between Melanoma Brain Metastases and Neurodegenerative Diseases.* May 2025. DOI: https://doi.org/10.1016/j.jid.2024.09.005

2. Leire Izagirre-Urizar, Luna Mora-Huerta, **Irene Soler-Saez**, Raquel Morales-Gallel, Maria-Jose Ulloa-Navas, Juan-Carlos Chara, Stefano Calovi, Cyrille Deboux, Laura Merino-Cacho, Citlalli Netzahualcoyotzi, Maria Domercq, José L. Zugaza, Luc Pellerin, Francisco Garcia-Garcia, Jose-Manuel Garcia-Verdugo, Carlos Matute, Brahim Nait-Oumesmar and Vanja Tepavcevic. *Monocarboxylate transporter 2 is required for the maintenance of myelin and axonal integrity by oligodendrocytes.* bioRxiv (preprint). January 2025. DOI: https://doi.org/10.1101/2025.01.10.632306

3. Fernando Gordillo-González, **Irene Soler-Sáez**, Cristina Galiana-Roselló, Marta R. Hidalgo, Borja Gómez-Cabañes, Rubén Grillo-Risco, Beatriz Dolader-Rabinad, Natalia del Rey Díez, Alejandro Virués-Morales, Natalia Yanguas-Casás, Franc Casanova Ferrer and Francisco García-García. *Uncovering sex differences in Parkinson's Disease through metaanalysis of single cell transcriptomic studies*. bioRxiv (preprint). December 2024. DOI: https://doi.org/10.1101/2024.12.20.628852

4. Adolfo López-Cerdán, Zoraida Andreu, Marta R Hidalgo, **Irene Soler-Sáez**, María de la Iglesia-Vayá, Akiko Mikozami, Franca R Guerini and Francisco García-García. *An integrated approach to identifying sex-specific genes, transcription factors, and pathways relevant to Alzheimer's disease*. September 2024. DOI: https://doi.org/10.1016/j.nbd.2024.106605

5. **Irene Soler-Sáez**, Borja Gómez-Cabañes, Rubén Grillo-Risco, Cristina Galiana-Roselló, Lucas Barea-Moya, Héctor Carceller, María de la Iglesia-Vayá, Sara Gil-Perotin, Vanja Tepavčević, Marta R. Hidalgo and Francisco García-García. *Single cell landscape of sex differences in the progression of multiple sclerosis*. bioRxiv (preprint). June 2024. DOI: https://doi.org/10.1101/2024.06.15.599139

6. Jaime Llera-Oyola, Héctor Carceller, Zoraida Andreu, Marta R. Hidalgo, **Irene Soler-Sáez**, Fernando Gordillo, Borja Gómez-Cabañes, Beatriz Roson, Maria de la Iglesia-Vayá, Roberta Mancuso, Franca R. Guerini, Akiko Mizokami and Francisco García-García. *The role of microRNAs in understanding sex-based differences in Alzheimer's disease.* Biology of sex Differences. January /2024. DOI: https://doi.org/10.1186/s13293-024-00588-1

7. José Francisco Català-Senent, Zoraida Andreu, Marta R Hidalgo, **Irene Soler-Sáez**, Francisco José Roig, Natalia Yanguas-Casás, Almudena Neva-Alejo, Adolfo López-Cerdán, María de la Iglesia-Vayá, Barbara E Stranger and Francisco García-García. *A deep transcriptome meta-analysis reveals molecular mechanisms sex-based differences in Multiple Esclerosis.* Neurobiology of Disease. June 2023. DOI: https://doi.org/10.1016/j.nbd.2023.106113

8. Adolfo López-Cerdán, Zoraida Andreu, Marta R. Hidalgo, Rubén Grillo-Risco, José Francisco Català-Senent, **Irene Soler-Sáez**, Almudena Neva-Alejo, Fernando Gordillo, María de la Iglesia-Vayá and Francisco García-García. *Unveiling sex-based differences in Parkinson Disease: a comprehensive meta-analysis of transcriptomic studies.* Biology of Sex Differences. November 2022. DOI: https://doi.org/10.1186/s13293-022-00477-5

9. Inés Rivero, Guillermo Jorge Gorines-Cordero, Luis A. Rubio-Rodríguez, **Irene Soler-Sáez**, Carla Perpiñá-Clérigues, Adrian García, Sara Monzón and Tamara Hernández-Beeftink. *ISCB RSG-Spain and highlights from the VIII Spanish Student Symposium in Bioinformatics and Computational Biology in 2021*. bioRxiv (preprint). August 2022. DOI: https://doi.org/10.1101/2022.08.19.504447

National and international conference presentations

1. *Gut microbiota variation modulates polyamine production in the context of multiple sclerosis.* Conference: The Barcelona Debates on The Human Microbiome. Date: June 2025. Authors: **Irene Soler-Sáez**, Gwen Falony, Georg Bündgen, Yong Fan, Oluf Pedersen, Tobias Bopp, Francisco García-García and Sara Vieira-Silva. Type of presentation: highlighted oral communication and poster

2. *Unveiling shared transcriptomic profiles of melanoma brain metastases and the neurodegenerative disorders Alzheimer's, Parkinson's and multiple sclerosis.* Conference: 1er Congreso de la Sociedad Española de Bioinformática y Biología Computacional. Date: October 2024. Authors: **Irene Soler-Sáez**, Alcida Karz, Marta R. Hidalgo, Borja Gómez-Cabañes, Adolfo López-Cerdán, José F. Català-Senent, María de la Iglesia-Vayá, Eva Hernando, Francisco García-García. Type of presentation: flash talk.

3. *Unveiling shared transcriptomic profiles of melanoma brain metastases and the neurodegenerative disorders Alzheimer's, Parkinson's and multiple sclerosis.* Conference: IX national student symposium of bioinformatics and computational biology. Date: October 2024. Authors: **Irene Soler-Sáez**, Alcida Karz, Marta R. Hidalgo, Borja Gómez-Cabañes, Adolfo López-Cerdán, José F. Català-Senent, María de la Iglesia-Vayá, Eva Hernando, Francisco García-García. Type of presentation: oral communication. Prize: second best oral communication.

4. *Shared signature of melanoma brain metastasis and Alzheimer's disease.* Conference: 5th Latin American Student Council Symposium (LA-SCS). Date: November 2023. Authors: **Irene Soler-Sáez,** Alcida Karz, Adolfo López-Cerdán, José Francisco Català-Senent, Miriam Poley-Gil, Antonio Porlán-Mingarro, Helena Gómez-Martínez, Macarena Pozo-Morales, María de la Iglesia-Vayá, Marta R. Hidalgo, Eva Hernando, Francisco García-García. Type of presentation: poster

5. *Common hallmarks between melanoma brain metastases and Alzheimer's disease.* Conference: Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB) 2023. Date: July 2023. Authors: **Irene Soler-Sáez**, Alcida Karz, Adolfo López-Cerdán, José F. Catalá-Senent, Gonzalo Antón-Bernat, María de la Iglesia-Vayá, Marta R. Hidalgo, Eva Hernando-Monge, Francisco García-García. Type of presentation: poster

6. *Molecular and functional atlas of sex-differences in secondary-progressive multiple sclerosis: neurons characterization.* Conference: Women in Bioinformatics and Data Science LA. Date: September 2022. Authors: **Irene Soler-Sáez**, Zoraida Andreu, José Francisco Català-Senent, Rubén Grillo-Risco, Adolfo López-Cerdán, Almudena Neva-Alejo, Borja Gómez-Cabañes, Héctor Carceller, María de la Iglesia-Vayá, Marta R. Hidalgo-García, Francisco García-García. Type of presentation: poster

7. *Molecular and functional atlas of sex-differences in multiple sclerosis subtypes analising single cell and single nucleus transcriptomic data.* Conference: 21st European Conference on Computational Biologý. Date: September 2022. Authors: **Irene Soler-Sáez**, Zoraida Andreu, José Francisco Català-Senent, Rubén Grillo-Risco, Adolfo López-Cerdán, Almudena Neva-Alejo, Borja Gómez-Cabañes, Héctor Carceller, María de la Iglesia-Vayá, Marta R. Hidalgo-García, Francisco García-García. Type of presentation: poster

8. *Molecular and functional atlas of sex-differences in multiple sclerosis subtypes analising single cell and single nucleus transcriptomic data.* Conference: 7th European Student Council Symposium (ESCS2022). Date: September 2022. Authors: **Irene Soler-Sáez**, Zoraida Andreu, José Francisco Català-Senent, Rubén Grillo-Risco, Adolfo López-Cerdán, Almudena Neva-Alejo, Borja Gómez-Cabañes, Héctor Carceller, María de la Iglesia-Vayá, Marta R. Hidalgo-García, Francisco García-García. Type of presentation: poster

9. *Atlas funcional de las diferencias de sexo en esclerosis múltiple mediante el análisis de datos transcriptómicos de célula única.* Conference: I Simposio de Estudiantes Hispanohablantes de Bioinformática y Biología Computacional (SEH2Bioinfo). Date: June 2022. Authors: **Irene Soler-Sáez**, Zoraida Andreu, José Francisco Català-Senent, Rubén Grillo-Risco, Adolfo López-Cerdán, Aludena Neva-Alejo, Héctor Carceller, María de la Iglesia-Vayá, Marta R. Hidalgo and Francisco García-García. Type of presentation: poster.

Teaching activities

1. Bachelor's Double Degree in Biotechnology and Agricultural Engineering. Institution: Polytechnic University of Valencia. Subject: Genetic Engineering. Date: academic years 2022-2023 and 2023-2024. Specific lessons: laboratory practicals.

2. Bachelor's Degree in Biotechnology. Institution: Polytechnic University of Valencia. Subject: Culture of animal cells and tissues. Date: academic years 2022–2023 and 2023–2024. Specific lessons: laboratory practicals.

3. External lecturer in Master's Degree in Bioinformatics. Institution: University of Valencia. Course: In silico Studies in Biomedicine. Date: from 2022 to present. Specific lessons: RNA-seq data analysis, scRNA-seq data analysis, signaling pathway analysis using transcriptomic data.

4. External lecturer in Master's Degree in Bioinformatics Applied to Personalized Medicine and Health. Institution: National School of Public Health, Carlos III Health Institute. Subject: Omics Data Analysis and Interpretation. Date: 2025. Specific lessons: Over-representation analysis; classifiers and predictors based on omics data.

5. CIPF Bioinformatics Training Courses. Institution: Principe Felipe Research Center. Date: July 2024. Specific classes: Analysis of Single-cell RNA-seq Data Using Web Tools.
6. Course Web-based Omics Data Analysis in Cardiovascular Diseases (WODA-CVD). Institution: Principe Felipe Research Center. Date: December 2022. Specific classes: Single-cell RNA-seq analysis.
7. Master's Degree in Bioinformatics, Computational Biology and Personalized Medicine.Institution: Polytechnic University of Valencia. Subject: Genomic Data Analysis. Date: 2022. Specific classes: RNA-seq data analysis.

Outreach activities

Scientific Training and Mentoring Committee. Principe Felipe Research Center. From January 2022 to date.

RSG-Spain active member. From 2021 to date. RSG-Spain is community for students interested in bioinformatics and computational biology, promoting diffusion, learning, and networking organizing online and in-person activities. National Secretary (2023-2024), National President (2024-2025).

Conference organization

1. National conference: X national student symposium of bioinformatics and computational biology. October 2025. Chair.
2. National conference: IX national student symposium of bioinformatics and computational biology. October 2024. Scientific committee.
3. International conference: 7th European Student Council Symposium (ESCS2022). September 2022. Finance and outreach committees.
4. International conference: I Simposio de Estudiantes Hispanohablantes de Bioinformática y Biología Computacional (SEH2Bioinfo). June 2022. Social media committee.

9. Bibliography

1. Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B. & Ciccarelli, O. Multiple sclerosis. *The Lancet* **391**, 1622–1636 (2018).
2. Ford, H. Clinical presentation and diagnosis of multiple sclerosis. *Clin. Med.* **20**, 380–383 (2020).
3. Baecher-Allan, C., Kaskow, B. J. & Weiner, H. L. Multiple Sclerosis: Mechanisms and Immunotherapy. *Neuron* **97**, 742–768 (2018).
4. Dobson, R. & Giovannoni, G. Multiple sclerosis - a review. *Eur. J. Neurol.* **26**, 27–40 (2019).
5. Walton, C. *et al.* Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult. Scler. J.* **26**, 1816–1821 (2020).
6. Leray, E., Moreau, T., Fromont, A. & Edan, G. Epidemiology of multiple sclerosis. *Rev. Neurol. (Paris)* **172**, 3–13 (2016).
7. Dymecka, J., Gerymski, R., Tataruch, R. & Bidzan, M. Fatigue, Physical Disability and Self-Efficacy as Predictors of the Acceptance of Illness and Health-Related Quality of Life in Patients with Multiple Sclerosis. *Int. J. Environ. Res. Public. Health* **18**, 13237 (2021).
8. Pintér, A., Cseh, D., Sárközi, A., Illigens, B. M. & Siepmann, T. Autonomic Dysregulation in Multiple Sclerosis. *Int. J. Mol. Sci.* **16**, 16920–16952 (2015).
9. Rao, S. M., Leo, G. J., Bernardin, L. & Unverzagt, F. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. *Neurology* **41**, 685–691 (1991).
10. Paz-Zulueta, M., Parás-Bravo, P., Cantarero-Prieto, D., Blázquez-Fernández, C. & Oterino-Durán, A. A literature review of cost-of-illness studies on the economic burden of multiple sclerosis. *Mult. Scler. Relat. Disord.* **43**, (2020).
11. Purves, D., *et al.* Neuroscience. 5th Edition. *Sinauer Associates* (2012). ISBN: (Hardcover) 978-0878936953.
12. Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. Immunobiology. 5th edition. *Garland Science* (2001). ISBN-10: 0-8153-3642-X
13. Jakimovski, D. *et al.* Multiple sclerosis. *The Lancet* **403**, 183–202 (2024).
14. Lanz, T. V. *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature* **603**, 321–327 (2022).
15. Bjornevik, K. *et al.* Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* **375**, 296–301 (2022).
16. Vietzen, H. *et al.* Ineffective control of Epstein-Barr-virus-induced autoimmunity increases the risk for multiple sclerosis. *Cell* **186**, 5705-5718.e13 (2023).
17. Ganta, K. K. *et al.* Acute infectious mononucleosis generates persistent, functional EBNA-1 antibodies with high cross-reactivity to alpha-crystalline beta. *Cell Rep.* **44**, 115709 (2025).
18. Liu, R. *et al.* Autoreactive lymphocytes in multiple sclerosis: Pathogenesis and treatment target. *Front. Immunol.* **13**, 996469 (2022).
19. van Langelaar, J., Rijvers, L., Smolders, J. & van Luijn, M. M. B and T Cells Driving Multiple Sclerosis: Identity, Mechanisms and Potential Triggers. *Front. Immunol.* **11**, (2020).
20. Lerma-Martin, C. *et al.* Cell type mapping reveals tissue niches and interactions in subcortical multiple sclerosis lesions. *Nat. Neurosci.* **27**, 2354–2365 (2024).
21. Lassmann, H. Multiple Sclerosis Pathology. *Cold Spring Harb. Perspect. Med.* **8**, a028936 (2018).
22. Bagnato, F. *et al.* Imaging chronic active lesions in multiple sclerosis: a consensus statement. *Brain* **147**, 2913–2933 (2024).
23. Siger, M. Magnetic Resonance Imaging in Primary Progressive Multiple Sclerosis Patients. *Clin. Neuroradiol.* **32**, 625–641 (2022).
24. Lublin, F. D. & Reingold, S. C. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology* **46**, 907–911 (1996).

25. Lublin, F. D. *et al.* Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology* **83**, 278–286 (2014).
26. Klineova, S. & Lublin, F. D. Clinical Course of Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **8**, a028928 (2018).
27. Krieger, S., Cook, K. & Hersh, C. M. Understanding multiple sclerosis as a disease spectrum: above and below the clinical threshold. *Curr. Opin. Neurol.* **37**, 189–201 (2024).
28. McGinley, M. P., Goldschmidt, C. H. & Rae-Grant, A. D. Diagnosis and Treatment of Multiple Sclerosis: A Review. *JAMA* **325**, 765–779 (2021).
29. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018).
30. Solomon, A. J. *et al.* Differential diagnosis of suspected multiple sclerosis: an updated consensus approach. *Lancet Neurol.* **22**, 750–768 (2023).
31. Hauser, S. L. & Cree, B. A. C. Treatment of Multiple Sclerosis: A Review. *Am. J. Med.* **133**, 1380-1390.e2 (2020).
32. Gajofatto, A. & Benedetti, M. D. Treatment strategies for multiple sclerosis: When to start, when to change, when to stop? *World J. Clin. Cases WJCC* **3**, 545–555 (2015).
33. Morrow, S. A. *et al.* Use of natalizumab in persons with multiple sclerosis: 2022 update. *Mult. Scler. Relat. Disord.* **65**, 103995 (2022).
34. McGinley, M. P. & Cohen, J. A. Sphingosine 1-phosphate receptor modulators in multiple sclerosis and other conditions. *Lancet Lond. Engl.* **398**, 1184–1194 (2021).
35. Giovannoni, G. & Mathews, J. Cladribine Tablets for Relapsing-Remitting Multiple Sclerosis: A Clinician's Review. *Neurol. Ther.* **11**, 571–595 (2022).
36. Wingerchuk, D. M. & Carter, J. L. Multiple sclerosis: current and emerging disease-modifying therapies and treatment strategies. *Mayo Clin. Proc.* **89**, 225–240 (2014).
37. Lamb, Y. N. Ocrelizumab: A Review in Multiple Sclerosis. *Drugs* **82**, 323–334 (2022).
38. Scott, L. J. Siponimod: A Review in Secondary Progressive Multiple Sclerosis. *CNS Drugs* **34**, 1191–1200 (2020).
39. Chataway, J. *et al.* Clinical trials for progressive multiple sclerosis: progress, new lessons learned, and remaining challenges. *Lancet Neurol.* **23**, 277–301 (2024).
40. Sriwastava, S. *et al.* Recent advances in the treatment of primary and secondary progressive Multiple Sclerosis. *J. Neuroimmunol.* **390**, 578315 (2024).
41. Torkildsen, Ø., Myhr, K.-M., Brugger-Synnes, P. & Bjørnevik, K. Antiviral therapy with tenofovir in MS. *Mult. Scler. Relat. Disord.* **83**, 105436 (2024).
42. Maple, P. A. *et al.* The Potential for EBV Vaccines to Prevent Multiple Sclerosis. *Front. Neurol.* **13**, 887794 (2022).
43. Campagnoli, L. I. M. *et al.* New therapeutic avenues in multiple sclerosis: Is there a place for gut microbiota-based treatments? *Pharmacol. Res.* **209**, 107456 (2024).
44. Hsu, S. & Bove, R. Hormonal Therapies in Multiple Sclerosis: a Review of Clinical Data. *Curr. Neurol. Neurosci. Rep.* **24**, 1–15 (2024).
45. DeLuca, J., Chiaravalloti, N. D. & Sandroff, B. M. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. *Nat. Rev. Neurol.* **16**, 319–332 (2020).
46. Haki, M., Al-Biati, H. A., Al-Tameemi, Z. S., Ali, I. S. & Al-Hussaniy, H. A. Review of multiple sclerosis: Epidemiology, etiology, pathophysiology, and treatment. *Medicine (Baltimore)* **103**, e37297 (2024).
47. Waubant, E. *et al.* Environmental and genetic risk factors for MS: an integrated review. *Ann. Clin. Transl. Neurol.* **6**, 1905–1922 (2019).
48. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

49. Baranzini, S. E. & Oksenberg, J. R. The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. *Trends Genet. TIG* **33**, 960–970 (2017).
50. Alfredsson, L. & Olsson, T. Lifestyle and Environmental Factors in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **9**, a028944 (2019).
51. Puthenparampil, M. *et al.* Multiple sclerosis epidemiological trends in Italy highlight the environmental risk factors. *J. Neurol.* **269**, 1817–1824 (2022).
52. Forsyth, K. S., Jiwrajka, N., Lovell, C. D., Toothacre, N. E. & Anguera, M. C. The conneXion between sex and immune responses. *Nat. Rev. Immunol.* **24**, 487–502 (2024).
53. Young, J. E., Wu, M. & Hunsberger, H. C. Editorial: Sex and gender differences in neurodegenerative diseases. *Front. Neurosci.* **17**, 1175674 (2023).
54. Mauvais-Jarvis, F. *et al.* Sex and gender: modifiers of health, disease, and medicine. *The Lancet* **396**, 565–582 (2020).
55. Regitz-Zagrosek, V. & Gebhard, C. Gender medicine: effects of sex and gender on cardiovascular disease manifestation and outcomes. *Nat. Rev. Cardiol.* **20**, 236–247 (2023).
56. Ortona, E. *et al.* Sex-based differences in autoimmune diseases. *Ann. Ist. Super. Sanita* **52**, 205–212 (2016).
57. D'hooghe, M. B., D'Hooghe, T. & De Keyser, J. Female gender and reproductive factors affecting risk, relapses and progression in multiple sclerosis. *Gynecol. Obstet. Invest.* **75**, 73–84 (2013).
58. Voskuhl, R. R. *et al.* Sex differences in brain atrophy in multiple sclerosis. *Biol. Sex Differ.* **11**, 49 (2020).
59. B, T. *et al.* Sex-specific differences in rim appearance of multiple sclerosis lesions on quantitative susceptibility mapping. *Mult. Scler. Relat. Disord.* **45**, 102317 (2020).
60. Leavitt, V. M., Dworkin, J. D., Kalina, T. & Ratzan, A. S. Sex differences in brain resilience of individuals with multiple sclerosis. *Mult. Scler. Relat. Disord.* **87**, 105646 (2024).
61. Golden, L. C. & Voskuhl, R. The importance of studying sex differences in disease: The example of multiple sclerosis. *J. Neurosci. Res.* **95**, 633–643 (2017).
62. Kalincik, T. *et al.* Sex as a determinant of relapse incidence and progressive course of multiple sclerosis. *Brain J. Neurol.* **136**, 3609–3617 (2013).
63. Koch, M., Kingwell, E., Rieckmann, P., Tremlett, H. & Neurologists, U. M. C. The natural history of secondary progressive multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* **81**, 1039–1043 (2010).
64. Li, R. *et al.* Sex differences in outcomes of disease-modifying treatments for multiple sclerosis: A systematic review. *Mult. Scler. Relat. Disord.* **12**, 23–28 (2017).
65. Houtchens, M. K. & Bove, R. A case for gender-based approach to multiple sclerosis therapeutics. *Front. Neuroendocrinol.* **50**, 123–134 (2018).
66. Alvarez-Sanchez, N. & Dunn, S. E. Potential biological contributers to the sex difference in multiple sclerosis progression. *Front. Immunol.* **14**, 1175874 (2023).
67. Carruth, L., Reisert, I. & Arnold, A. Sex chromosome genes directly affect brain sexual differentiation. *Nat Neurosci* **5**, 933–934 (2002).
68. Arnold, A. P. & Chen, X. What does the "four core genotypes" mouse model tell us about sex differences in the brain and other tissues? *Front. Neuroendocrinol.* **30**, 1–9 (2009).
69. Moore, S., Patel, R., Hannsun, G., Yang, J. & Tiwari-Woodruff, S. K. Sex chromosome complement influences functional callosal myelination. *Neuroscience* **245**, 166–178 (2013).
70. Doss, P. M. I. A. *et al.* Male sex chromosomal complement exacerbates the pathogenicity of Th17 cells in a chronic model of central nervous system autoimmunity. *Cell Rep.* **34**, (2021).
71. Voskuhl, R. R., Sawalha, A. H. & Itoh, Y. Sex chromosome contributions to sex differences in multiple sclerosis susceptibility and progression. *Mult. Scler. J.* **24**, 22–31 (2018).
72. Murgia, F. *et al.* Sex Hormones as Key Modulators of the Immune Response in Multiple Sclerosis: A Review. *Biomedicines* **10**, 3107 (2022).

73. Confavreux, C., Hutchinson, M., Hours, M. M., Cortinovis-Tourniaire, P. & Moreau, T. Rate of pregnancy-related relapse in multiple sclerosis. Pregnancy in Multiple Sclerosis Group. *N. Engl. J. Med.* **339**, 285–291 (1998).
74. Avila, M., Bansal, A., Culberson, J. & Peiris, A. N. The Role of Sex Hormones in Multiple Sclerosis. *Eur. Neurol.* **80**, 93–99 (2018).
75. Bove, R. & Chitnis, T. The role of gender and sex hormones in determining the onset and outcome of multiple sclerosis. *Mult. Scler. Houndmills Basingstoke Engl.* **20**, 520–526 (2014).
76. Keane, J. T. *et al.* Gender and the Sex Hormone Estradiol Affect Multiple Sclerosis Risk Gene Expression in Epstein-Barr Virus-Infected B Cells. *Front. Immunol.* **12**, (2021).
77. Shayestehfar, M. *et al.* Sex hormone therapy in Multiple Sclerosis: A systematic review of randomized clinical trials. *J. Cent. Nerv. Syst. Dis.* **16**, 11795735231223411 (2024).
78. Krysko, K. M. *et al.* Sex effects across the lifespan in women with multiple sclerosis. *Ther. Adv. Neurol. Disord.* **13**, 1756286420936166 (2020).
79. Català-Senent, J. F. *et al.* A deep transcriptome meta-analysis reveals sex differences in multiple sclerosis. *Neurobiol. Dis.* **181**, 106113 (2023).
80. Itoh, N., Itoh, Y., Stiles, L. & Voskuhl, R. Sex differences in the neuronal transcriptome and synaptic mitochondrial function in the cerebral cortex of a multiple sclerosis model. *Front. Neurol.* **14**, (2023).
81. Magistretti, P. J. & Allaman, I. A Cellular Perspective on Brain Energy Metabolism and Functional Imaging. *Neuron* **86**, 883–901 (2015).
82. Gadhave, D. G. *et al.* Neurodegenerative disorders: Mechanisms of degeneration and therapeutic approaches with their clinical relevance. *Ageing Res. Rev.* **99**, 102357 (2024).
83. Salvador, A. F. M. & Kipnis, J. Immune response after central nervous system injury. *Semin. Immunol.* **59**, 101629 (2022).
84. Johnson, H. J. & Koshy, A. A. Understanding neuroinflammation through central nervous system infections. *Curr. Opin. Neurobiol.* **76**, 102619 (2022).
85. Grinda, T., Aizer, A. A., Lin, N. U. & Sammons, S. L. Central Nervous System Metastases in Breast Cancer. *Curr. Treat. Options Oncol.* **26**, 14–35 (2025).
86. Schaff, L. R. & Grommes, C. Primary central nervous system lymphoma. *Blood* **140**, 971–979 (2022).
87. Steinmetz, J. D. *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.* **23**, 344–381 (2024).
88. Mitchell, A. *et al.* Estimating the Economic Impact of Direct Health Expenditure on Brain Disorders, Globally and in the United States (P11-4.009). *Neurology* **102**, 6468 (2024).
89. Coleman, M. P. & Höke, A. Programmed axon degeneration: from mouse to mechanism to medicine. *Nat. Rev. Neurosci.* **21**, 183–196 (2020).
90. Coleman, M. Axon degeneration mechanisms: commonality amid diversity. *Nat. Rev. Neurosci.* **6**, 889–898 (2005).
91. Wilson, D. M. *et al.* Hallmarks of neurodegenerative diseases. *Cell* **186**, 693–714 (2023).
92. Kang, G., Moo, E. K., Banton, R., Petel, O. E. & Harris, A. R. Cellular mechanisms of traumatic brain injury. *Npj Biol. Phys. Mech.* **2**, 16 (2025).
93. Swanson, P. A. & McGavern, D. B. Viral Diseases of the Central Nervous System. *Curr. Opin. Virol.* **11**, 44–54 (2015).
94. Osswald, M. *et al.* Brain tumour cells interconnect to a functional and resistant network. *Nature* **528**, 93–98 (2015).
95. Kölliker-Frers, R. *et al.* Neuroinflammation: An Integrating Overview of Reactive-Neuroimmune Cell Interactions in Health and Disease. *Mediators Inflamm.* **2021**, 9999146 (2021).
96. Heneka, M. T. *et al.* Neuroinflammation in Alzheimer disease. *Nat. Rev. Immunol.* **25**, 321–352 (2025).

97. Xiong, Y., Mahmood, A. & Chopp, M. Current understanding of neuroinflammation after traumatic brain injury and cell-based therapeutic opportunities. *Chin. J. Traumatol.* **21**, 137–151 (2018).

98. Tohidpour, A. *et al.* Neuroinflammation and Infection: Molecular Mechanisms Associated with Dysfunction of Neurovascular Unit. *Front. Cell. Infect. Microbiol.* **7**, 276 (2017).

99. Wen, J., Liu, D., Zhu, H. & Shu, K. Microenvironmental regulation of tumor-associated neutrophils in malignant glioma: from mechanism to therapy. *J. Neuroinflammation* **21**, 226 (2024).

100. Doron, H., Pukrop, T. & Erez, N. A Blazing Landscape: Neuroinflammation Shapes Brain Metastasis. *Cancer Res.* **79**, 423–436 (2019).

101. Scheltens, P. *et al.* Alzheimer's disease. *The Lancet* **397**, 1577–1590 (2021).

102. De Strooper, B. & Karran, E. The Cellular Phase of Alzheimer's Disease. *Cell* **164**, 603–615 (2016).

103. Long, J. M. & Holtzman, D. M. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell* **179**, 312–339 (2019).

104. Peng, S. *et al.* Global, regional and national burden of Parkinson's disease in people over 55 years of age: a systematic analysis of the global burden of disease study. *BMC Neurol* **178**, 1991–2021 (2025).

105. Kalia, L. V. & Lang, A. E. Parkinson's disease. *Lancet Lond. Engl.* **386**, 896–912 (2015).

106. Dorsey, E. R. & Bloem, B. R. The Parkinson Pandemic-A Call to Action. *JAMA Neurol.* **75**, 9–10 (2018).

107. Long, G. V., Swetter, S. M., Menzies, A. M., Gershenwald, J. E. & Scolyer, R. A. Cutaneous melanoma. *Lancet Lond. Engl.* **402**, 485–502 (2023).

108. Abbasi, N. R. *et al.* Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA* **292**, 2771–2776 (2004).

109. Rebecca, V. W., Sondak, V. K. & Smalley, K. S. M. A Brief History of Melanoma: From Mummies to Mutations. *Melanoma Res.* **22**, 114–122 (2012).

110. Internò, V. *et al.* Melanoma Brain Metastases: A Retrospective Analysis of Prognostic Factors and Efficacy of Multimodal Therapies. *Cancers* **15**, 1542 (2023).

111. Vailati-Riboni, M., Palombo, V. & Loor, J. J. Periparturient Diseases of Dairy Cows: A Systems Biology Approach, Chapter: What Are Omics Sciences? *Springer International Publishing.* (2017). Online ISBN: 978-3-319-43033-1

112. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* **8**, (2017).

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

114. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

115. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, e1005457 (2017).

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

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

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

119. Jin, J., Liu, X. & Shiroguchi, K. Long journey of 16S rRNA-amplicon sequencing toward cell-based functional bacterial microbiota characterization. *iMetaOmics* **1**, e9 (2024).

120. Pérez-Cobas, A. E., Gomez-Valero, L. & Buchrieser, C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microb. Genomics* **6**, mgen000409 (2020).

121. Shahi, S. K., Freedman, S. N. & Mangalam, A. K. Gut microbiome in multiple sclerosis: The players involved and the roles they play. *Gut Microbes* **8**, 607–615 (2017).

122. All about data sharing. *Nat. Cancer* **2**, 475–475 (2021).

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

124. Numajiri, H. & Hayashi, T. Analysis on open data as a foundation for data-driven research. *Scientometrics* **129**, 6315–6332 (2024).

125. Digital Science et al. The State of Open Data 2023. https://doi.org/10.6084/m9.figshare.24428194.v1 (2023).

126. Hahnel, M., Smith, G. & Campbell, A. The State of Open Data 2024: Special Report Bridging Policy and Practice in Data Sharing. https://doi.org/10.6084/m9.figshare.27337476.v2 (2024).

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

128. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99 (2013).

129. Schneider, M. V. & Orchard, S. Omics technologies, data and bioinformatics principles. *Methods Mol. Biol.* **719**, 3–30 (2011).

130. Peccoud, J. Data sharing policies: share well and you shall be rewarded. *Synth. Biol. Oxf. Engl.* **6**, ysab028 (2021).

131. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).

132. Sarkans, U. *et al.* From ArrayExpress to BioStudies. *Nucleic Acids Res.* **49**, D1502–D1506 (2020).

133. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).

134. O'Cathail, C. *et al.* The European Nucleotide Archive in 2024. *Nucleic Acids Res.* **53**, D49–D55 (2024).

135. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* **6**, e1000097 (2009).

136. Page, M. J. *et al.* PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **372**, n160 (2021).

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

138. Polanin, J. R., Pigott, T. D., Espelage, D. L. & Grotpeter, J. K. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Res. Synth. Methods* **10**, 330–342 (2019).

139. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).

140. Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R. & Lagisz, M. Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environ. Evid.* **12**, 8 (2023).

141. Toro-Domínguez, D. *et al.* A survey of gene expression meta-analysis: methods and applications. *Brief. Bioinform.* **22**, 1694–1705 (2021).

142. Wang, X. *et al.* A brief introduction of meta-analyses in clinical practice and research. *J. Gene Med.* **23**, e3312 (2021).

143. Kleffman, K. *et al.* Melanoma-Secreted Amyloid Beta Suppresses Neuroinflammation and Promotes Brain Metastasis. *Cancer Discov.* **12**, 1314–1335 (2022).

144. Ellen, O. *et al.* The Heterogeneous Multiple Sclerosis Lesion: How Can We Assess and Modify a Degenerating Lesion? *Int. J. Mol. Sci.* **24**, 11112 (2023).

145. Caire, M. J., Reddy, V. & Varacallo, M. A. Physiology, Synapse. *StatPearls* (2023). Available from: https://www.ncbi.nlm.nih.gov/books/NBK526047/

146. Jürgens, T. *et al.* Reconstruction of single cortical projection neurons reveals primary spine loss in multiple sclerosis. *Brain* **139**, 39–46 (2016).

147. Kapell, H. *et al.* Neuron-oligodendrocyte potassium shuttling at nodes of Ranvier protects against inflammatory demyelination. *J. Clin. Invest.* **133**, (2023).

148. Barakat-Alrashdi, A., Dawod, B., Schampel, A., Tacke, S., Kuerten, S., Marshall, J. S. *et al.* Nav1.6 promotes inflammation and neuronal degeneration in a mouse model of multiple sclerosis. *J. Neuroinflammation* **16**, 215 (2019).

149. Schattling, B. *et al.* Activity of Nav1.2 promotes neurodegeneration in an animal model of multiple sclerosis. *JCI Insight* **1**, (2016).

150. Macrez, R., Stys, P. K., Vivien, D., Lipton, S. A. & Docagne, F. Mechanisms of glutamate toxicity in multiple sclerosis: biomarker and therapeutic opportunities. *Lancet Neurol.* **15**, 1089–1102 (2016).

151. Pitt, D., Werner, P. & Raine, C. S. Glutamate excitotoxicity in a model of multiple sclerosis. *Nat. Med.* **6**, 67–70 (2000).

152. Woo, M. S., Engler, J. B. & Friese, M. A. The neuropathobiology of multiple sclerosis. *Nat. Rev. Neurosci.* **25**, 493–513 (2024).

153. Azevedo, C. J. *et al.* In vivo evidence of glutamate toxicity in multiple sclerosis. *Ann. Neurol.* **76**, 269–278 (2014).

154. Rossi, S. *et al.* Impaired striatal GABA transmission in experimental autoimmune encephalomyelitis. *Brain. Behav. Immun.* **25**, 947–956 (2011).

155. Kiljan, S. *et al.* Enhanced GABAergic Immunoreactivity in Hippocampal Neurons and Astroglia of Multiple Sclerosis Patients. *J. Neuropathol. Exp. Neurol.* **78**, 480–491 (2019).

156. Madsen, M. A. *et al.* Association of Cortical Lesions With Regional Glutamate, GABA, N-Acetylaspartate, and Myoinositol Levels in Patients With Multiple Sclerosis. *Neurology* **103**, e209543 (2024).

157. Zhang, L., Verkhratsky, A. & Shi, F.-D. Astrocytes and microglia in multiple sclerosis and neuromyelitis optica. In *Handbook of Clinical Neurology*, Vol. 210, 133–145. Elsevier BV (2025).

158. Lassmann, H. Pathogenic Mechanisms Associated With Different Clinical Courses of Multiple Sclerosis. *Front. Immunol.* **9**, 3116 (2018).

159. Soroush, A. & Dunn, J. F. A Hypoxia-Inflammation Cycle and Multiple Sclerosis: Mechanisms and Therapeutic Implications. *Curr. Treat. Options Neurol.* **27**, 6 (2025).

160. Correale, J. & Farez, M. F. The Role of Astrocytes in Multiple Sclerosis Progression. *Front. Neurol.* **6**, (2015).

161. Rawji, K. S., Martinez, G. A. G., Sharma, A. & Franklin, R. J. M. The Role of Astrocytes in Remyelination. *Trends Neurosci.* **43**, 596–607 (2020).

162. Skripuletz, T. *et al.* Astrocytes regulate myelin clearance through recruitment of microglia during cuprizone-induced demyelination. *Brain J. Neurol.* **136**, 147–167 (2013).

163. Zhang, X. *et al.* Microglia in the context of multiple sclerosis. *Front. Neurol.* **14**, (2023).

164. Gluck, L., Gerstein, B. & Kaunzner, U. W. Repair mechanisms of the central nervous system: From axon sprouting to remyelination. *Neurotherapeutics* **22**, (2025).

165. Han, S., Gim, Y., Jang, E.-H. & Hur, E.-M. Functions and dysfunctions of oligodendrocytes in neurodegenerative diseases. *Front. Cell. Neurosci.* **16**, (2022).

166. Tepavčević, V. & Lubetzki, C. Oligodendrocyte progenitor cell recruitment and remyelination in multiple sclerosis: the more, the merrier? *Brain* **145**, 4178–4192 (2022).

167. Wang, Q. *et al.* Oligodendroglial precursor cells modulate immune response and early demyelination in a murine model of multiple sclerosis. *Sci. Transl. Med.* **17**(792), eadn9980 (2025).

168. Pishesha, N., Harmand, T. J. & Ploegh, H. L. A guide to antigen processing and presentation. *Nat. Rev. Immunol.* **22**, 751–764 (2022).

169. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).

170. Hemmer, B., Kerschensteiner, M. & Korn, T. Role of the innate and adaptive immune responses in the course of multiple sclerosis. *Lancet Neurol.* **14**, 406–419 (2015).

171. Jacobs, B. M., Giovannoni, G., Cuzick, J. & Dobson, R. Systematic review and meta-analysis of the association between Epstein-Barr virus, multiple sclerosis and other risk factors. *Mult. Scler. Houndmills Basingstoke Engl.* **26**, 1281–1297 (2020).

172. Jilek, S. *et al.* Strong EBV-specific CD8+ T-cell response in patients with early multiple sclerosis. *Brain J. Neurol.* **131**, 1712–1721 (2008).

173. Ascherio, A. *et al.* Epstein-Barr virus antibodies and risk of multiple sclerosis: a prospective study. *JAMA* **286**, 3083–3088 (2001).

174. Rasheed, A. & Khan, G. Epstein-Barr virus, vitamin D and the immune response: connections with consequences for multiple sclerosis. *Front. Immunol.* **15**, (2024).

175. Giovannoni, G. *et al.* The case for targeting latent and lytic Epstein-Barr virus infection in multiple sclerosis. *Brain J. Neurol.* awaf170 (2025).

176. Khan, Z., Mehan, S., Gupta, G. D. & Narula, A. S. Immune System Dysregulation in the Progression of Multiple Sclerosis: Molecular Insights and Therapeutic Implications. *Neuroscience* **548**, 9–26 (2024).

177. Domínguez-Mozo, M. I. *et al.* Mitochondrial Impairments in Peripheral Blood Mononuclear Cells of Multiple Sclerosis Patients. *Biology* **11**, 1633 (2022).

178. Acquaviva, M. *et al.* Inferring Multiple Sclerosis Stages from the Blood Transcriptome via Machine Learning. *Cell Rep. Med.* **1**, 100053 (2020).

179. Mokaram Doust Delkhah, A. Integrated transcriptomics of multiple sclerosis peripheral blood mononuclear cells explored potential biomarkers for the disease. *Biochem. Biophys. Rep.* **42**, 102022 (2025).

180. Gross, C. C. *et al.* Multiple sclerosis endophenotypes identified by high-dimensional blood signatures are associated with distinct disease trajectories. *Sci. Transl. Med.* **16**, eade8560 (2024).

181. Yang, J. H., Rempe, T., Whitmire, N., Dunn-Pirio, A. & Graves, J. S. Therapeutic Advances in Multiple Sclerosis. *Front. Neurol.* **13**, (2022).

182. Sempik, I., Dziadkowiak, E., Moreira, H., Zimny, A. & Pokryszko-Dragan, A. Primary Progressive Multiple Sclerosis—A Key to Understanding and Managing Disease Progression. *Int. J. Mol. Sci.* **25**, 8751 (2024).

183. von Essen, M. R. *et al.* Intrathecal CD8+CD20+ T Cells in Primary Progressive Multiple Sclerosis. *Neurol. Neuroimmunol. Neuroinflammation* **10**, e200140 (2023).

184. Holloman, J. P., Axtell, R. C., Monson, N. L. & Wu, G. F. The Role of B Cells in Primary Progressive Multiple Sclerosis. *Front. Neurol.* **12**, (2021).

185. Canto-Gomes, J. *et al.* People with Primary Progressive Multiple Sclerosis Have a Lower Number of Central Memory T Cells and HLA-DR+ Tregs. *Cells* **12**, 439 (2023).

186. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).

187. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

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

189. Burel, J. G. & Peters, B. Discovering transcriptional signatures of disease for diagnosis versus mechanism. *Nat. Rev. Immunol.* **18**, 289–290 (2018).

190. Qi, G. *et al.* Single-cell allele-specific expression analysis reveals dynamic and cell-type-specific regulatory effects. *Nat. Commun.* **14**, 6317 (2023).

191. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

192. The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

193. Wick, H. C. *et al.* DFLAT: functional annotation for human development. *BMC Bioinformatics* **15**, 45 (2014).

194. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

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

196. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

197. Amadoz, A., Hidalgo, M. R., Çubuk, C., Carbonell-Caballero, J. & Dopazo, J. A comparison of mechanistic signaling pathway activity analysis methods. *Brief. Bioinform.* **20**, 1655–1668 (2019).

198. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

199. Jaakkola, M. K. & Elo, L. L. Empirical comparison of structure-based pathway methods. *Brief. Bioinform.* **17**, 336–345 (2016).

200. Liu, Z., Sun, D. & Wang, C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol.* **23**, 218 (2022).

201. Shao, X., Lu, X., Liao, J., Chen, H. & Fan, X. New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell* **11**, 866–880 (2020).

202. Du, S. *et al.* XY sex chromosome complement, compared with XX, in the CNS confers greater neurodegeneration during experimental autoimmune encephalomyelitis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2806–2811 (2014).

203. Du, L. *et al.* Innate gender-based proclivity in response to cytotoxicity and programmed cell death pathway. *J. Biol. Chem.* **279**, 38563–38570 (2004).

204. Kniffin, A. R. & Briand, L. A. Sex differences in glutamate transmission and plasticity in reward related regions. *Front. Behav. Neurosci.* **18**, (2024).

205. Yokomaku, D. *et al.* Estrogen enhances depolarization-induced glutamate release through activation of phosphatidylinositol 3-kinase and mitogen-activated protein kinase in cultured hippocampal neurons. *Mol. Endocrinol. Baltim. Md* **17**, 831–844 (2003).

206. Pozzo Miller, L. D. & Aoki, A. Stereological analysis of the hypothalamic ventromedial nucleus. II. Hormone-induced changes in the synaptogenic pattern. *Brain Res. Dev. Brain Res.* **61**, 189–196 (1991).

207. Chisholm, N. C. & Sohrabji, F. Astrocytic response to cerebral ischemia is influenced by sex differences and impaired by aging. *Neurobiol. Dis.* **85**, 245–253 (2016).

208. Villa, A., Della Torre, S. & Maggi, A. Sexual differentiation of microglia. *Front. Neuroendocrinol.* **52**, 156–164 (2019).

209. Tariq, M. B., Lee, J. & McCullough, L. D. Sex differences in the inflammatory response to stroke. *Semin. Immunopathol.* **45**, 295–313 (2023).

210. Wiedrick, J. *et al.* Sex differences in EAE reveal common and distinct cellular and molecular components. *Cell. Immunol.* **359**, 104242 (2021).

211. Tassoni, A. *et al.* The astrocyte transcriptome in EAE optic neuritis shows complement activation and reveals a sex difference in astrocytic C3 expression. *Sci. Rep.* **9**, 10010 (2019).

212. McGill, M. M. *et al.* p38 MAP Kinase Signaling in Microglia Plays a Sex-Specific Protective Role in CNS Autoimmunity and Regulates Microglial Transcriptional States. *Front. Immunol.* **12**, 715311 (2021).

213. Mayrhofer, F. *et al.* Reduction in CD11c+ microglia correlates with clinical progression in chronic experimental autoimmune demyelination. *Neurobiol. Dis.* **161**, 105556 (2021).

214. Luchetti, S. *et al.* Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: a retrospective autopsy cohort analysis. *Acta Neuropathol. (Berl.)* **135**, 511–528 (2018).

215. Frischer, J. M. *et al.* Clinical and pathological insights into the dynamic nature of the white matter multiple sclerosis plaque. *Ann. Neurol.* **78**, 710–721 (2015).

216. Guneykaya, D. *et al.* Transcriptional and Translational Differences of Microglia from Male and Female Brains. *Cell Rep.* **24**, 2773-2783.e6 (2018).

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

218. Abi Ghanem, C. *et al.* Long-lasting masculinizing effects of postnatal androgens on myelin governed by the brain androgen receptor. *PLoS Genet.* **13**, e1007049 (2017).

219. Yasuda, K. *et al.* Sex-specific differences in transcriptomic profiles and cellular characteristics of oligodendrocyte precursor cells. *Stem Cell Res.* **46**, 101866 (2020).

220. Zahaf, A. *et al.* Androgens show sex-dependent differences in myelination in immune and non-immune murine models of CNS demyelination. *Nat. Commun.* **14**, 1592 (2023).

221. Diem, L., Hammer, H., Hoepner, R., Pistor, M., Remlinger, J. & Salmen, A. Sex and gender differences in autoimmune demyelinating CNS disorders: Multiple sclerosis (MS), neuromyelitis optica spectrum disorder (NMOSD) and myelin-oligodendrocyte-glycoprotein antibody associated disorder (MOGAD). *Int. Rev. Neurobiol.* **164**, 129–178 (2022).

222. Taylor, L. C., Puranam, K., Gilmore, W., Ting, J. P.-Y. & Matsushima, G. K. 17beta-estradiol protects male mice from cuprizone-induced demyelination and oligodendrocyte loss. *Neurobiol. Dis.* **39**, 127–137 (2010).

223. Collongues, N., Patte-Mensah, C., De Seze, J., Mensah-Nyagan, A.-G. & Derfuss, T. Testosterone and estrogen in multiple sclerosis: from pathophysiology to therapeutics. *Expert Rev. Neurother.* **18**, 515–522 (2018).

224. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).

225. Umair, M., Fazazi, M. R. & Rangachari, M. Biological Sex As a Critical Variable in CD4+ Effector T Cell Function in Preclinical Models of Multiple Sclerosis. *Antioxid. Redox Signal.* **37**, 135–149 (2022).

226. Tejera-Alhambra, M. *et al.* Perforin expression by CD4+ regulatory T cells increases at multiple sclerosis relapse: sex differences. *Int. J. Mol. Sci.* **13**, 6698–6710 (2012).

227. Layug, P. J., Vats, H., Kannan, K. & Arsenio, J. Sex differences in CD8+ T cell responses during adaptive immunity. *WIREs Mech. Dis.* **16**, e1645 (2024).

228. Karrenbauer, V. D., Bedri, S. K., Hillert, J. & Manouchehrinia, A. Cerebrospinal fluid oligoclonal immunoglobulin gamma bands and long-term disability progression in multiple sclerosis: a retrospective cohort study. *Sci. Rep.* **11**, 14987 (2021).

229. Castellazzi, M. *et al.* The Sexual Dimorphism in Cerebrospinal Fluid Protein Content Does Not Affect Intrathecal IgG Synthesis in Multiple Sclerosis. *J. Pers. Med.* **12**, 977 (2022).

230. Voskuhl, R. R. & Gold, S. M. Sex-related Factors in Multiple Sclerosis: Genetic, Hormonal and Environmental Contributions. *Nat. Rev. Neurol.* **8**, 255–263 (2012).

231. Dhaeze, T. *et al.* Sex-dependent factors encoded in the immune compartment dictate relapsing or progressive phenotype in demyelinating disease. *JCI Insight* **4**, e124885, 124885 (2019).

232. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

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

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

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

236. Speir, M. L. *et al.* UCSC Cell Browser: visualize your single-cell data. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab503.

237. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).

238. Kim, G. D., Lim, C. & Park, J. A practical handbook on single-cell RNA sequencing data quality control and downstream analysis. *Mol. Cells* **47**, 100103 (2024).

239. Hong, R. *et al.* Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data. *Nat. Commun.* **13**, 1688 (2022).

240. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2021).

241. Subramanian, A., Alperovich, M., Yang, Y. & Li, B. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biol.* **23**, 267 (2022).

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

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

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

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

246. Blighe, K. & Lun, A. *PCAtools: Everything Principal Components Analysis*. R package version 2.4.0 (2021). https://github.com/kevinblighe/PCAtools, last accessed September 14, 2025.

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

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

249. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal* Complex Systems, 1695 (2005).

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

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

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

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

254. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinforma. Oxf. Engl.* **22**, 1600–1607 (2006).

255. Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human.* R package. https://bioconductor.org/packages/org.Hs.eg.db

256. Alexa, A. & Rahnenfuhrer, J. *topGO: Enrichment analysis for Gene Ontology*. R package. https://bioconductor.org/packages/topGO

257. Gu, Z. & Hübschmann, D. simplifyEnrichment: A Bioconductor Package for Clustering and Visualizing Functional Enrichment Results. *Genomics Proteomics Bioinformatics* **21**, 190–202 (2023).

258. Hidalgo, M. R. *et al.* High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget* **8**, 5160–5178 (2017).

259. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).

260. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat. Protoc.* **20**, 180–219 (2025).

261. Chang, W., *et al. shiny: Web Application Framework for R*. R package https://shiny.posit.co/.

262. Khoy, K. *et al.* Natalizumab in Multiple Sclerosis Treatment: From Biological Effects to Immune Monitoring. *Front. Immunol.* **11**, 549842 (2020).

263. Schwarz, K. & Schmitz, F. Synapse Dysfunctions in Multiple Sclerosis. *Int. J. Mol. Sci.* **24**, 1639 (2023).

264. Mahmoud, S., Gharagozloo, M., Simard, C. & Gris, D. Astrocytes Maintain Glutamate Homeostasis in the CNS by Controlling the Balance between Glutamate Uptake and Release. *Cells* **8**, 184 (2019).

265. Halder, S. K. & Milner, R. Hypoxia in multiple sclerosis; is it the chicken or the egg? *Brain J. Neurol.* **144**, 402–410 (2021).

266. An, H., Williams, N. G. & Shelkovnikova, T. A. NEAT1 and paraspeckles in neurodegenerative diseases: A missing lnc found? *Non-Coding RNA Res.* **3**, 243–252 (2018).

267. Simchovitz, A. *et al.* A lncRNA survey finds increases in neuroprotective LINC-PINT in Parkinson's disease substantia nigra. *Aging Cell* **19**, e13115 (2020).

268. Li, W.-W., Penderis, J., Zhao, C., Schumacher, M. & Franklin, R. J. M. Females remyelinate more efficiently than males following demyelination in the aged but not young adult CNS. *Exp. Neurol.* **202**, 250–254 (2006).

269. Hughes, A. N. & Appel, B. Oligodendrocytes express synaptic proteins that modulate myelin sheath formation. *Nat. Commun.* **10**, 4125 (2019).

270. Chen, Y., Tian, D., Ku, L., Osterhout, D. J. & Feng, Y. The selective RNA-binding protein quaking I (QKI) is necessary and sufficient for promoting oligodendroglia differentiation. *J. Biol. Chem.* **282**, 23553–23560 (2007).

271. Zhou, X. *et al.* Mature myelin maintenance requires Qki to coactivate PPARβ-RXRα-mediated lipid metabolism. *J. Clin. Invest.* **130**, 2220–2236 (2020).

272. Watzlawik, J. O., Warrington, A. E. & Rodriguez, M. PDGF is required for remyelination-promoting IgM stimulation of oligodendrocyte progenitor cell proliferation. *PloS One* **8**, e55149 (2013).

273. Vana, A. C. *et al.* Platelet-derived growth factor promotes repair of chronically demyelinated white matter. *J. Neuropathol. Exp. Neurol.* **66**, 975–988 (2007).

274. Zeitelhofer, M. *et al.* Blocking PDGF-CC signaling ameliorates multiple sclerosis-like neuroinflammation by inhibiting disruption of the blood-brain barrier. *Sci. Rep.* **10**, 22383 (2020).

275. Cantuti-Castelvetri, L. *et al.* Defective cholesterol clearance limits remyelination in the aged central nervous system. *Science* **359**, 684–688 (2018).

276. Wilkinson, K. & El Khoury, J. Microglial scavenger receptors and their roles in the pathogenesis of Alzheimer's disease. *Int. J. Alzheimers Dis.* **2012**, 489456 (2012).

277. Qi, G. *et al.* ApoE4 Impairs Neuron-Astrocyte Coupling of Fatty Acid Metabolism. *Cell Rep.* **34**, 108572 (2021).

278. Pinholt, M., Frederiksen, J. L. & Christiansen, M. The association between apolipoprotein E and multiple sclerosis. *Eur. J. Neurol.* **13**, 573–580 (2006).

279. Shi, J., Zhao, C. B., Vollmer, T. L., Tyry, T. M. & Kuniyoshi, S. M. APOE ε4 allele is associated with cognitive impairment in patients with multiple sclerosis. *Neurology* **70**, 185–190 (2008).

280. Trenova, A. G., Slavov, G. S., Manova, M. G., Kostadinova, I. I. & Vasileva, T. V. Female sex hormones and cytokine secretion in women with multiple sclerosis. *Neurol. Res.* **35**, 95–99 (2013).

281. Atsaves, V., Leventaki, V., Rassidakis, G. Z. & Claret, F. X. AP-1 Transcription Factors as Regulators of Immune Responses in Cancer. *Cancers* **11**, 1037 (2019).

282. Zhou, Y. *et al.* Nuclear Factor κB (NF-κB)-Mediated Inflammation in Multiple Sclerosis. *Front. Immunol.* **11**, 391 (2020).

283. Webb, P. *et al.* The estrogen receptor enhances AP-1 activity by two distinct mechanisms with different requirements for receptor transactivation functions. *Mol. Endocrinol. Baltim. Md* **13**, 1672–1685 (1999).

284. Xing, D. *et al.* Estrogen modulates NFκB signaling by enhancing IκBα levels and blocking p65 binding at the promoters of inflammatory genes via estrogen receptor-β. *PloS One* **7**, e36890 (2012).

285. Al-Kafaji, G., Alwehaidah, M. S., Alsabbagh, M. M., Alharbi, M. A. & Bakhiet, M. Mitochondrial DNA haplogroup analysis in Saudi Arab patients with multiple sclerosis. *PloS One* **17**, e0279237 (2022).

286. Leppert, D. *et al.* Blood Neurofilament Light in Progressive Multiple Sclerosis. *Neurology* **98**, e2120–e2131 (2022).

287. Pender, M. P., Csurhes, P. A., Pfluger, C. M. & Burrows, S. R. Deficiency of CD8+ effector memory T cells is an early and persistent feature of multiple sclerosis. *Mult. Scler. Houndmills Basingstoke Engl.* **20**, 1825–1832 (2014).

288. Serafini, B., Rosicarelli, B., Veroni, C. & Aloisi, F. Tissue-resident memory T cells in the multiple sclerosis brain and their relationship to Epstein-Barr virus infected B cells. *J. Neuroimmunol.* **376**, 578036 (2023).

289. Machado-Santos, J. *et al.* The compartmentalized inflammatory response in the multiple sclerosis brain is composed of tissue-resident CD8+ T lymphocytes and B cells. *Brain J. Neurol.* **141**, 2066–2082 (2018).

290. Wang, P.-F. *et al.* Mitochondrial and metabolic dysfunction of peripheral immune cells in multiple sclerosis. *J. Neuroinflammation* **21**, 28 (2024).

291. Weinberg, S. E. & Jennings, L. J. HLA and Autoimmune Disease. *Adv. Mol. Pathol.* **3**, 207–219 (2020).

292. Stein, M. M. *et al.* Sex-specific differences in peripheral blood leukocyte transcriptional response to LPS are enriched for HLA region and X chromosome genes. *Sci. Rep.* **11**, 1107 (2021).

293. Reichert, S., Stein, J., Gautsch, A., Schaller, H.-G. & Machulla, H. K. G. Gender differences in HLA phenotype frequencies found in German patients with generalized aggressive periodontitis and chronic periodontitis. *Oral Microbiol. Immunol.* **17**, 360–368 (2002).

294. Enz, L. S. *et al.* Increased HLA-DR expression and cortical demyelination in MS links with HLA-DR15. *Neurol. Neuroimmunol. Neuroinflammation* **7**, e656 (2020).

295. Martin, R., Sospedra, M., Eiermann, T. & Olsson, T. Multiple sclerosis: doubling down on MHC. *Trends Genet. TIG* **37**, 784–797 (2021).

296. Caillier, S. J. *et al.* Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. *J. Immunol. Baltim. Md 1950* **181**, 5473–5480 (2008).

297. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat. Genet.* **47**, 1107–1113 (2015).

298. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).

299. Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G. & Doxey, A. C. An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genomics* **6**, e000341 (2020).

300. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596 (2013).

301. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

302. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

303. García-López, M. *et al.* Analysis of 1,000 Type-Strain Genomes Improves Taxonomic Classification of Bacteroidetes. *Front. Microbiol.* **10**, (2019).

304. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci.* **103**, 626–631 (2006).

305. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

306. Minich, J. J. *et al.* Host biology, ecology and the environment influence microbial biomass and diversity in 101 marine fish species. *Nat. Commun.* **13**, 6978 (2022).

307. Qu, Q. *et al.* Population-level gut microbiome and its associations with environmental factors and metabolic disorders in Southwest China. *Npj Biofilms Microbiomes* **11**, 24 (2025).

308. Ranheim Sveen, T., Hannula, S. E. & Bahram, M. Microbial regulation of feedbacks to ecosystem change. *Trends Microbiol.* **32**, 68–78 (2024).

309. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

310. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

311. Hold, G. L., Pryde, S. E., Russell, V. J., Furrie, E. & Flint, H. J. Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. *FEMS Microbiol. Ecol.* **39**, 33–39 (2002).

312. Gill, S. R. *et al.* Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **312**, 1355–1359 (2006).

313. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).

314. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

315. Ruan, W., Engevik, M. A., Spinler, J. K. & Versalovic, J. Healthy Human Gastrointestinal Microbiome: Composition and Function After a Decade of Exploration. *Dig. Dis. Sci.* **65**, 695–705 (2020).

316. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

317. Winter, S. E. & Bäumler, A. J. Gut dysbiosis: Ecological causes and causative effects on human disease. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2316579120 (2023).

318. Mann, E. R., Lam, Y. K. & Uhlig, H. H. Short-chain fatty acids: linking diet, the microbiome and immunity. *Nat. Rev. Immunol.* **24**, 577–595 (2024).

319. Singh, V. *et al.* Butyrate producers, "The Sentinel of Gut": Their intestinal significance with and beyond butyrate, and prospective use as microbial therapeutics. *Front. Microbiol.* **13**, 1103836 (2023).

320. Rios-Covian, D., Salazar, N., Gueimonde, M. & de los Reyes-Gavilan, C. G. Shaping the Metabolism of Intestinal Bacteroides Population through Diet to Improve Human Health. *Front. Microbiol.* **8**, 376 (2017).

321. Pokusaeva, K., Fitzgerald, G. F. & van Sinderen, D. Carbohydrate metabolism in Bifidobacteria. *Genes Nutr.* **6**, 285–306 (2011).

322. Hoskisson, P. A. & Fernández-Martínez, L. T. Regulation of specialised metabolites in Actinobacteria – expanding the paradigms. *Environ. Microbiol. Rep.* **10**, 231–238 (2018).

323. Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C. & Gasbarrini, A. Proteobacteria: A Common Factor in Human Diseases. *BioMed Res. Int.* **2017**, 9351507 (2017).

324. Chen, L. *et al.* The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302-2315.e12 (2021).

325. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).

326. Kruger, K. *et al.* Evaluation of inter- and intra-variability in gut health markers in healthy adults using an optimised faecal sampling and processing method. *Sci. Rep.* **14**, 24580 (2024).

327. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

328. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).

329. Johnson, A. J. *et al.* A Guide to Diet-Microbiome Study Design. *Front. Nutr.* **7**, (2020).

330. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

331. Parizadeh, M. & Arrieta, M.-C. The global human gut microbiome: genes, lifestyles, and diet. *Trends Mol. Med.* **29**, 789–801 (2023).

332. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).

333. Madsen, L., Myrmel, L. S., Fjære, E., Liaset, B. & Kristiansen, K. Links between Dietary Protein Sources, the Gut Microbiota, and Obesity. *Front. Physiol.* **8**, (2017).

334. Falony, G., Vieira-Silva, S. & Raes, J. Richness and ecosystem development across faecal snapshots of the gut microbiota. *Nat. Microbiol.* **3**, 526–528 (2018).

335. Procházková, N. *et al.* Advancing human gut microbiota research by considering gut transit time. *Gut* **72**, 180–191 (2023).

336. Lewis, S. J. & Heaton, K. W. Stool Form Scale as a Useful Guide to Intestinal Transit Time. *Scand. J. Gastroenterol.* **32**, 920–924 (1997).

337. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).

338. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).

339. Valles-Colomer, M. *et al.* Variation and transmission of the human gut microbiota across multiple familial generations. *Nat. Microbiol.* **7**, 87–96 (2022).

340. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).

341. Vandeputte, D. *et al.* Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nat. Commun.* **12**, 6740 (2021).

342. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).

343. Alili, R. *et al.* Characterization of the Gut Microbiota in Individuals with Overweight or Obesity during a Real-World Weight Loss Dietary Program: A Focus on the Bacteroides 2 Enterotype. *Biomedicines* **10**, 16 (2021).

344. Bah, Y. R. *et al.* Bacteroides- and Prevotella-enriched gut microbial clusters associate with metabolic risks. *Gut Pathog.* **17**, 55 (2025).

345. Knights, D. *et al.* Rethinking "Enterotypes". *Cell Host Microbe* **16**, 433–437 (2014).

346. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103 (2016).

347. Nicholson, J. K. *et al.* Host-gut microbiota metabolic interactions. *Science* **336**, 1262–1267 (2012).

348. Carías Domínguez, A. M. *et al.* Intestinal Dysbiosis: Exploring Definition, Associated Symptoms, and Perspectives for a Comprehensive Understanding - a Scoping Review. *Probiotics Antimicrob. Proteins* **17**, 440–449 (2025).

349. Hou, K. *et al.* Microbiota in health and diseases. *Signal Transduct. Target. Ther.* **7**, 135 (2022).

350. Joos, R. *et al.* Examining the healthy human microbiome concept. *Nat. Rev. Microbiol.* **23**, 192–205 (2025).

351. Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8**, 51 (2016).

352. Shanahan, F., Ghosh, T. S. & O'Toole, P. W. The Healthy Microbiome—What Is the Definition of a Healthy Gut Microbiome? *Gastroenterology* **160**, 483–494 (2021).

353. Rinninella, E. *et al.* What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* **7**, 14 (2019).

354. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).

355. Zhao, H. *et al.* Systematic review and meta-analysis of the role of Faecalibacterium prausnitzii alteration in inflammatory bowel disease. *J. Gastroenterol. Hepatol.* **36**, 320–328 (2021).

356. Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J. & Martinez-Medina, M. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *ISME J.* **11**, 841–852 (2017).

357. Caenepeel, C. *et al.* Dysbiosis and Associated Stool Features Improve Prediction of Response to Biological Therapy in Inflammatory Bowel Disease. *Gastroenterology* **166**, 483–495 (2024).

358. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).

359. Henke, M. T. *et al.* Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci.* **116**, 12672–12677 (2019).

360. Valeri, F. & Endres, K. How biological sex of the host shapes its gut microbiota. *Front. Neuroendocrinol.* **61**, 100912 (2021).

361. Gancz, N. N., Levinson, J. A. & Callaghan, B. L. Sex and gender as critical and distinct contributors to the human brain-gut-microbiome axis. *Brain Res. Bull.* **199**, 110665 (2023).

362. Jovanovic, N. *et al.* A gender perspective on diet, microbiome, and sex hormone interplay in cardiovascular disease. *Acta Physiol.* **240**, e14228 (2024).

363. Flak, M. B., Neves, J. F. & Blumberg, R. S. Welcome to the Microgenderome. *Science* **339**, 1044–1045 (2013).

364. Mulak, A., Larauche, M. & Taché, Y. Sexual Dimorphism in the Gut Microbiome: Microgenderome or Microsexome? *J. Neurogastroenterol. Motil.* **28**, 332–333 (2022).

365. Ma, Z. (Sam) & Li, W. How and Why Men and Women Differ in Their Microbiomes: Medical Ecology and Network Analyses of the Microgenderome. *Adv. Sci.* **6**, 1902054 (2019).

366. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, e1002533 (2016).

367. Jung, H.-K., Kim, D.-Y. & Moon, I.-H. Effects of Gender and Menstrual Cycle on Colonic Transit Time in Healthy Subjects. *Korean J. Intern. Med.* **18**, 181–186 (2003).

368. Kim, Y. S., Unno, T., Kim, B.-Y. & Park, M.-S. Sex Differences in Gut Microbiota. *World J. Mens Health* **38**, 48–60 (2020).

369. Mueller, S. *et al.* Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: a Cross-Sectional Study. *Appl. Environ. Microbiol.* **72**, 1027–1033 (2006).

370. Ma, Z. S. Revisiting microgenderome: detecting and cataloguing sexually unique and enriched species in human microbiomes. *BMC Biol.* **22**, 284 (2024).

371. Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. Prevotella diversity, niches and interactions with the human host. *Nat. Rev. Microbiol.* **19**, 585–599 (2021).

372. Takagi, T. *et al.* Differences in gut microbiota associated with age, sex, and stool consistency in healthy Japanese subjects. *J. Gastroenterol.* **54**, 53–63 (2019).

373. Yoon, K. & Kim, N. Roles of Sex Hormones and Gender in the Gut Microbiota. *J. Neurogastroenterol. Motil.* **27**, 314–325 (2021).

374. Cox, L. M., Abou-El-Hassan, H., Maghzi, A. H., Vincentini, J. & Weiner, H. L. The sex-specific interaction of the microbiome in neurodegenerative diseases. *Brain Res.* **1724**, 146385 (2019).

375. Ekpruke, C. D., Alford, R., Parker, E. & Silveyra, P. Gonadal sex and chromosome complement influence the gut microbiome in a mouse model of allergic airway inflammation. *Physiol. Genomics* **56**, 417–425 (2024).

376. Hooper, L. V., Littman, D. R. & Macpherson, A. J. Interactions between the microbiota and the immune system. *Science* **336**(6086), 1268-1273 (2012).

377. Kieser, K. J. & Kagan, J. C. Multi-receptor detection of individual bacterial products by the innate immune system. *Nat. Rev. Immunol.* **17**, 376–390 (2017).

378. Gomez, A., Luckey, D. & Taneja, V. The gut microbiome in autoimmunity: Sex matters. *Clin. Immunol.* **159**, 154–162 (2015).

379. Fransen, F. *et al.* The Impact of Gut Microbiota on Gender-Specific Differences in Immunity. *Front. Immunol.* **8**, 754 (2017).

380. Huang, M.-X. *et al.* Gut microbiota contributes to sexual dimorphism in murine autoimmune cholangitis. *J. Leukoc. Biol.* **110**, 1121–1130 (2021).

381. Loh, J. S. *et al.* Microbiota–gut–brain axis and its therapeutic applications in neurodegenerative diseases. *Signal Transduct. Target. Ther.* **9**, 37 (2024).

382. Qu, S. *et al.* Gut microbiota modulates neurotransmitter and gut-brain signaling. *Microbiol. Res.* **287**, 127858 (2024).

383. Hokanson, K. C., Hernández, C., Deitzler, G. E., Gaston, J. E. & David, M. M. Sex shapes gut–microbiota–brain communication and disease. *Trends Microbiol.* **32**, 151–161 (2024).

384. Zhang, S., Cai, H., Wang, C., Zhu, J. & Yu, Y. Sex-dependent gut microbiota-brain-cognition associations: a multimodal MRI study. *BMC Neurol.* **23**, 169 (2023).

385. Caldarelli, M. *et al.* Gut–Brain Axis: Focus on Sex Differences in Neuroinflammation. *Int. J. Mol. Sci.* **25**, 5377 (2024).

386. Manosso, L. M. *et al.* Sex-related patterns of the gut-microbiota-brain axis in the neuropsychiatric conditions. *Brain Res. Bull.* **171**, 196–208 (2021).

387. Wang, M. *et al.* Alteration of gut microbiota-associated epitopes in children with autism spectrum disorders. *Brain. Behav. Immun.* **75**, 192–199 (2019).

388. Saha, P. & Sisodia, S. S. Role of the gut microbiome in mediating sex-specific differences in the pathophysiology of Alzheimer's disease. *Neurotherapeutics* **21**, e00426 (2024).

389. Zhou, X. *et al.* The intestinal microbiota exerts a sex-specific influence on neuroinflammation in a Parkinson's disease mouse model. *Neurochem. Int.* **173**, 105661 (2024).

390. Zhang, M. *et al.* Exploring the alteration of gut microbiota and brain function in gender-specific Parkinson's disease based on metagenomic sequencing. *Front. Aging Neurosci.* **15**, 1148546 (2023).

391. Wang, Y. & LêCao, K.-A. Managing batch effects in microbiome data. *Brief. Bioinform.* **21**, 1954–1970 (2020).

392. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, (2017).

393. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of Microbiome Data in the Presence of Excess Zeros. *Front. Microbiol.* **8**, (2017).

394. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput. Biol.* **10**, e1003531 (2014).

395. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* **12**, 3562 (2021).

396. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).

397. Zhou, R., Ng, S. K., Sung, J. J. Y., Goh, W. W. B. & Wong, S. H. Data pre-processing for analyzing microbiome data – A mini review. *Comput. Struct. Biotechnol. J.* **21**, 4804–4815 (2023).

398. Tito, R. Y. *et al.* Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development. *Nat. Med.* **30**, 1339–1348 (2024).

399. Srivastava, A., Akhter, Y. & Verma, D. A step-by-step procedure for analysing the 16S rRNA-based microbiome diversity using QIIME 2 and comprehensive PICRUSt2 illustration for functional prediction. *Arch. Microbiol.* **206**, 467 (2024).

400. Berer, K. *et al.* Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature* **479**, 538–541 (2011).

401. Ochoa-Repáraz, J. *et al.* Role of gut commensal microflora in the development of experimental autoimmune encephalomyelitis. *J. Immunol. Baltim. Md 1950* **183**, 6041–6050 (2009).

402. Cekanaviciute, E. *et al.* Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. *Proc. Natl. Acad. Sci.* **114**, 10713–10718 (2017).

403. Berer, K. *et al.* Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice. *Proc. Natl. Acad. Sci.* **114**, 10719–10724 (2017).

404. Correale, J., Hohlfeld, R. & Baranzini, S. E. The role of the gut microbiota in multiple sclerosis. *Nat. Rev. Neurol.* **18**, 544–558 (2022).

405. Shahi, S. K., Yadav, M., Ghimire, S. & Mangalam, A. K. Role of the gut microbiome in multiple sclerosis: from etiology to therapeutics. *Int. Rev. Neurobiol.* **167**, 185–215 (2022).

406. Chen, J. *et al.* Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Sci. Rep.* **6**, 28484 (2016).

407. Galluzzo, P. *et al.* Comparison of the Intestinal Microbiome of Italian Patients with Multiple Sclerosis and Their Household Relatives. *Life* **11**, 620 (2021).

408. Moles, L. *et al.* Microbial dysbiosis and lack of SCFA production in a Spanish cohort of patients with multiple sclerosis. *Front. Immunol.* **13**, 960761 (2022).

409. iMSMS Consortium. Electronic address: sergio.baranzini@ucsf.edu & iMSMS Consortium. Gut microbiome of multiple sclerosis patients and paired household healthy controls reveal associations with disease risk and course. *Cell* **185**, 3467-3486.e16 (2022).

410. Yadav, S. K., Chen, C., Dhib-Jalbut, S. & Ito, K. The mechanism of disease progression by aging and age-related gut dysbiosis in multiple sclerosis. *Neurobiol. Dis.* **212**, 106956 (2025).

411. Jangi, S. *et al.* Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* **7**, 12015 (2016).

412. Elsayed, N. S. *et al.* Genetic risk score in multiple sclerosis is associated with unique gut microbiome. *Sci. Rep.* **13**, 16269 (2023).

413. Montgomery, T. L. *et al.* Identification of commensal gut microbiota signatures as predictors of clinical severity and disease progression in multiple sclerosis. *Sci. Rep.* **14**, 15292 (2024).

414. Jiang, J. *et al.* Efficacy of probiotics in multiple sclerosis: a systematic review of preclinical trials and meta-analysis of randomized controlled trials. *Food Funct.* **12**, 2354–2377 (2021).

415. Li, Q., Wang, G., Zhao, J., Chen, W. & Tian, P. Gut microbiota and myelination: Crosstalk across the lifespan and microbiota-based modulation strategies. *Microbiol. Res.* **300**, 128286 (2025).

416. Chen, T., Noto, D., Hoshino, Y., Mizuno, M. & Miyake, S. Butyrate suppresses demyelination and enhances remyelination. *J. Neuroinflammation* **16**, 165 (2019).

417. Mirza, A. *et al.* The multiple sclerosis gut microbiota: A systematic review. *Mult. Scler. Relat. Disord.* **37**, (2020).

418. Ordoñez-Rodriguez, A., Roman, P., Rueda-Ruzafa, L., Campos-Rios, A. & Cardona, D. Changes in Gut Microbiota and Multiple Sclerosis: A Systematic Review. *Int. J. Environ. Res. Public. Health* **20**, 4624 (2023).

419. Jette, S., Schaetzen, C. de, Tsai, C.-C. & Tremlett, H. The multiple sclerosis gut microbiome and disease activity: A systematic review. *Mult. Scler. Relat. Disord.* **92**, (2024).

420. Lin, Q. *et al.* Meta-analysis identifies common gut microbiota associated with multiple sclerosis. *Genome Med.* **16**, 94 (2024).

421. Zhang, X., Wei, Z., Liu, Z., Yang, W. & Huai, Y. Changes in Gut Microbiota in Patients with Multiple Sclerosis Based on 16s rRNA Gene Sequencing Technology: A Review and Meta-Analysis. *J. Integr. Neurosci.* **23**, 127 (2024).

422. D'Anca, M. *et al.* Why Is Multiple Sclerosis More Frequent in Women? Role of the Immune System and of Oral and Gut Microbiota. *Appl. Sci.* **13**, 5881 (2023).

423. Miller, P. G., Bonn, M. B., Franklin, C. L., Ericsson, A. C. & McKarns, S. C. TNFR2 Deficiency Acts in Concert with Gut Microbiota To Precipitate Spontaneous Sex-Biased Central Nervous System Demyelinating Autoimmune Disease. *J. Immunol. Baltim. Md 1950* **195**, 4668–4684 (2015).

424. Benedek, G. *et al.* Estrogen protection against EAE modulates the microbiota and mucosal-associated regulatory cells. *J. Neuroimmunol.* **310**, 51–59 (2017).

425. Peng, H.-R. *et al.* Intestinal epithelial dopamine receptor signaling drives sex-specific disease exacerbation in a mouse model of multiple sclerosis. *Immunity* **56**, 2773-2789.e8 (2023).

426. Yu, S. & Zeng, M. Y. Sex biased gut dopamine signaling in multiple sclerosis. *Immunity* **56**, 2674–2676 (2023).

427. Becker, A. *et al.* Short-chain fatty acids and intestinal inflammation in multiple sclerosis: modulation of female susceptibility by microbial products? *Auto- Immun. Highlights* **12**, 7 (2021).

428. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

429. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

430. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

431. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).

432. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

433. Goldfarb, T. *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res.* **53**, D243–D257 (2025).

434. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **48**, D84–D86 (2020).

435. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

436. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

437. Kleine Bardenhorst, S., Vital, M., Karch, A. & Rübsamen, N. Richness estimation in microbiome data obtained from denoising pipelines. *Comput. Struct. Biotechnol. J.* **20**, 508–520 (2022).

438. Chakraborty, J., Palit, K. & Das, S. Metagenomic approaches to study the culture-independent bacterial diversity of a polluted environment — a case study on north-eastern coast of Bay of Bengal, India. In *Microbial Biodegradation and Bioremediation* 81–107. Elsevier (2022).

439. Willis, A. D. Rarefaction, Alpha Diversity, and Statistics. *Front. Microbiol.* **10**, (2019).

440. Marcon, E., Scotti, I., Hérault, B., Rossi, V. & Lang, G. Generalization of the partitioning of shannon diversity. *PloS One* **9**, e90289 (2014).

441. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217 (2013).

442. Su, X. Elucidating the Beta-Diversity of the Microbiome: from Global Alignment to Local Alignment. *mSystems* **6**, 10.1128/msystems.00363-21 (2021).

443. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).

444. Legendre, P. & Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280 (2001).

445. Borcard, D., Gillet, F. & Legendre, P. *Numerical Ecology with R*. (Springer International Publishing, Cham, 2018). doi:10.1007/978-3-319-71404-2.

446. Oksanen, J., *et al. vegan: Community Ecology Package*. R package https://vegandevs.github.io/vegan/.

447. Blanchet, F. G., Legendre, P. & Borcard, D. Forward selection of explanatory variables. *Ecology* **89**, 2623–2632 (2008).

448. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLOS ONE* **7**, e30126 (2012).

449. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).

450. Morgan, M. *et al. DirichletMultinomial: Dirichlet-Multinomial Mixture Models for Microbiome Data*. R package https://bioconductor.org/packages/DirichletMultinomial.

451. McHugh, M. L. The chi-square test of independence. *Biochem. Medica* **23**, 143–149 (2013).

452. R Core Team. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* (2021).

453. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1**, 80–83 (1945).

454. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).

455. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).

456. Dunn, O. J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).

457. Dinno, A. *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. R package. https://CRAN.R-project.org/package=dunn.test

458. Looney, S. W. & Hagan, J. L. Statistical methods for assessing biomarkers and analyzing biomarker data. In *Handbook of Statistics: Epidemiology and Medical Statistics*, Vol. 27, 27–65. Elsevier (2011).

459. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

460. Hothorn, T., Winell, H., Hornik, K., Van de Wiel, M. A. & Zeileis, A. *coin: Conditional Inference Procedures in a Permutation Test Framework*. R package. https://doi.org/10.32614/CRAN.package.coin.

461. Kassambara, A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. R package. https://doi.org/10.32614/CRAN.package.rstatix

462. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application*. Cambridge Univ. Press (1997).

463. McKenzie, J. E. & Veroniki, A. A. A brief note on the random-effects meta-analysis model and its relationship to other models. *J. Clin. Epidemiol.* **174**, 111492 (2024).

464. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).

465. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).

466. Belsley, D. A., Kuh, E. & Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley (1980).

467. Cook, R. D. & Weisberg, S. *Residuals and Influence in Regression*. New York: Chapman and Hall (1982).

468. Sterne, J. A. C. & Harbord, R. M. Funnel Plots in Meta-analysis. *Stata J.* **4**, 127–141 (2004).

469. Cox, L. M. *et al.* The Gut Microbiome in Progressive Multiple Sclerosis. *Ann. Neurol.* **89**, 1195–1211 (2021).

470. Yadav, M. *et al.* Multiple sclerosis patients have an altered gut mycobiome and increased fungal to bacterial richness. *PLoS ONE* **17**, e0264556 (2022).

471. Choileáin, S. N. *et al.* CXCR3+ T cells in multiple sclerosis correlate with reduced diversity of the gut microbiome. *J. Transl. Autoimmun.* **3**, 100032 (2020).

472. Kozhieva, M. *et al.* Primary progressive multiple sclerosis in a Russian cohort: relationship with gut bacterial diversity. *BMC Microbiol.* **19**, 309 (2019).

473. Ashraf, H. *et al.* Mycobacterium avium subspecies paratuberculosis (MAP) infection, and its impact on gut microbiome of individuals with multiple sclerosis. *Sci. Rep.* **14**, 24027 (2024).

474. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio* **9**, 10.1128/mbio.00525-18 (2018).

475. Forry, S. P. *et al.* Variability and bias in microbiome metagenomic sequencing: an interlaboratory study comparing experimental protocols. *Sci. Rep.* **14**, 9785 (2024).

476. Scepanovic, P. *et al.* A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. *Microbiome* **7**, 130 (2019).

477. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).

478. Vandeputte, D. *et al.* Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57-62 (2016).

479. García-López, R. *et al.* Doing More with Less: A Comparison of 16S Hypervariable Regions in Search of Defining the Shrimp Microbiota. *Microorganisms* **8**, 134 (2020).

480. Radjabzadeh, D. *et al.* Gut microbiome-wide association study of depressive symptoms. *Nat. Commun.* **13**, 7128 (2022).

481. Bonnechère, B., Amin, N. & van Duijn, C. What Are the Key Gut Microbiota Involved in Neurological Diseases? A Systematic Review. *Int. J. Mol. Sci.* **23**, 13665 (2022).

482. Hong, J. *et al.* Gut microbiome changes associated with chronic pancreatitis and pancreatic cancer: a systematic review and meta-analysis. *Int. J. Surg. Lond. Engl.* **110**, 5781–5794 (2024).

483. Wang, X. *et al.* Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. *Gut* **69**, 2131–2142 (2020).

484. Alexander, M. *et al.* Human gut bacterial metabolism drives Th17 activation and colitis. *Cell Host Microbe* **30**, 17-30.e9 (2022).

485. Shin, Y.-H., Bang, S., Xavier, R. & Clardy, J. Eggerthella lenta Produces a Cryptic Pro-inflammatory Lipid. *J. Am. Chem. Soc.* **147**, 25180–25183 (2025).

486. Amir, I., Bouvet, P., Legeay, C., Gophna, U. & Weinberger, A. Eisenbergiella tayi gen. nov., sp. nov., isolated from human blood. *Int. J. Syst. Evol. Microbiol.* **64**, 907–914 (2014).

487. Zheng, X. *et al.* Unveiling genetic links between gut microbiota and asthma: a Mendelian randomization. *Front. Microbiol.* **15**, 1448629 (2024).

488. Liu, B. *et al.* Assessing the relationship between gut microbiota and irritable bowel syndrome: a two-sample Mendelian randomization analysis. *BMC Gastroenterol.* **23**, 150 (2023).

489. Rodriguez-Castaño, G. P., Rey, F. E., Caro-Quintero, A. & Acosta-González, A. Gut-derived Flavonifractor species variants are differentially enriched during in vitro incubation with quercetin. *PloS One* **15**, e0227724 (2020).

490. Mikami, A. *et al.* Oral Administration of Flavonifractor plautii, a Bacteria Increased With Green Tea Consumption, Promotes Recovery From Acute Colitis in Mice via Suppression of IL-17. *Front. Nutr.* **7**, 610946 (2020).

491. Ogita, T. *et al.* Oral Administration of Flavonifractor plautii Strongly Suppresses Th2 Immune Responses in Mice. *Front. Immunol.* **11**, 379 (2020).

492. Saha, S. *et al.* Longitudinal analysis of gut microbiome and metabolome correlates of response and toxicity with idecabtagene vicleucel. *Blood Adv.* **9**, 3429–3440 (2025).

493. Berding, K. *et al.* Diet and the Microbiota–Gut–Brain Axis: Sowing the Seeds of Good Mental Health. *Adv. Nutr.* **12**, 1239–1285 (2021).

494. DONG, T. S., PETERS, K., GUPTA, A., JACOBS, J. P. & CHANG, L. Prevotella Is Associated With Sex-Based Differences in Irritable Bowel Syndrome. *Gastroenterology* **167**, 1221-1224.e8 (2024).

495. Shahi, S. K. *et al.* Prevotella histicola, A Human Gut Commensal, Is as Potent as COPAXONE® in an Animal Model of Multiple Sclerosis. *Front. Immunol.* **10**, (2019).

496. Larsen, J. M. The immune response to Prevotella bacteria in chronic inflammatory disease. *Immunology* **151**, 363–374 (2017).

497. Ramos Meyers, G., Samouda, H. & Bohn, T. Short Chain Fatty Acid Metabolism in Relation to Gut Microbiota and Genetic Variability. *Nutrients* **14**, 5361 (2022).

498. Yoon, H. *et al.* Multiple sclerosis and gut microbiota: Lachnospiraceae from the ileum of MS twins trigger MS-like disease in germfree transgenic mice-An unbiased functional study. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2419689122 (2025).

499. Gan, L., Cookson, M. R., Petrucelli, L. & La Spada, A. R. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nat. Neurosci.* **21**, 1300–1309 (2018).

500. Fu, H., Hardy, J. & Duff, K. E. Selective vulnerability in neurodegenerative diseases. *Nat. Neurosci.* **21**, 1350–1358 (2018).

501. Kampmann, M. Molecular and cellular mechanisms of selective vulnerability in neurodegenerative diseases. *Nat. Rev. Neurosci.* **25**, 351–371 (2024).

502. Ahmad, F., Javed, M., Athar, M. & Shahzadi, S. Determination of affected brain regions at various stages of Alzheimer's disease. *Neurosci. Res.* **192**, 77–82 (2023).

503. Damier, P., Hirsch, E. C., Agid, Y. & Graybiel, A. M. The substantia nigra of the human brain. II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain J. Neurol.* **122** (Pt 8), 1437–1448 (1999).

504. Surmeier, D. J., Guzman, J. N. & Sanchez-Padilla, J. Calcium, cellular aging, and selective neuronal vulnerability in Parkinson's disease. *Cell Calcium* **47**, 175–182 (2010).

505. Palop, J. J. & Mucke, L. Network abnormalities and interneuron dysfunction in Alzheimer disease. *Nat. Rev. Neurosci.* **17**, 777–792 (2016).

506. Iovino, L., Tremblay, M. E. & Civiero, L. Glutamate-induced excitotoxicity in Parkinson's disease: The role of glial cells. *J. Pharmacol. Sci.* **144**, 151–164 (2020).

507. Mandolesi, G. *et al.* Synaptopathy connects inflammation and neurodegeneration in multiple sclerosis. *Nat. Rev. Neurol.* **11**, 711–724 (2015).

508. Bustamante-Barrientos, F. A. *et al.* Mitochondrial dysfunction in neurodegenerative disorders: Potential therapeutic application of mitochondrial transfer to central nervous system-residing cells. *J. Transl. Med.* **21**, 613 (2023).

509. Derevyanko, A., Tao, T. & Allen, N. J. Common alterations to astrocytes across neurodegenerative disorders. *Curr. Opin. Neurobiol.* **90**, 102970 (2025).

510. Gao, C., Jiang, J., Tan, Y. & Chen, S. Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. *Signal Transduct. Target. Ther.* **8**, 359 (2023).

511. Das, S., Zhang, Z. & Ang, L. C. Clinicopathological overlap of neurodegenerative diseases: A comprehensive review. *J. Clin. Neurosci.* **78**, 30–33 (2020).

512. Etemadifar, M., Afshar, F., Nasr, Z. & Kheradmand, M. Parkinsonism associated with multiple sclerosis: A report of eight new cases and a review on the literature. *Iran. J. Neurol.* **13**, 88–93 (2014).

513. Cottrill, R. *et al.* Alzheimer's disease (AD) in multiple sclerosis (MS): A systematic review of published cases, mechanistic links between AD and MS, and possible clinical evaluation of AD in MS. *J. Alzheimers Dis. Rep.* **9**, 25424823251316134 (2025).

514. Rajput, A. H., Rozdilsky, B. & Rajput, A. Alzheimer's disease and idiopathic Parkinson's disease coexistence. *J. Geriatr. Psychiatry Neurol.* **6**, 170–176 (1993).

515. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).

516. Steeg, P. S. Targeting metastasis. *Nat. Rev. Cancer* **16**, 201–218 (2016).

517. Wu, A., Colón, G. R. & Lim, M. Quality of Life and Role of Palliative and Supportive Care for Patients With Brain Metastases and Caregivers: A Review. *Front. Neurol.* **13**, (2022).

518. Yri, O. E. *et al.* Survival and quality of life after first-time diagnosis of brain metastases: a multicenter, prospective, observational study. *Lancet Reg. Health – Eur.* **49**, (2025).

519. Achrol, A. S. *et al.* Brain metastases. *Nat. Rev. Dis. Primer* **5**, 5 (2019).

520. Giannoudis, A. *et al.* Breast cancer brain metastases genomic profiling identifies alterations targetable by immune-checkpoint and PARP inhibitors. *Npj Precis. Oncol.* **8**, 282 (2024).

521. Yoo, J. *et al.* Spatial Transcriptomic Landscape of Brain Metastases from Triple-Negative Breast Cancer: Comparison of Primary Tumor and Brain Metastases Using Spatial Analysis. *Cancer Res. Treat.* (2025).

522. Geissler, M. *et al.* The Brain Pre-Metastatic Niche: Biological and Technical Advancements. *Int. J. Mol. Sci.* **24**, 10055 (2023).

523. Celià-Terrassa, T. & Kang, Y. Metastatic niche functions and therapeutic opportunities. *Nat. Cell Biol.* **20**, 868–877 (2018).

524. Powell, A. M. *et al.* The epigenetic landscape of brain metastasis. *Oncogene* **44**, 2227–2239 (2025).

525. Charles, N. A., Holland, E. C., Gilbertson, R., Glass, R. & Kettenmann, H. The brain tumor microenvironment. *Glia* **60**, 502–514 (2012).

526. Yuzhalin, A. E. & Yu, D. Critical functions of extracellular matrix in brain metastasis seeding. *Cell. Mol. Life Sci. CMLS* **80**, 297 (2023).

527. Wang, Z., Wu, X., Chen, H.-N. & Wang, K. Amino acid metabolic reprogramming in tumor metastatic colonization. *Front. Oncol.* **13**, (2023).

528. Deshpande, K. *et al.* SRRM4-mediated REST to REST4 dysregulation promotes tumor growth and neural adaptation in breast cancer leading to brain metastasis. *Neuro-Oncol.* **26**, 309–322 (2024).

529. Ruan, X. *et al.* Breast cancer cell-secreted miR-199b-5p hijacks neurometabolic coupling to promote brain metastasis. *Nat. Commun.* **15**, 4549 (2024).

530. Deshpande, K. *et al.* Neuronal exposure induces neurotransmitter signaling and synaptic mediators in tumors early in brain metastasis. *Neuro-Oncol.* **24**, 914–924 (2021).

531. Xing, X. *et al.* Pan-cancer human brain metastases atlas at single-cell resolution. *Cancer Cell* **43**, 1242-1260.e9 (2025).

532. Strickland, M. R., Alvarez-Breckenridge, C., Gainor, J. F. & Brastianos, P. K. Tumor immune microenvironment of brain metastases: towards unlocking anti-tumor immunity. *Cancer Discov.* **12**, 1199–1216 (2022).

533. Xing, F. *et al.* Reactive astrocytes promote the metastatic growth of breast cancer stem-like cells by activating Notch signalling in brain. *EMBO Mol. Med.* **5**, 384–396 (2013).

534. Valiente, M. *et al.* Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell* **156**, 1002–1016 (2014).

535. Burn, L., Gutowski, N., Whatmore, J., Giamas, G. & Pranjol, M. Z. I. The role of astrocytes in brain metastasis at the interface of circulating tumour cells and the blood brain barrier. *Front. Biosci. Landmark Ed.* **26**, 590–601 (2021).

536. Schulz, M. & Prinz, M. Brain metastasis: From etiology to ecotypes. *Cancer Cell* **43**, 1193–1195 (2025).

537. Vilariño, N., Bruna, J., Bosch-Barrera, J., Valiente, M. & Nadal, E. Immunotherapy in NSCLC patients with brain metastases. Understanding brain tumor microenvironment and dissecting outcomes from immune checkpoint blockade in the clinic. *Cancer Treat. Rev.* **89**, 102067 (2020).

538. Li, Y. D. *et al.* Tumor-induced peripheral immunosuppression promotes brain metastasis in patients with non-small cell lung cancer. *Cancer Immunol. Immunother. CII* **68**, 1501–1513 (2019).

539. Tas, F. Metastatic Behavior in Melanoma: Timing, Pattern, Survival, and Influencing Factors. *J. Oncol.* **2012**, 647684 (2012).

540. Nguyen, D. X., Bos, P. D. & Massagué, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **9**, 274–284 (2009).

541. Smedby, K. E., Brandt, L., Bäcklund, M. L. & Blomqvist, P. Brain metastases admissions in Sweden between 1987 and 2006. *Br. J. Cancer* **101**, 1919–1924 (2009).

542. Davies, M. A. *et al.* Prognostic factors for survival in melanoma patients with brain metastases. *Cancer* **117**, 1687–1696 (2011).

543. Vosoughi, E. *et al.* Survival and clinical outcomes of patients with melanoma brain metastasis in the era of checkpoint inhibitors and targeted therapies. *BMC Cancer* **18**, 490 (2018).

544. Bander, E. D. *et al.* Melanoma Brain Metastasis Presentation, Treatment and Outcomes in the Age of Targeted- and Immuno-therapies. *Cancer* **127**, 2062–2073 (2021).

545. Redmer, T. Deciphering mechanisms of brain metastasis in melanoma - the gist of the matter. *Mol. Cancer* **17**, 106 (2018).

546. Rabbie, R. *et al.* The mutational landscape of melanoma brain metastases presenting as the first visceral site of recurrence. *Br. J. Cancer* **124**, 156–160 (2021).

547. Kumar, S. *et al.* Integrated analysis of molecular and clinical features associated with overall survival in melanoma patients with brain metastasis. *Acta Neuropathol. Commun.* **13**, 75 (2025).

548. Biermann, J. *et al.* Dissecting the treatment-naive ecosystem of human melanoma brain metastasis. *Cell* **185**, 2591-2608.e30 (2022).

549. Salvati, L., Mandalà, M. & Massi, D. Melanoma Brain Metastases: Review of Histopathological Features and Immune-Molecular Aspects. *Melanoma Manag.* **7**, 4–10 (2020).

550. Radke, J. *et al.* Decoding molecular programs in melanoma brain metastases. *Nat. Commun.* **13**, 7304 (2022).

551. Fischer, G. M. *et al.* Molecular Profiling Reveals Unique Immune and Metabolic Features of Melanoma Brain Metastases. *Cancer Discov.* **9**, 628–645 (2019).

552. Vasudevan, H. N. *et al.* Molecular Features of Resected Melanoma Brain Metastases, Clinical Outcomes, and Responses to Immunotherapy. *JAMA Netw. Open* **6**, e2329186 (2023).

553. Phadke, M., Ozgun, A., Eroglu, Z. & Smalley, K. S. M. Melanoma brain metastases: Biological basis and novel therapeutic strategies. *Exp. Dermatol.* **31**, 31–42 (2022).

554. Kennedy, L. B. *et al.* A comprehensive, multi-center, immunogenomic analysis of melanoma brain metastases. *Acta Neuropathol. Commun.* **13**, 123 (2025).

555. In, G. K. *et al.* Multi-omic profiling reveals discrepant immunogenic properties and a unique tumor microenvironment among melanoma brain metastases. *Npj Precis. Oncol.* **7**, 120 (2023).

556. LeBleu, V. S. *et al.* PGC-1α mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis. *Nat. Cell Biol.* **16**, 992–1003 (2014).

557. Zhang, G. *et al.* Targeting mitochondrial biogenesis to overcome drug resistance to MAPK inhibitors. *J. Clin. Invest.* **126**, 1834–1856 (2016).

558. Iyengar, B. & Singh, A. V. Patterns of neural differentiation in melanomas. *J. Biomed. Sci.* **17**, 87 (2010).

559. Vidal, A. & Redmer, T. Tracking of Melanoma Cell Plasticity by Transcriptional Reporters. *Int. J. Mol. Sci.* **23**, 1199 (2022).

560. Radke, J., Roßner, F. & Redmer, T. CD271 determines migratory properties of melanoma cells. *Sci. Rep.* **7**, 9834 (2017).

561. Seo, J. & Park, M. Molecular crosstalk between cancer and neurodegenerative diseases. *Cell. Mol. Life Sci. CMLS* **77**, 2659–2680 (2019).

562. Driver, J. A., *et al.* Inverse association between cancer and Alzheimer's disease: results from the Framingham Heart Study. *BMJ* **344**, e1442 (2012).

563. Driver, J. A., Logroscino, G., Buring, J. E., Gaziano, J. M. & Kurth, T. A prospective cohort study of cancer incidence following the diagnosis of Parkinson's disease. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **16**, 1260–1265 (2007).

564. Wang, F., Chen, L., Nie, M. & Li, Z. Integrative analysis of causal associations between neurodegenerative diseases and colorectal cancer. *Heliyon* **10**, (2024).

565. Kuiper, J. G., Overbeek, J. A., Foch, C., Boutmy, E. & Sabidó, M. Incidence of malignancies in patients with multiple sclerosis versus a healthy matched cohort: A population-based cohort study in the Netherlands using the PHARMO Database Network. *J. Clin. Neurosci.* **103**, 49–55 (2022).

566. Jiang, M. *et al.* Incidence and Characteristics of Melanoma in Multiple Sclerosis Patients Treated With Fingolimod: A Systematic Review. *Curr. Dermatol. Rep.* **12**, 300–313 (2023).

567. Carbone, M. L. *et al.* Multiple Sclerosis Treatment and Melanoma Development. *Int. J. Mol. Sci.* **21**, 2950 (2020).

568. Hemond, C. C. *et al.* Exacerbation of Multiple Sclerosis by BRAF/MEK Treatment for Malignant Melanoma: The Central Vein Sign to Distinguish Demyelinating Lesions From Metastases. *J. Investig. Med. High Impact Case Rep.* **9**, 23247096211033047 (2021).

569. Pape, K. *et al.* Case Report: Balancing immune responses – multiple sclerosis disease exacerbation under BRAF/MEK treatment for malignant melanoma. *Front. Oncol.* **13**, (2023).

570. Sugier, P.-E. *et al.* Investigation of Shared Genetic Risk Factors Between Parkinson's Disease and Cancers. *Mov. Disord. Off. J. Mov. Disord. Soc.* **38**, 604–615 (2023).

571. Ye, Q., Wen, Y., Al-Kuwari, N. & Chen, X. Association Between Parkinson's Disease and Melanoma: Putting the Pieces Together. *Front. Aging Neurosci.* **12**, 60 (2020).

572. Forés-Martos, J. *et al.* Transcriptomic and Genetic Associations between Alzheimer's Disease, Parkinson's Disease, and Cancer. *Cancers* **13**, 2990 (2021).

573. Zhang, X., Guarin, D., Mohammadzadehhonarvar, N., Chen, X. & Gao, X. Parkinson's disease and cancer: a systematic review and meta-analysis of over 17 million participants. *BMJ Open* **11**, e046329 (2021).

574. Arnold, M. R. *et al.* Alpha-synuclein regulates nucleolar DNA double-strand break repair in melanoma. *Sci. Adv.* **11**, eadq2519 (2025).

575. Israeli, E. *et al.* α-Synuclein expression selectively affects tumorigenesis in mice modeling Parkinson's disease. *PloS One* **6**, e19622 (2011).

576. Shekoohi, S. *et al.* Knocking out alpha-synuclein in melanoma cells dysregulates cellular iron metabolism and suppresses tumor growth. *Sci. Rep.* **11**, 5267 (2021).

577. Kwon, H. S. & Koh, S.-H. Neuroinflammation in neurodegenerative disorders: the roles of microglia and astrocytes. *Transl. Neurodegener.* **9**, 42 (2020).

578. Stephenson, J., Nutma, E., van der Valk, P. & Amor, S. Inflammation in CNS neurodegenerative diseases. *Immunology* **154**, 204–219 (2018).

579. López-Cerdán, A. *et al.* An integrated approach to identifying sex-specific genes, transcription factors, and pathways relevant to Alzheimer's disease. *Neurobiol. Dis.* **199**, 106605 (2024).

580. López-Cerdán, A. *et al.* Unveiling sex-based differences in Parkinson's disease: a comprehensive meta-analysis of transcriptomic studies. *Biol. Sex Differ.* **13**, 68 (2022).

581. Cosgrove, N. *et al.* Mapping molecular subtype specific alterations in breast cancer brain metastases identifies clinically relevant vulnerabilities. *Nat. Commun.* **13**, 514 (2022).

582. Varešlija, D. *et al.* Transcriptome Characterization of Matched Primary Breast and Brain Metastatic Tumors to Detect Novel Actionable Targets. *J. Natl. Cancer Inst.* **111**, 388–398 (2019).

583. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

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

585. Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. vol. Springer-Verlag New York (2016).

586. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinforma. Oxf. Engl.* **33**, 2938–2940 (2017).

587. Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022).

588. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. Camb. Mass* **2**, 100141 (2021).

589. Li N, F. S., Carlson M & Pagès H. *AnnotationDbi: Manipulation of SQLite-Based Annotations in Bioconductor*. (2024).

590. Monje, M. *et al.* Roadmap for the Emerging Field of Cancer Neuroscience. *Cell* **181**, 219–222 (2020).

591. Winkler, F. *et al.* Cancer neuroscience: State of the field, emerging directions. *Cell* **186**, 1689–1707 (2023).

592. Shi, M., Chu, F., Zhu, F. & Zhu, J. Impact of Anti-amyloid-β Monoclonal Antibodies on the Pathology and Clinical Profile of Alzheimer's Disease: A Focus on Aducanumab and Lecanemab. *Front. Aging Neurosci.* **14**, 870517 (2022).

593. Mohan, V., Edamakanti, C. R. & Pathak, A. Editorial: Role of extracellular matrix in neurodevelopment and neurodegeneration. *Front. Cell. Neurosci.* **17**, 1135555 (2023).

594. Matsunuma, R. *et al.* DPYSL3 modulates mitosis, migration, and epithelial-to-mesenchymal transition in claudin-low breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11978–E11987 (2018).

595. Veillon, L., Fakih, C., Abou-El-Hassan, H., Kobeissy, F. & Mechref, Y. Glycosylation Changes in Brain Cancer. *ACS Chem. Neurosci.* **9**, 51–72 (2018).

596. Ying, Z. *et al.* Biglycan gene connects metabolic dysfunction with brain disorder. *Biochim. Biophys. Acta Mol. Basis Dis.* **1864**, 3679–3687 (2018).

597. Sunderland, A. *et al.* Biglycan and reduced glycolysis are associated with breast cancer cell dormancy in the brain. *Front. Oncol.* **13**, 1191980 (2023).

598. Haldrup, C. *et al.* Biomarker potential of ST6GALNAC3 and ZNF660 promoter hypermethylation in prostate cancer tissue and liquid biopsies. *Mol. Oncol.* **12**, 545–560 (2018).

599. Sung, P.-S., Lin, P.-Y., Liu, C.-H., Su, H.-C. & Tsai, K.-J. Neuroinflammation and Neurogenesis in Alzheimer's Disease and Potential Therapeutic Approaches. *Int. J. Mol. Sci.* **21**, 701 (2020).

600. Muste Sadurni, M. & Saponaro, M. Deregulations of RNA Pol II Subunits in Cancer. *Appl. Biosci.* **2**, 459–476 (2023).

601. Doi, T., Ogawa, H., Tanaka, Y., Hayashi, Y. & Maniwa, Y. Bex1 significantly contributes to the proliferation and invasiveness of malignant tumor cells. *Oncol. Lett.* **20**, 362 (2020).

602. Liang, A., Zhou, B. & Sun, W. Integrated genomic characterization of cancer genes in glioma. *Cancer Cell Int.* **17**, 90 (2017).

603. Yang, T., Li, X.-N., Li, X.-G., Li, M. & Gao, P.-Z. DNAJC6 promotes hepatocellular carcinoma progression through induction of epithelial-mesenchymal transition. *Biochem. Biophys. Res. Commun.* **455**, 298–304 (2014).

604. Brettrager, E. J., Meehan, A. W., Falany, C. N. & van Waardenburg, R. C. A. M. Sulfotransferase 4A1 activity facilitates sulfate-dependent cellular protection to oxidative stress. *Sci. Rep.* **12**, 1625 (2022).

605. Hossain, M. I. *et al.* SULT4A1 Protects Against Oxidative-Stress Induced Mitochondrial Dysfunction in Neuronal Cells. *Drug Metab. Dispos.* **47**, 949–953 (2019).

606. Newington, J. T. *et al.* Amyloid beta resistance in nerve cell lines is mediated by the Warburg effect. *PloS One* **6**, e19191 (2011).

607. Song, Q. *et al.* Single-cell sequencing reveals the landscape of the human brain metastatic microenvironment. *Commun. Biol.* **6**, 760 (2023).

608. Kim, S. *et al.* Upregulation of extracellular proteins in a mouse model of Alzheimer's disease. *Sci. Rep.* **13**, 6998 (2023).

609. Bowley, T. Y. *et al.* The RPL/RPS gene signature of melanoma CTCs associates with brain metastasis. *Cancer Res. Commun.* **2**, 1436–1448 (2022).

610. Garcia-Esparcia, P. *et al.* Altered machinery of protein synthesis is region- and stage-dependent and is associated with α-synuclein oligomers in Parkinson's disease. *Acta Neuropathol. Commun.* **3**, 76 (2015).

611. Martin, I. *et al.* Ribosomal protein s15 phosphorylation mediates LRRK2 neurodegeneration in Parkinson's disease. *Cell* **157**, 472–485 (2014).

612. Oh, M., Choi, I.-K. & Kwon, H. J. Inhibition of histone deacetylase1 induces autophagy. *Biochem. Biophys. Res. Commun.* **369**, 1179–1183 (2008).

613. Mulcahy Levy, J. M. & Thorburn, A. Autophagy in cancer: moving from understanding mechanism to improving therapy responses in patients. *Cell Death Differ.* **27**, 843–857 (2020).

614. Scrivo, A., Bourdenx, M., Pampliega, O. & Cuervo, A. M. Selective autophagy as a potential therapeutic target for neurodegenerative disorders. *Lancet Neurol.* **17**, 802–815 (2018).

615. Nurzat, Y. *et al.* Identification of Therapeutic Targets and Prognostic Biomarkers Among Integrin Subunits in the Skin Cutaneous Melanoma Microenvironment. *Front. Oncol.* **11**, 751875 (2021).

616. Abela, L. *et al.* Neurodevelopmental and synaptic defects in DNAJC6 parkinsonism, amenable to gene therapy. *Brain* **147**(6), 2023-2037 (2024).

617. Vidyadhara, D. J. *et al.* Dopamine transporter and synaptic vesicle sorting defects underlie auxilin-associated Parkinson's disease. *Cell Rep.* **42**, 112231 (2023).

618. Golden, L. C. & Voskuhl, R. The Importance of Studying Sex Differences in Disease: The Example of Multiple Sclerosis. *J. Neurosci. Res.* **95**, 633–643 (2017).

619. Pozzilli, C. *et al.* Diagnosis and treatment of progressive multiple sclerosis: A position paper. *Eur. J. Neurol.* **30**, 9–21 (2023).

620. Ahmed, M., Kim, H. J. & Kim, D. R. Maximizing the utility of public data. *Front. Genet.* **14**, (2023).

621. Yoong, S. L., Turon, H., Grady, A., Hodder, R. & Wolfenden, L. The benefits of data sharing and ensuring open sources of systematic review data. *J. Public Health* **44**, e582–e587 (2022).

622. Rahimzadeh, V. *et al.* Benefits of sharing neurophysiology data from the BRAIN Initiative Research Opportunities in Humans Consortium. *Neuron* **111**, 3710–3715 (2023).

623. Hendriks, S., Ramos, K. M. & Grady, C. Survey of Investigators About Sharing Human Research Data in the Neurosciences. *Neurology* **99**, e1314–e1325 (2022).

624. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).

625. Gomes, D. G. E. *et al.* Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. R. Soc. B Biol. Sci.* **289**, 20221113 (2022).

626. Xue, B., Khoroshevskyi, O., Gomez, R. A. & Sheffield, N. C. Opportunities and challenges in sharing and reusing genomic interval data. *Front. Genet.* **14**, 1155809 (2023).

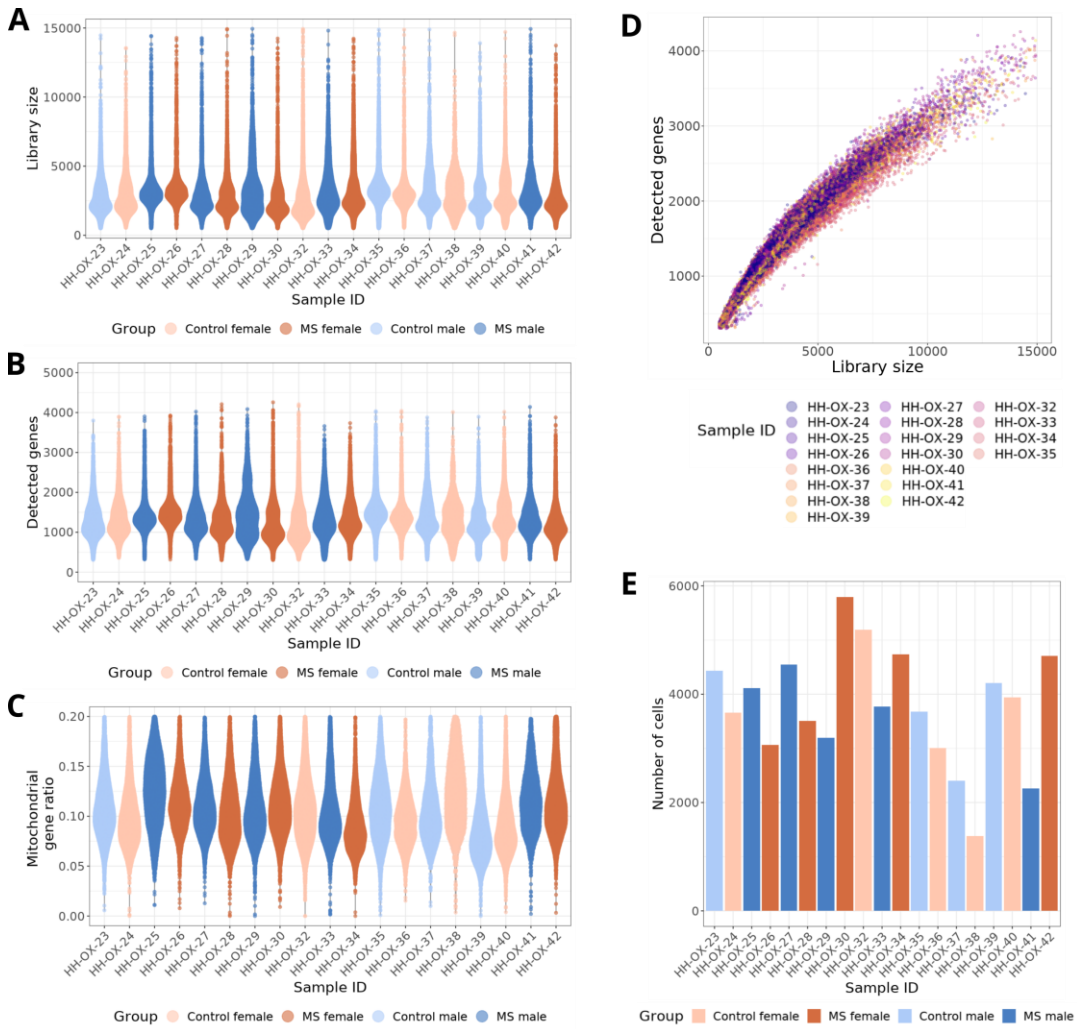
627. Ugochukwu, A. I. & Phillips, P. W. B. Open data ownership and sharing: Challenges and opportunities for application of FAIR principles and a checklist for data managers. *J. Agric. Food Res.* **16**, 101157 (2024).

628. Cervantes-Gracia, K., Chahwan, R. & Husi, H. Integrative OMICS Data-Driven Procedure Using a Derivatized Meta-Analysis Approach. *Front. Genet.* **13**, (2022).

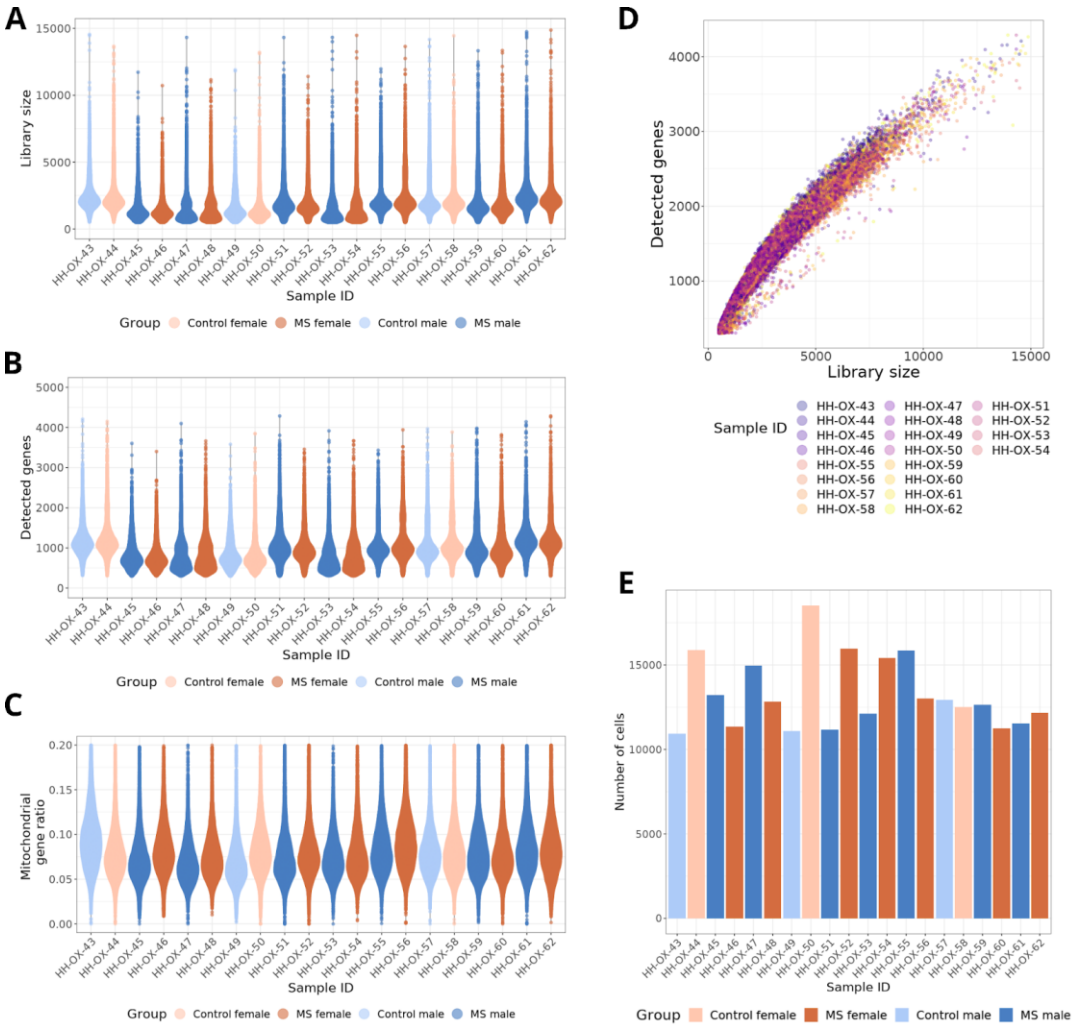
629. Vennou, K. E., Piovani, D., Kontou, P. I., Bonovas, S. & Bagos, P. G. Methods for multiple outcome meta-analysis of gene-expression data. *MethodsX* **7**, 100834 (2020).

10. Annexes

10.1. Supplementary material Study I

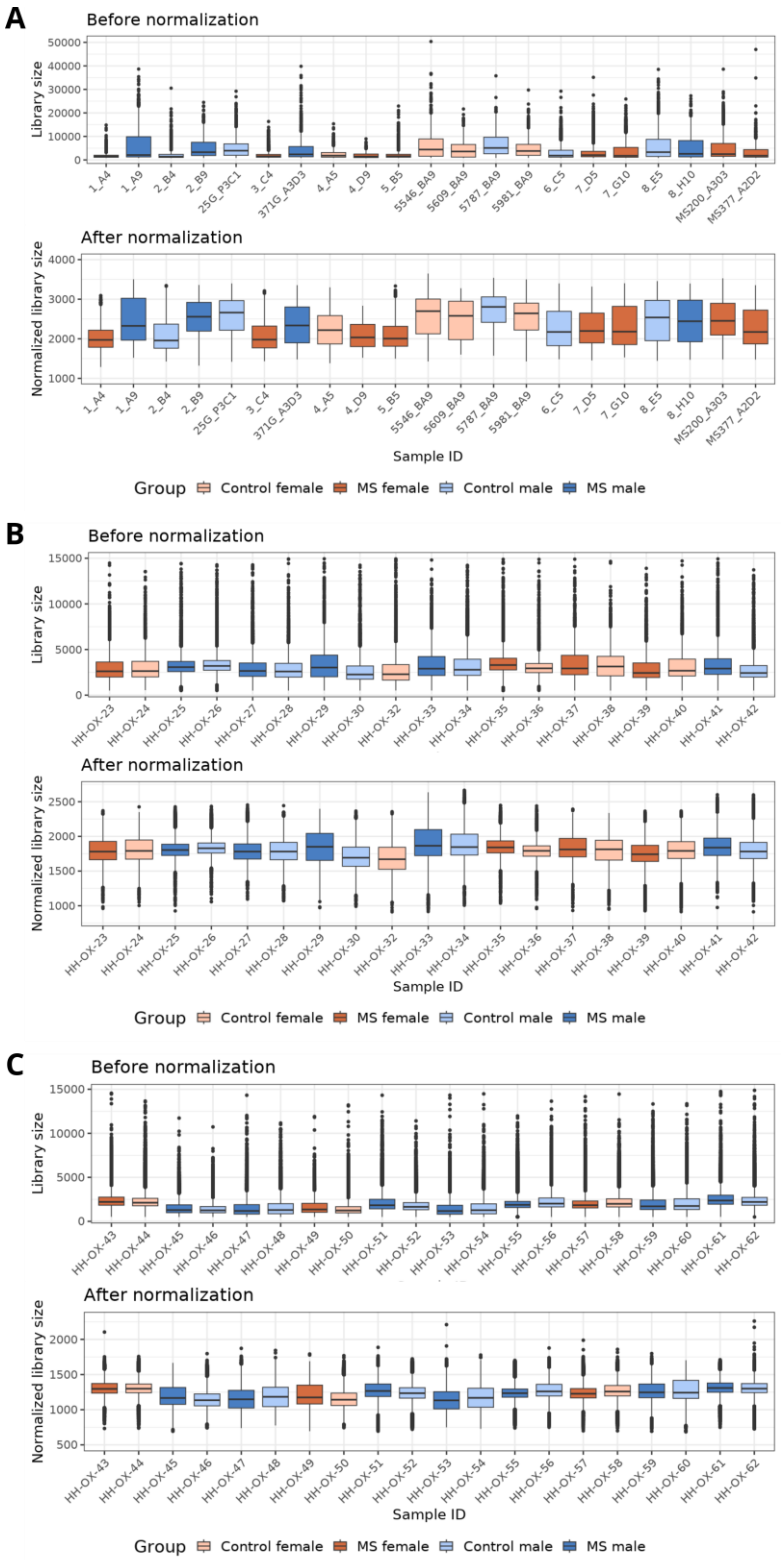


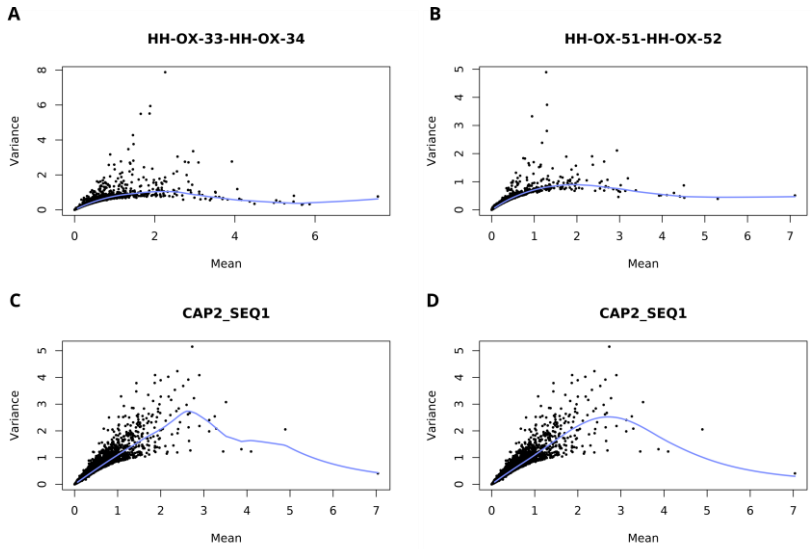
Supplementary Figure 3.S1. Quality control assessment of the RRMS-PBMCs dataset before and after filtering. (A-C) Representation of cells (dots) based on the sample of origin (X-axis) and according to the quality control metric value (Y-axis): (A) library size, (B) number of detected genes and (C) mitochondrial gene expression ratio. (D) Scatter plot of the library size versus the number of detected genes. Each dot represents a cell colored by the sample of origin. (E) Bar plot for the number of evaluated cells per sample. *ID: Identifier; MS: multiple sclerosis.*



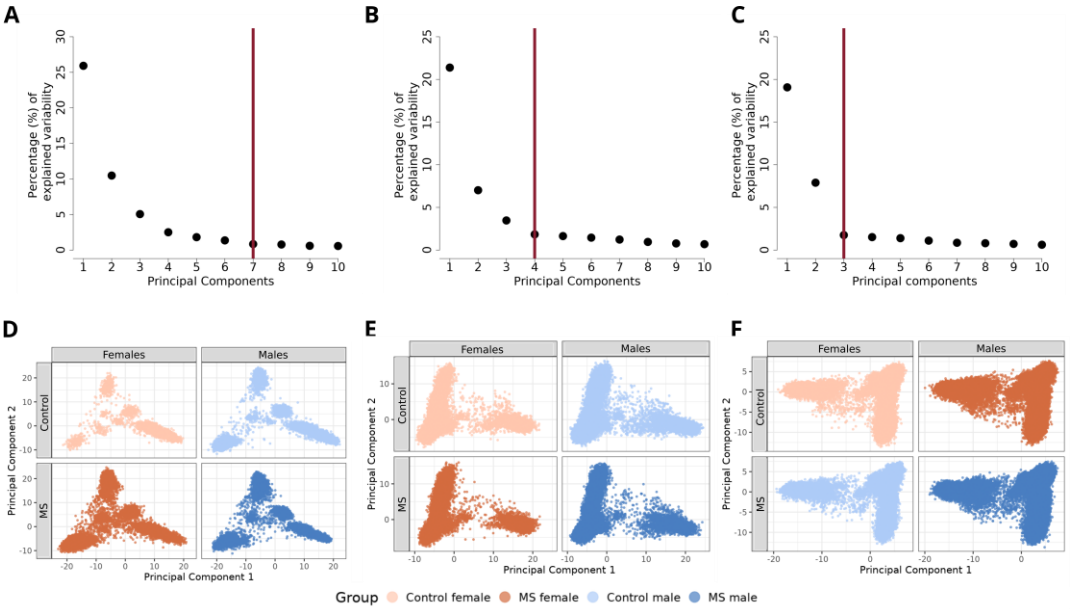
Supplementary Figure 3.S2. Quality control assessment of the PPMS-PBMCs dataset before and after filtering. (A-C) Representation of cells (dots) based on the sample of origin (X-axis) and according to the quality control metric value (Y-axis): (A) library size, (B) number of detected genes and (C) mitochondrial gene expression ratio. (D) Scatter plot of the library size versus the number of detected genes. Each dot represents a cell colored by the sample of origin. (E) Bar plot for the number of evaluated cells per sample. *ID: Identifier; MS: multiple sclerosis.*

Supplementary Figure 3-S3. Normalization results for (A) SPMS-CNS, (B) RRMS-PBMCs and (C) PPMS-PBMCs datasets. (*Next page*) Distribution of library sizes for the cells comprising each sample, shown before (top) and after (bottom) applying log-transformed deconvolution-based normalization. Colors are established based on the condition and sex of the individual. *MS: multiple sclerosis.*

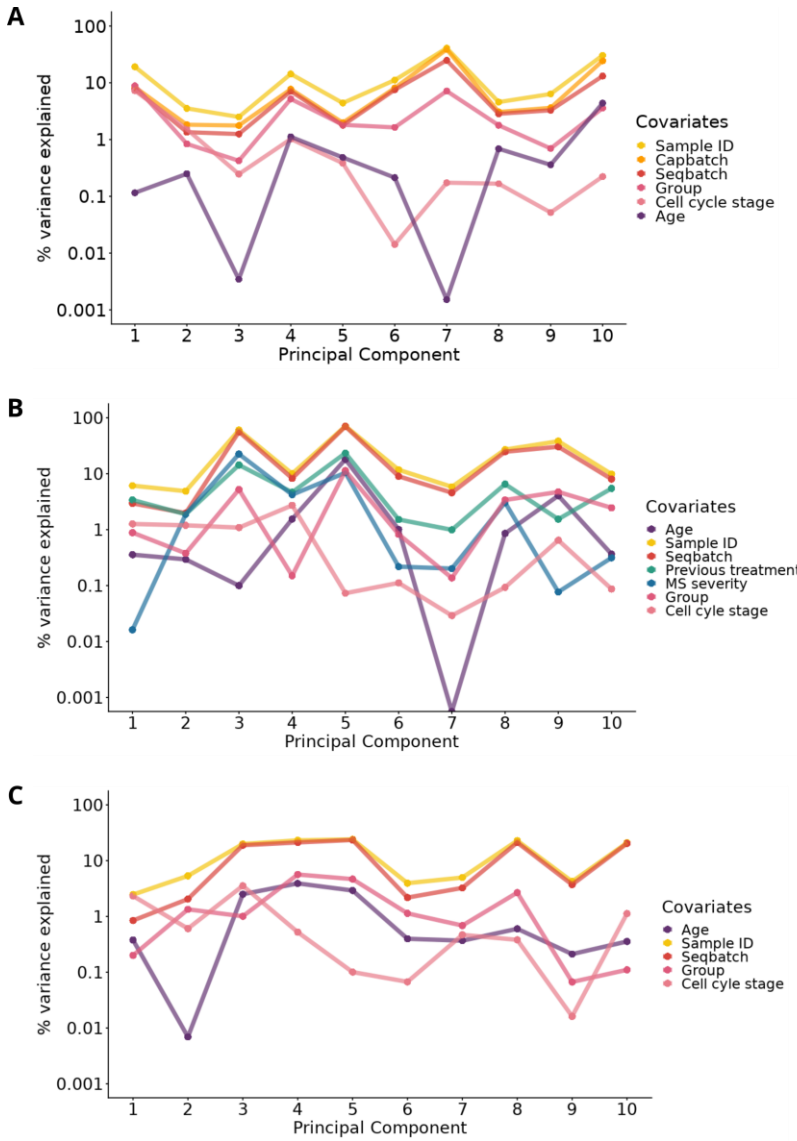

A
Before normalization
Library size
50000
40000
30000
20000
10000
0
1_A4
1_A9
2_B4
2_B9
25G_P3C1
3_C4
371G_A3D3
4_A5
4_D9
5_B5
5546_BA9
5609_BA9
5787_BA9
5981_BA9
6_C5
7_D5
7_G10
8_E5
8_H10
MS200_A303
MS377_A2D2
After normalization
Normalized library size
4000
3000
2000
1000
Sample ID
Group
Control female
MS female
Control male
MS male
B
Before normalization
Library size
15000
10000
5000
0
HH-OX-23
HH-OX-24
HH-OX-25
HH-OX-26
HH-OX-27
HH-OX-28
HH-OX-29
HH-OX-30
HH-OX-32
HH-OX-33
HH-OX-34
HH-OX-35
HH-OX-36
HH-OX-37
HH-OX-38
HH-OX-39
HH-OX-40
HH-OX-41
HH-OX-42
After normalization
Normalized library size
2500
2000
1500
1000
Sample ID
Group
Control female
MS female
Control male
MS male
C
Before normalization
Library size
15000
10000
5000
0
HH-OX-43
HH-OX-44
HH-OX-45
HH-OX-46
HH-OX-47
HH-OX-48
HH-OX-49
HH-OX-50
HH-OX-51
HH-OX-52
HH-OX-53
HH-OX-54
HH-OX-55
HH-OX-56
HH-OX-57
HH-OX-58
HH-OX-59
HH-OX-60
HH-OX-61
HH-OX-62
After normalization
Normalized library size
2000
1500
1000
500
Sample ID
Group
Control female
MS female
Control male
MS male



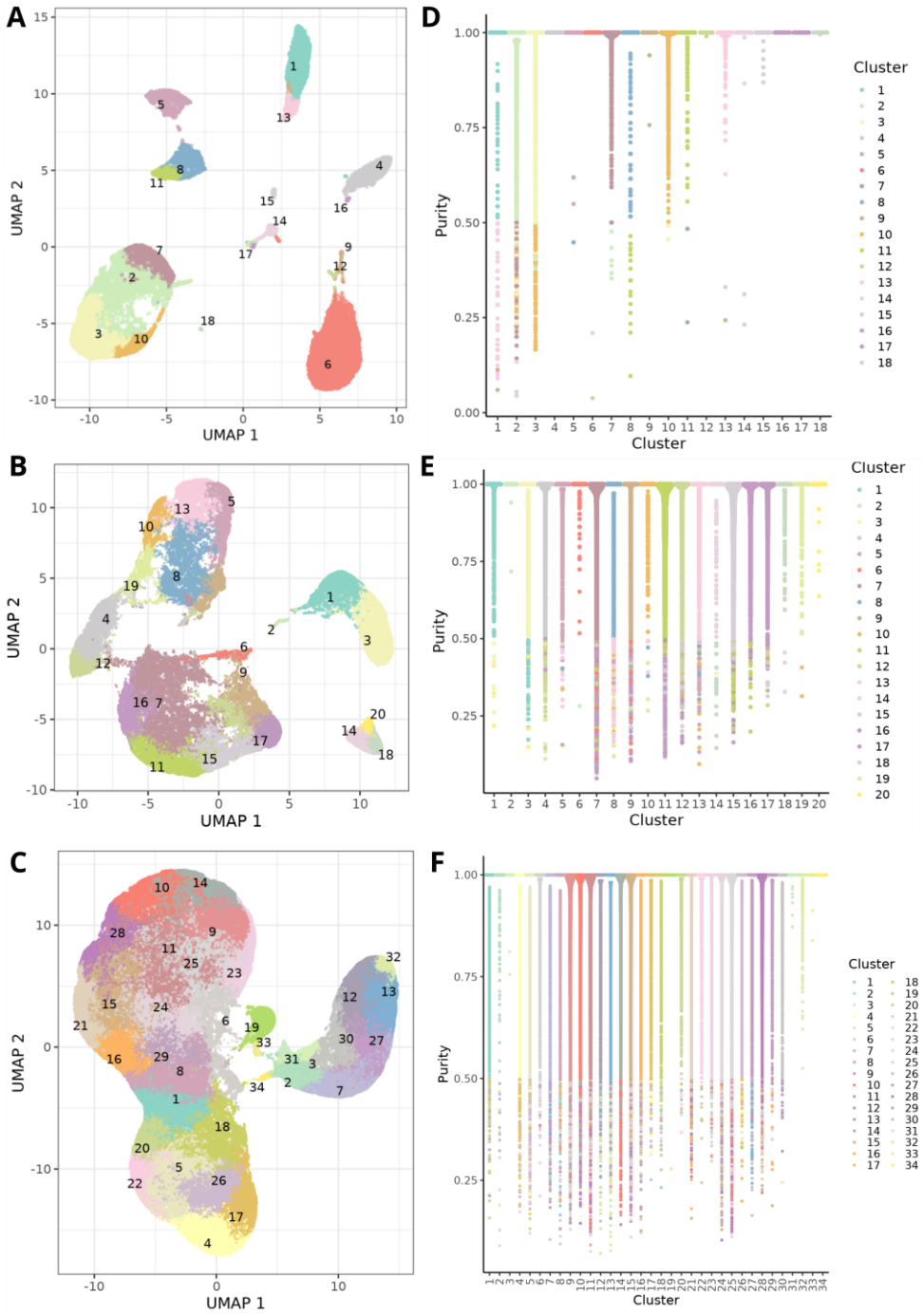
Supplementary Figure 3.S4. Graphical representations of the gene variance versus the gene mean levels from log-normalized data. Results are shown for the processing batches named (A) HH-OX-33-HH-OX-34 from RRMS-PBMCs dataset, (B) HH-OX-51-HH-OX-42 from PPMS-PBMCs dataset, and CAP2_SEQ1 from SPMS-CNS calculating technical variance trend (C) with and (D) without density weights. Blue solid line: fitted trend estimating the technical component of gene expression variance.



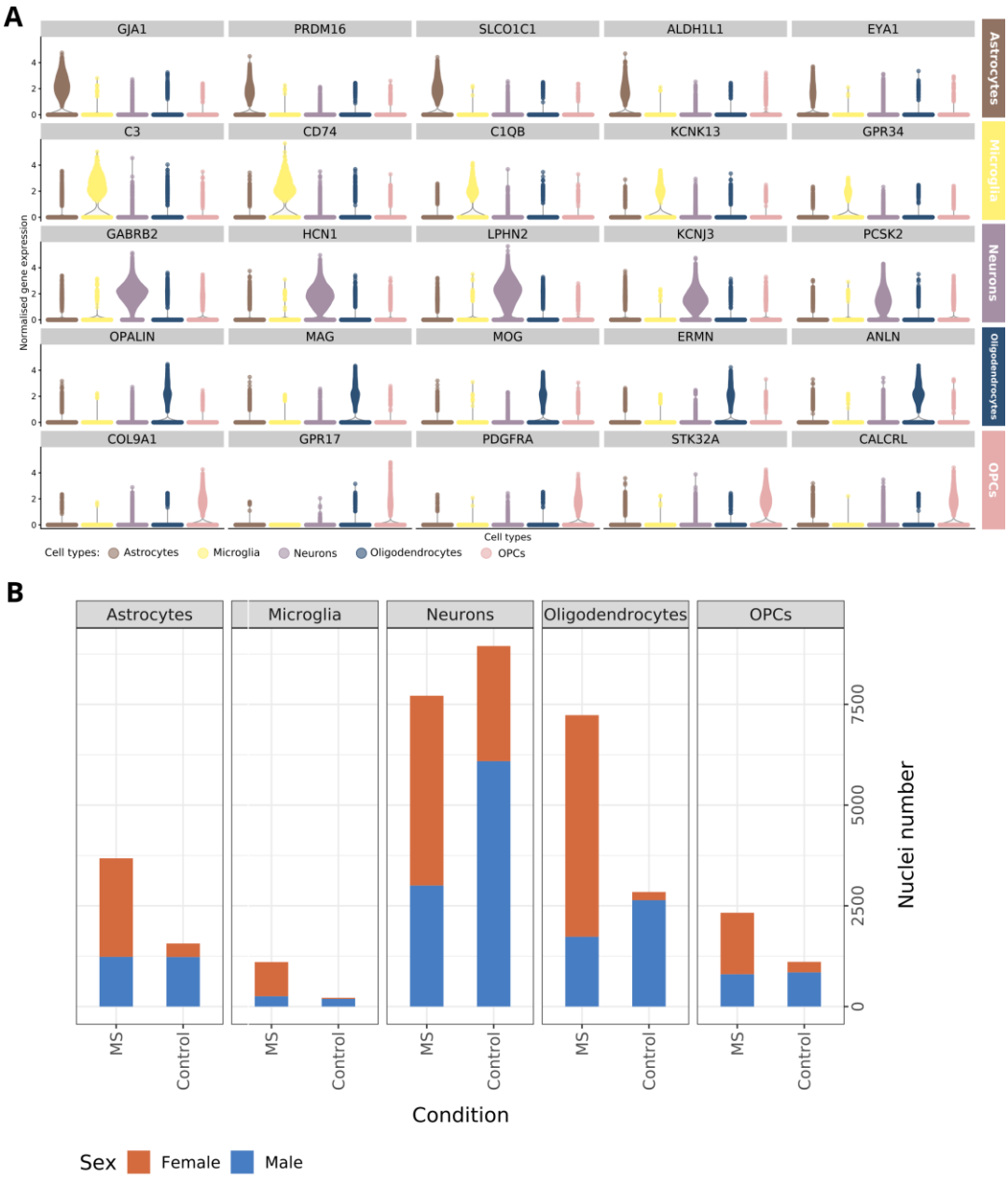
Supplementary Figure 3.S5. Scree plots and cell distribution based on the first two principal components. (A-C) Percentage of variance explained by each principal component for (A) SPMS-CNS, (B) RRMS-PBMCs and (C) PPMS-PBMCs datasets. Red horizontal line: variance cut-off point established by the elbow method. (D-F) Dot plot for the first (X-axis) and second (Y-axis) principal components. Each dot represents a cell. Cell distribution by condition and sex of the individual for (D) SPMS-CNS, (E) RRMS-PBMCs and (F) PPMS-PBMCs datasets. *MS: multiple sclerosis.*



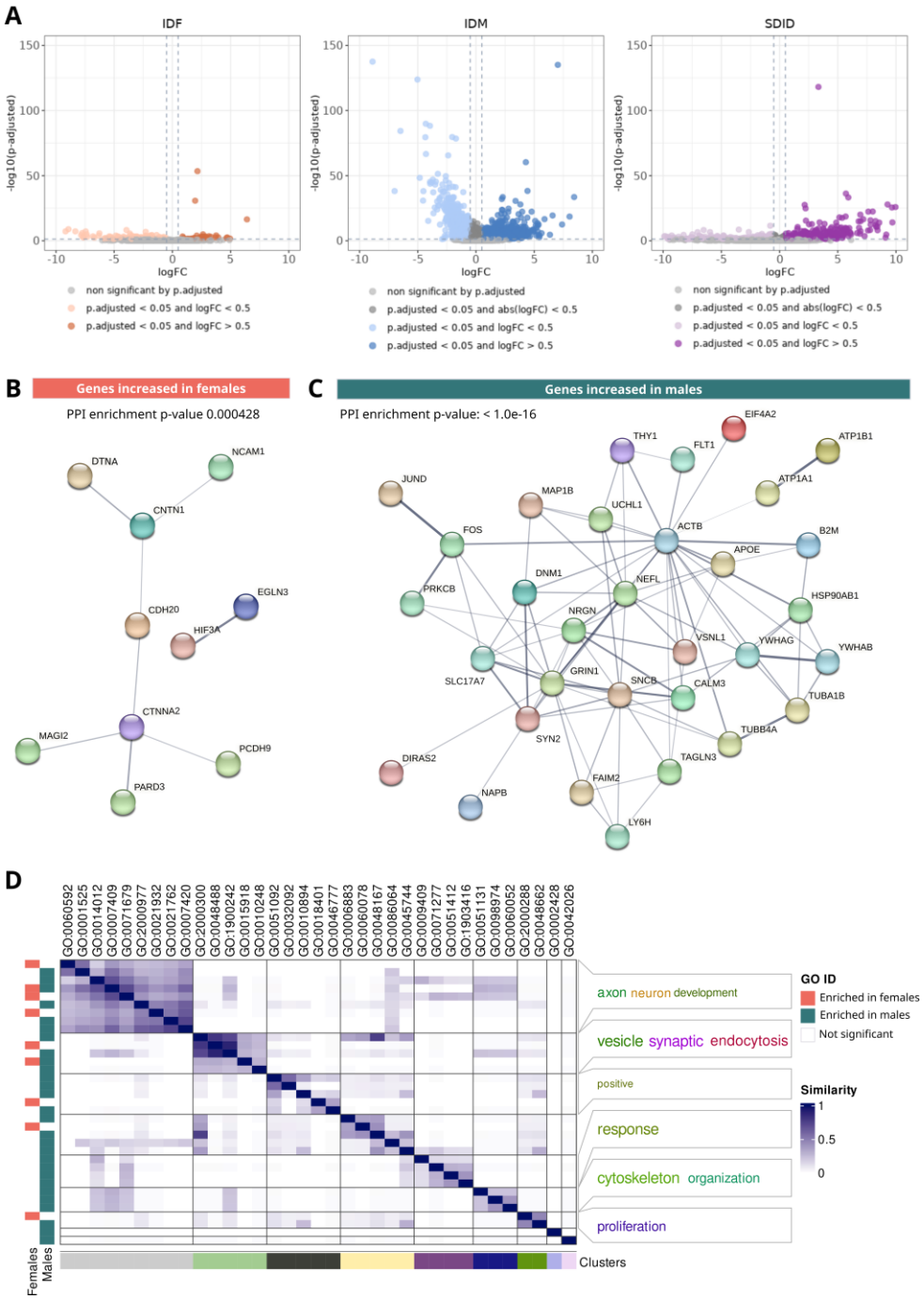
Supplementary Figure 3.S6. Proportion of variability explained in principal components by variables for (A) SPMS-CNS, (B) RRMS-PBMCs, and (C) PPMS-PBMCs datasets. Dot plots representing the percentage of variance (Y-axis) that can be explained by each reported variable across the principal components (X-axis). The variables assessed were sample identifier, lesion state, age, affected cerebral region, capture batch, sequencing batch and cell cycle for SPMS-CNS; sample identifier, age, previous treatments, MS severity, batch effect and cell cycle for RRMS-PBMCs; and sample identifier, age, batch effect and cell cycle for PPMS-PBMCs. *ID: identifier.*



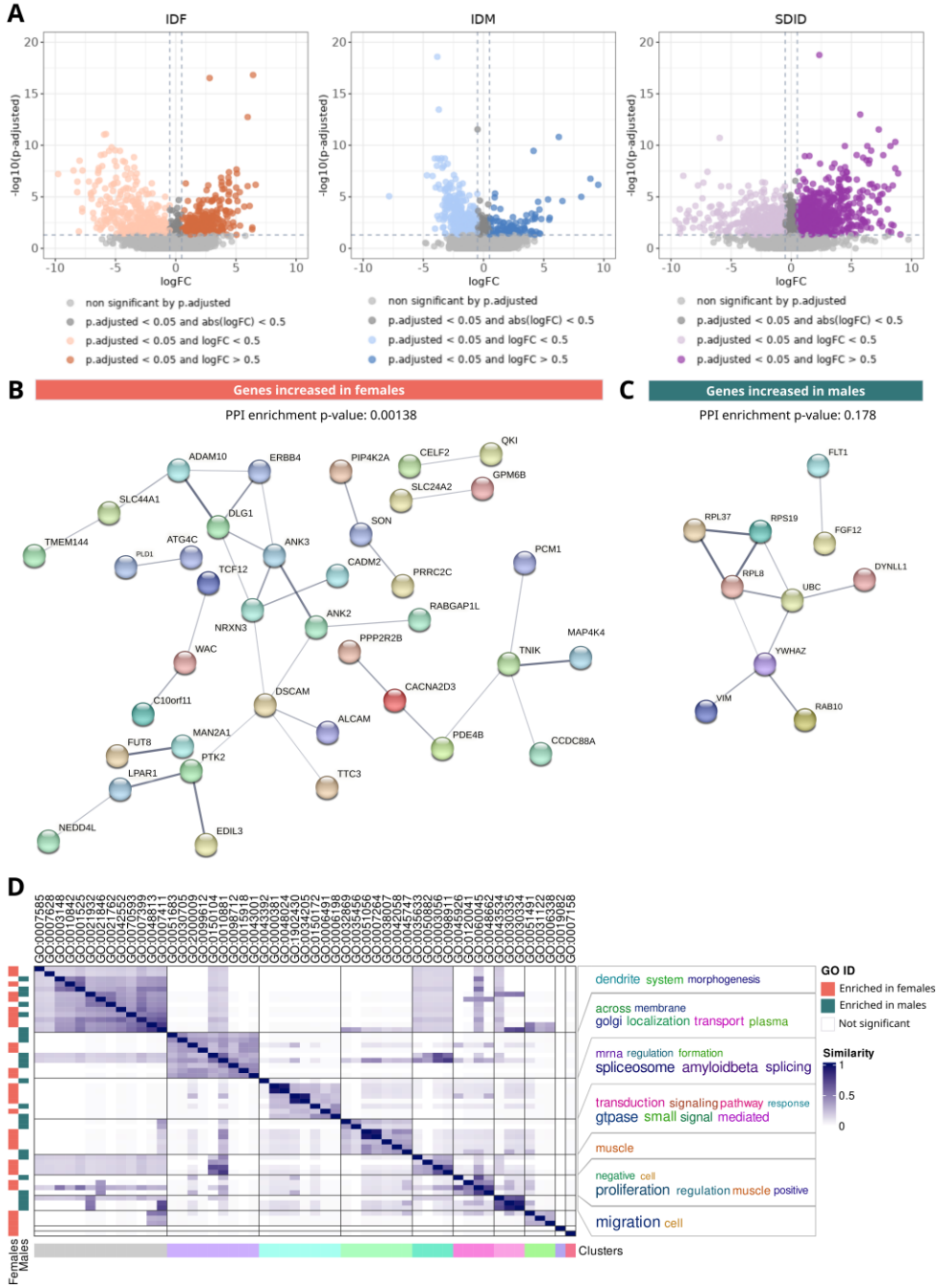
Supplementary Figure 3.S7. Cell identities assignment. (A) UMAP of the cellular landscape after clustering assignment. Each dot represents an individual cell, colored according to its assigned cell identity. (B) Purity characterization by cluster. Each dot represents a cell group based on their cell identity. The color is assigned based on the cluster most of its neighbors belong to. *UMAP: Uniform Manifold Approximation and Projection*.



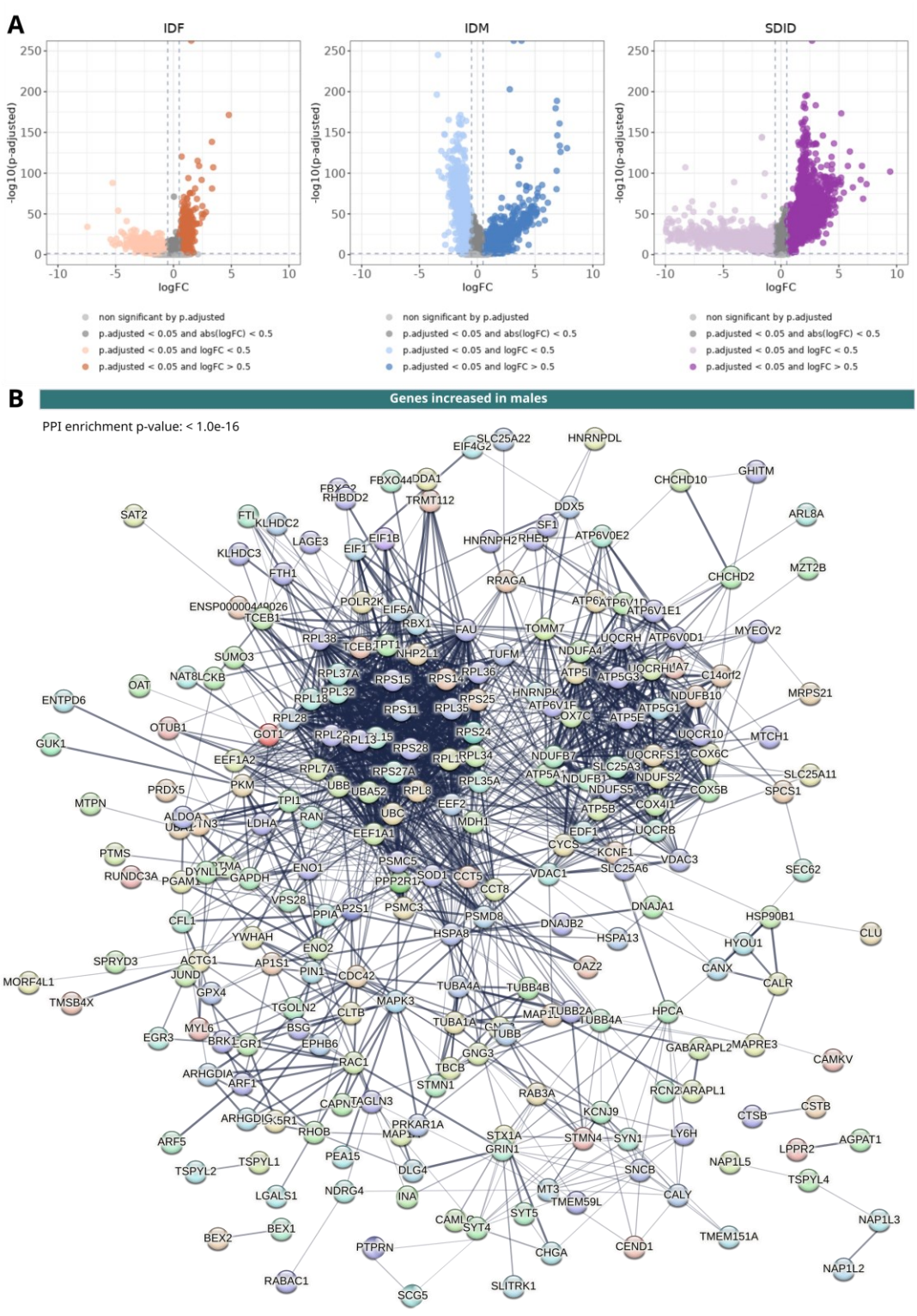
Supplementary Figure 3.S8. Cell type description for SPMS-CNS. (A) Expression pattern of marker genes (Y-axis) in astrocytes, microglia, neurons, oligodendrocytes and oligodendrocyte precursor cells (OPCs) (Y-axis). Each dot represents a cell, colored according to the cell type noted. Marker genes have been selected from bibliographic research. (B) Number of cells identified for each cell type (X-axis) based on condition (Y-axis) and sex *(color). MS: multiple sclerosis; OPCs: oligodendrocyte precursor cells.*

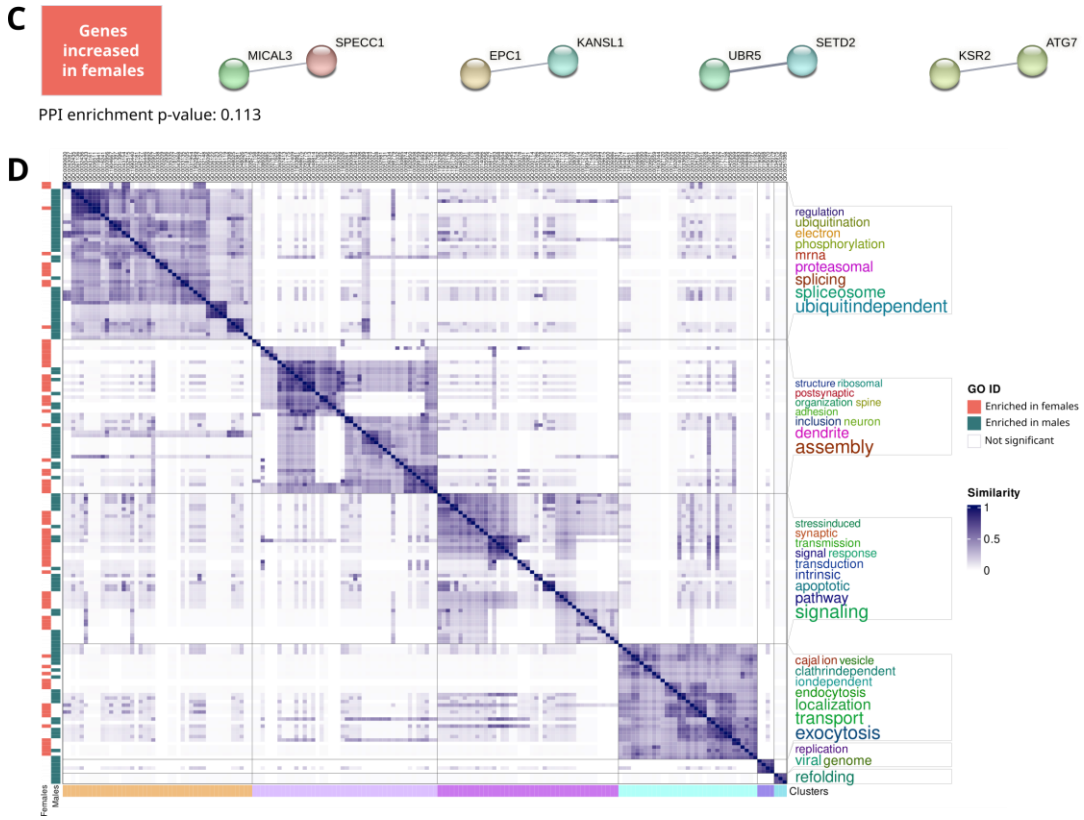


Supplementary Figure 3.S9. Astrocyte sex-differential atlas in secondary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.

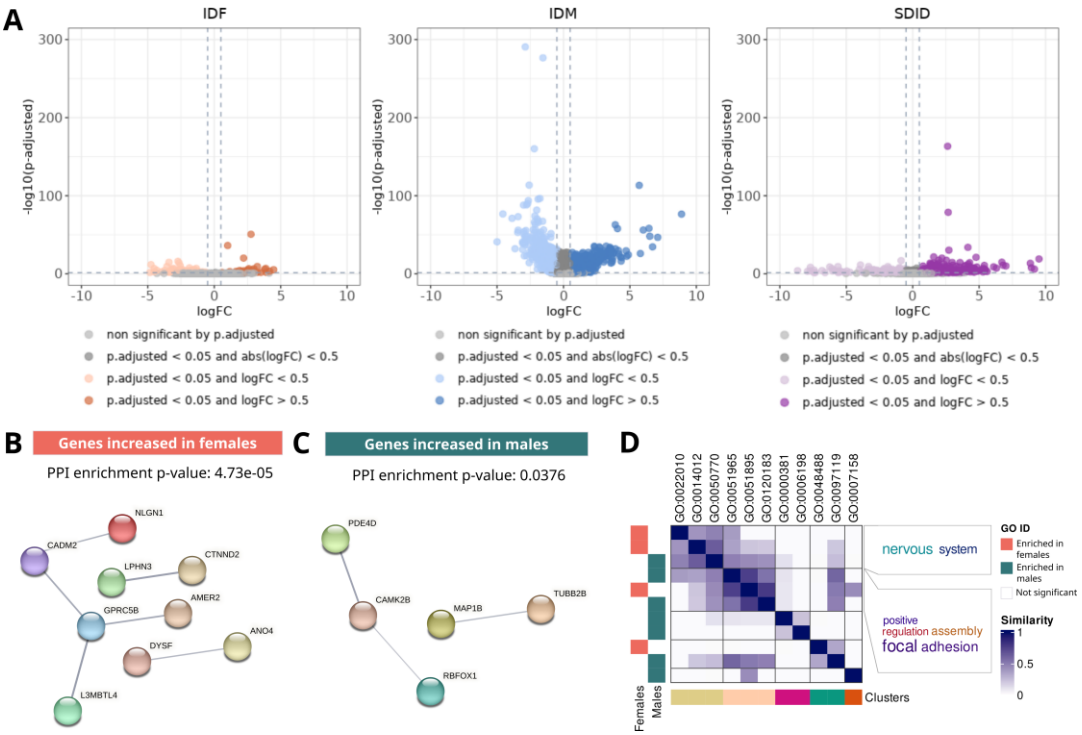


Supplementary Figure 3.S10. Microglia sex-differential atlas in secondary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.

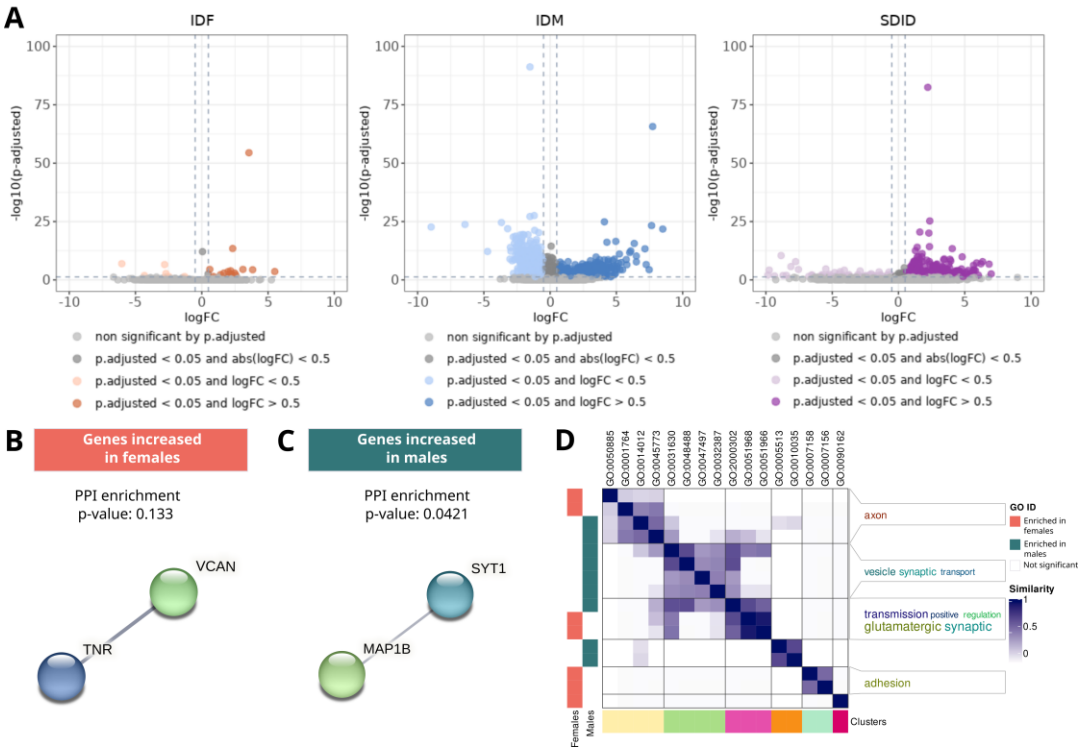

A
IDF
IDM
SDID
-log10(p-adjusted)
logFC
non significant by p.adjusted
p.adjusted < 0.05 and abs(logFC) < 0.5
p.adjusted < 0.05 and logFC < 0.5
p.adjusted < 0.05 and logFC > 0.5
B
Genes increased in males
PPI enrichment p-value: < 1.0e-16



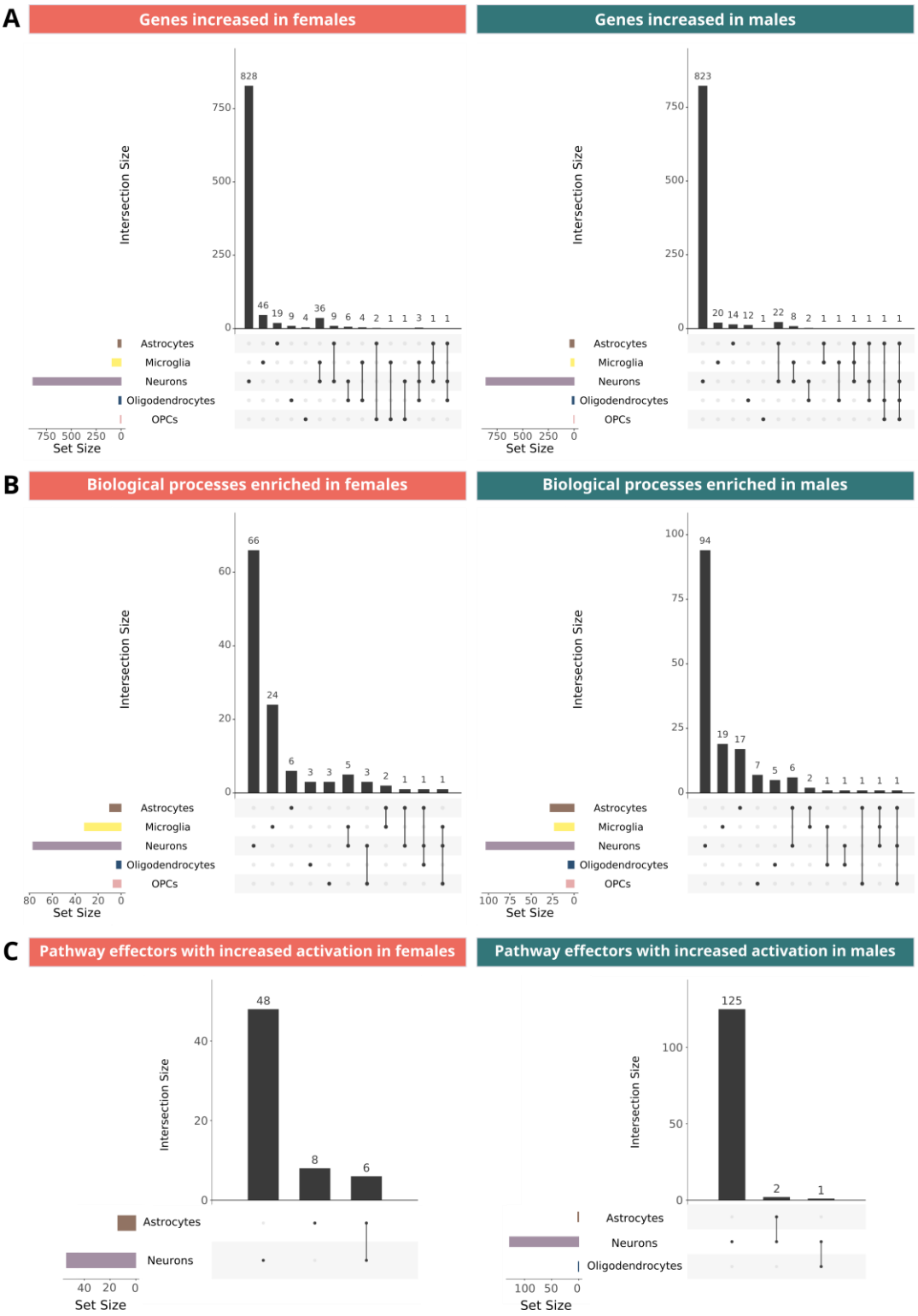
Supplementary Figure 3.S11. Neurons sex-differential atlas in secondary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in males (B, green) and females (C, orange) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* Due to the high number of significant results, subclusters were calculated with the following specifications: 1) community search with the fastgreedy algorithm, 2) discarding of subclusters with a size smaller than 0.7, 3) disable text mining option and 4) genes set with an absolute magnitude of change higher of 4. (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



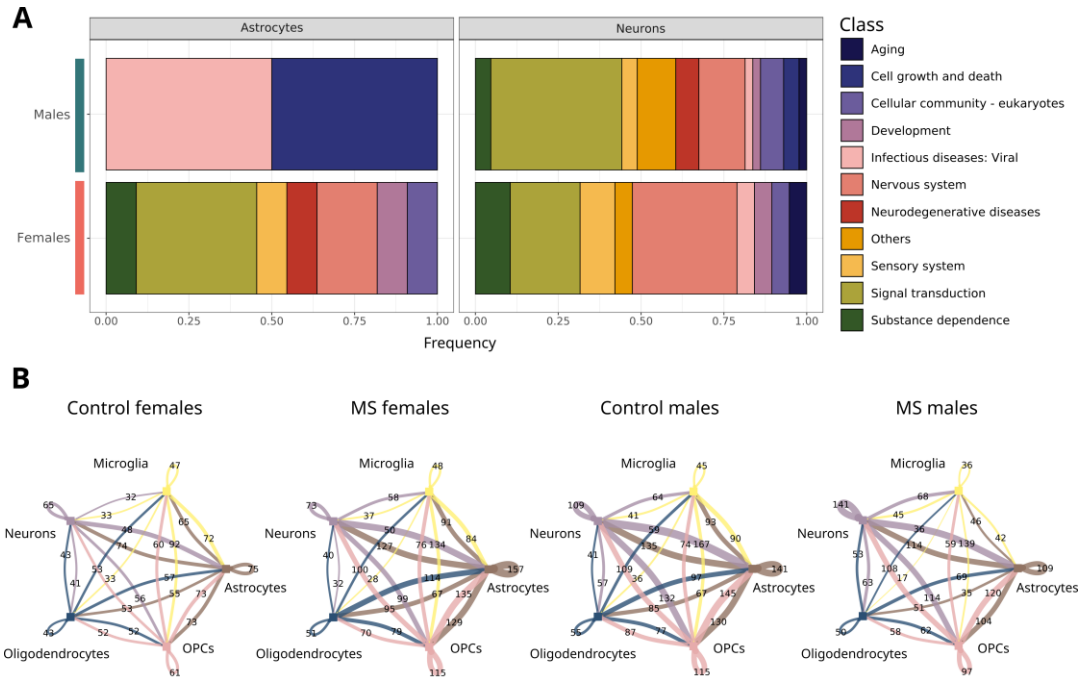
Supplementary Figure 3.S12. Oligodendrocytes sex-differential atlas in secondary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



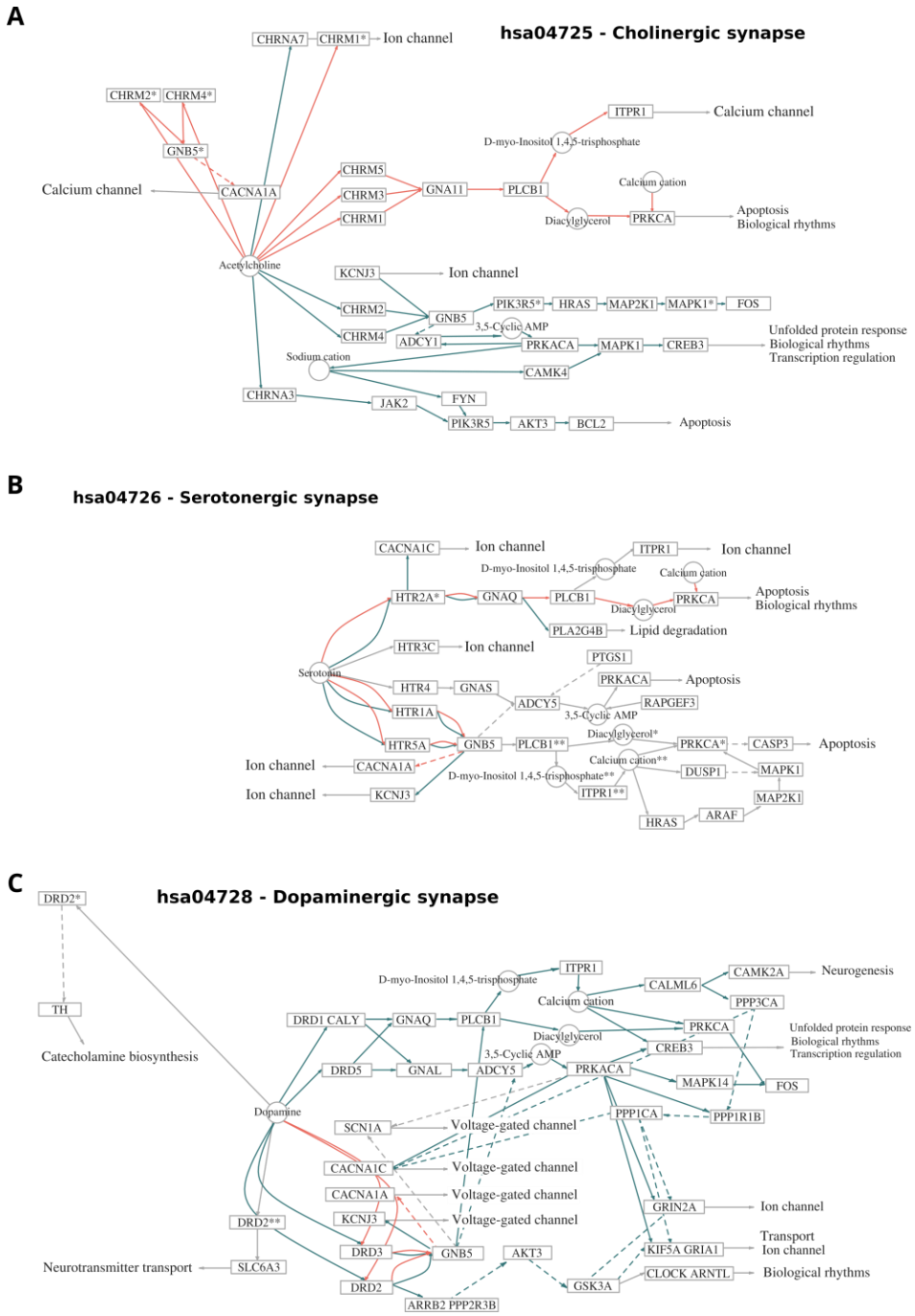
Supplementary Figure 3.S13. Oligodendrocyte precursor cells sex-differential atlas in secondary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.


A
Genes increased in females
Intersection Size
828
46
19
9
4
36
9
6
4
2
1
1
3
1
1
Astrocytes
Microglia
Neurons
Oligodendrocytes
OPCs
750 500 250 0
Set Size
Genes increased in males
Intersection Size
823
20
14
12
1
22
8
2
1
1
1
1
1
Astrocytes
Microglia
Neurons
Oligodendrocytes
OPCs
750 500 250 0
Set Size
B
Biological processes enriched in females
Intersection Size
66
24
6
3
3
5
3
2
1
1
1
Astrocytes
Microglia
Neurons
Oligodendrocytes
OPCs
80 60 40 20 0
Set Size
Biological processes enriched in males
Intersection Size
94
19
17
7
5
6
2
1
1
1
1
1
Astrocytes
Microglia
Neurons
Oligodendrocytes
OPCs
100 75 50 25 0
Set Size
C
Pathway effectors with increased activation in females
Intersection Size
48
8
6
Astrocytes
Neurons
40 20 0
Set Size
Pathway effectors with increased activation in males
Intersection Size
125
2
1
Astrocytes
Neurons
Oligodendrocytes
100 50 0
Set Size

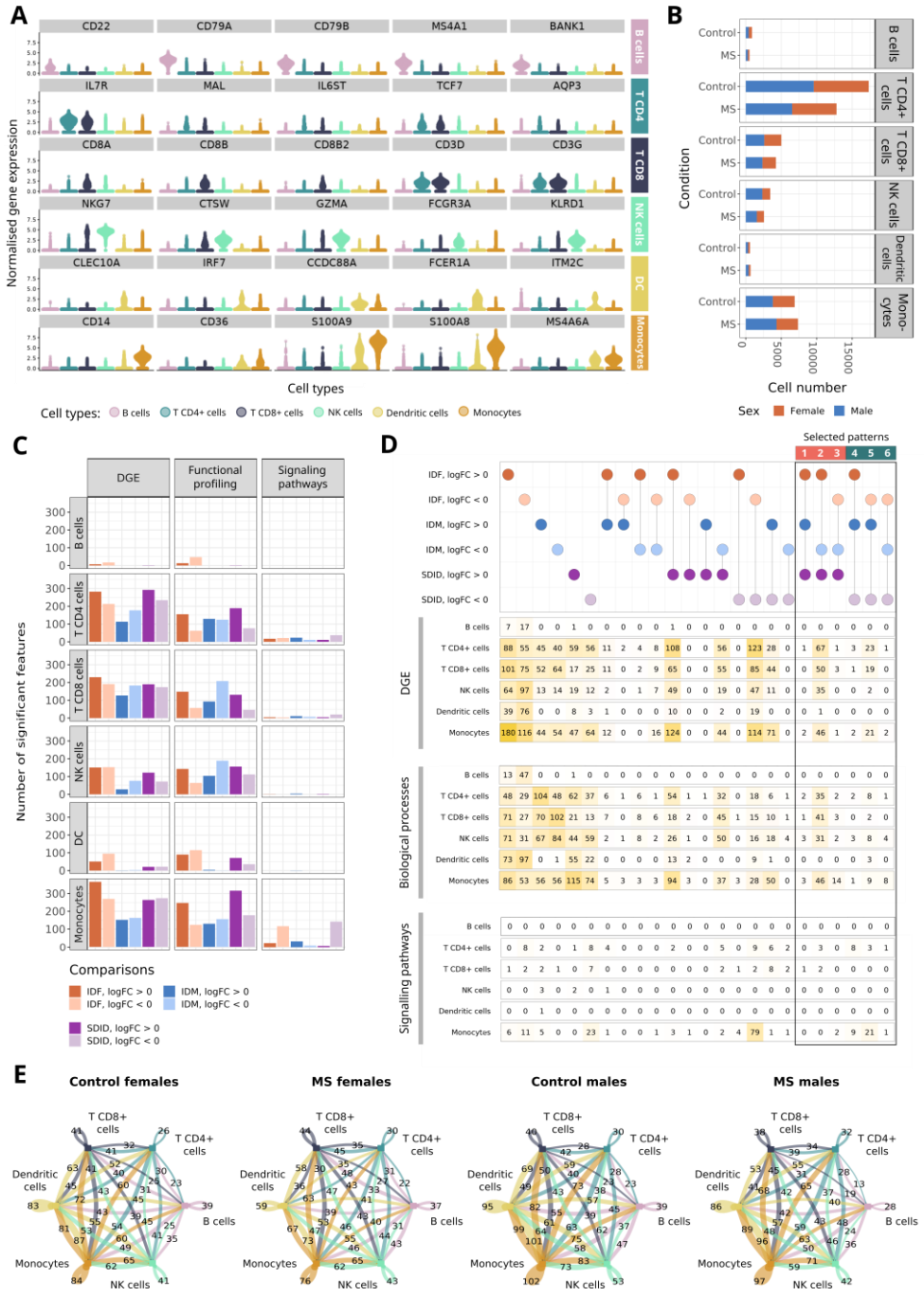
Supplementary Figure 3.S14. Intersection of significant results among brain cell types for secondary progressive multiple sclerosis. (*Previous page*) Upset plots for significant sex-differential (A) genes, (C) functions and (C) effectors of signaling pathways in the three assessed comparisons: IDF: impact of disease in females (MS females vs control females); IDM: impact of disease in males (MS males vs control males); SDID: sex differential impact of disease ((MS females vs control females) - (MS males vs control males)). *MS: multiple sclerosis; OPCs: oligodendrocyte precursor cells.*



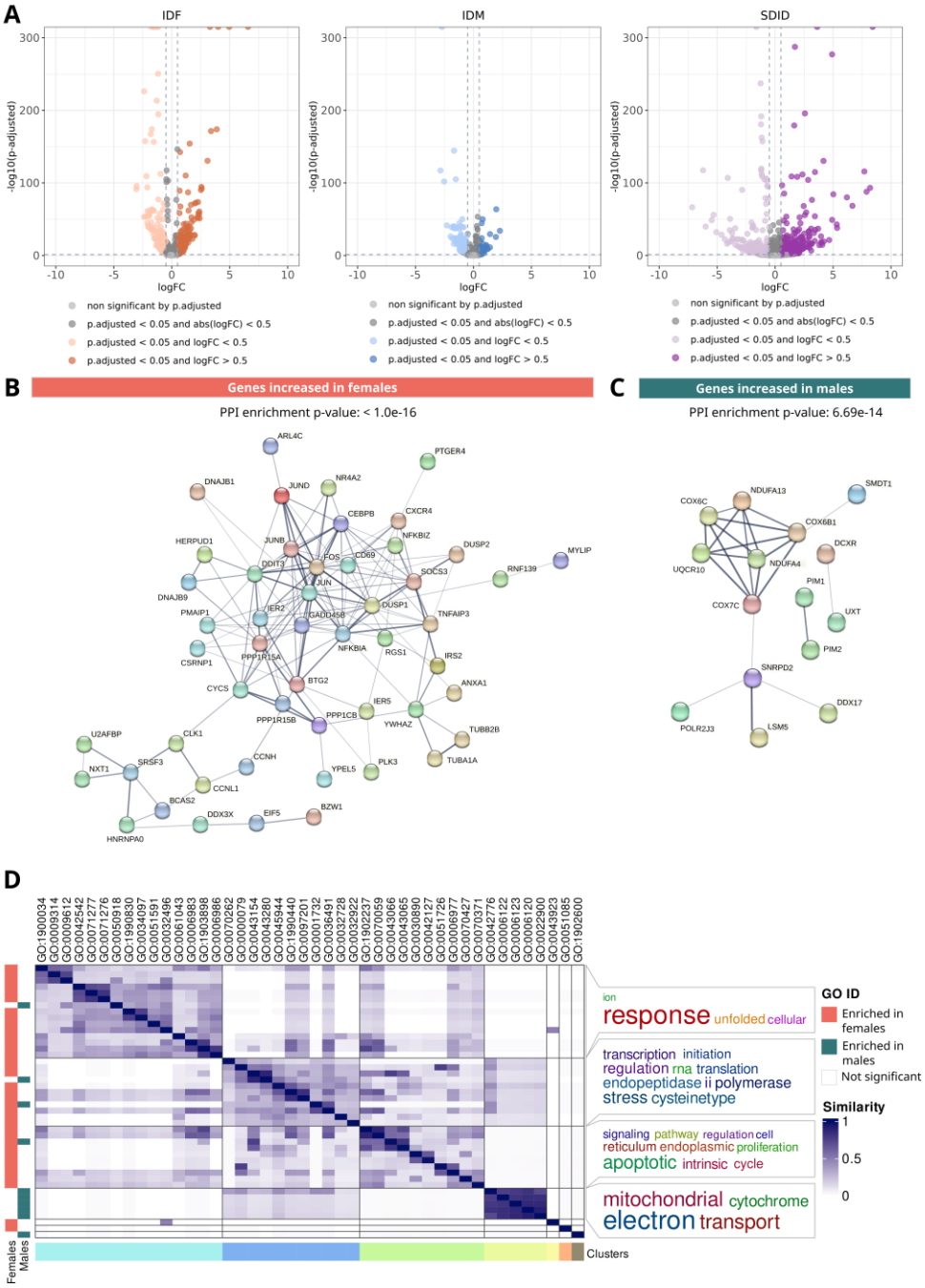
Supplementary Figure 3.S15. Significant results in the analyses of (A) signaling pathway effector activation and (B) cell-cell communication for the secondary progressive form of multiple sclerosis. (A) Relative frequency distribution of KEGG signaling pathways significantly activated in females (orange) and males (green) by cell type. Individual terms have been classified into general categories (see legend). No significant results have been obtained for microglia and oligodendrocyte precursor cells. Oligodendrocytes presented more activated the *Tight junction* pathway in males. (D) Number of significant cell-cell interactions by group. Color indicates the cell type providing the ligand protein. The thickness of the interaction corresponds to the magnitude of interactions. *KEGG: Kyoto Encyclopedia of Genes and Genomes; MS: multiple sclerosis; OPCs: oligodendrocyte precursor cells.*



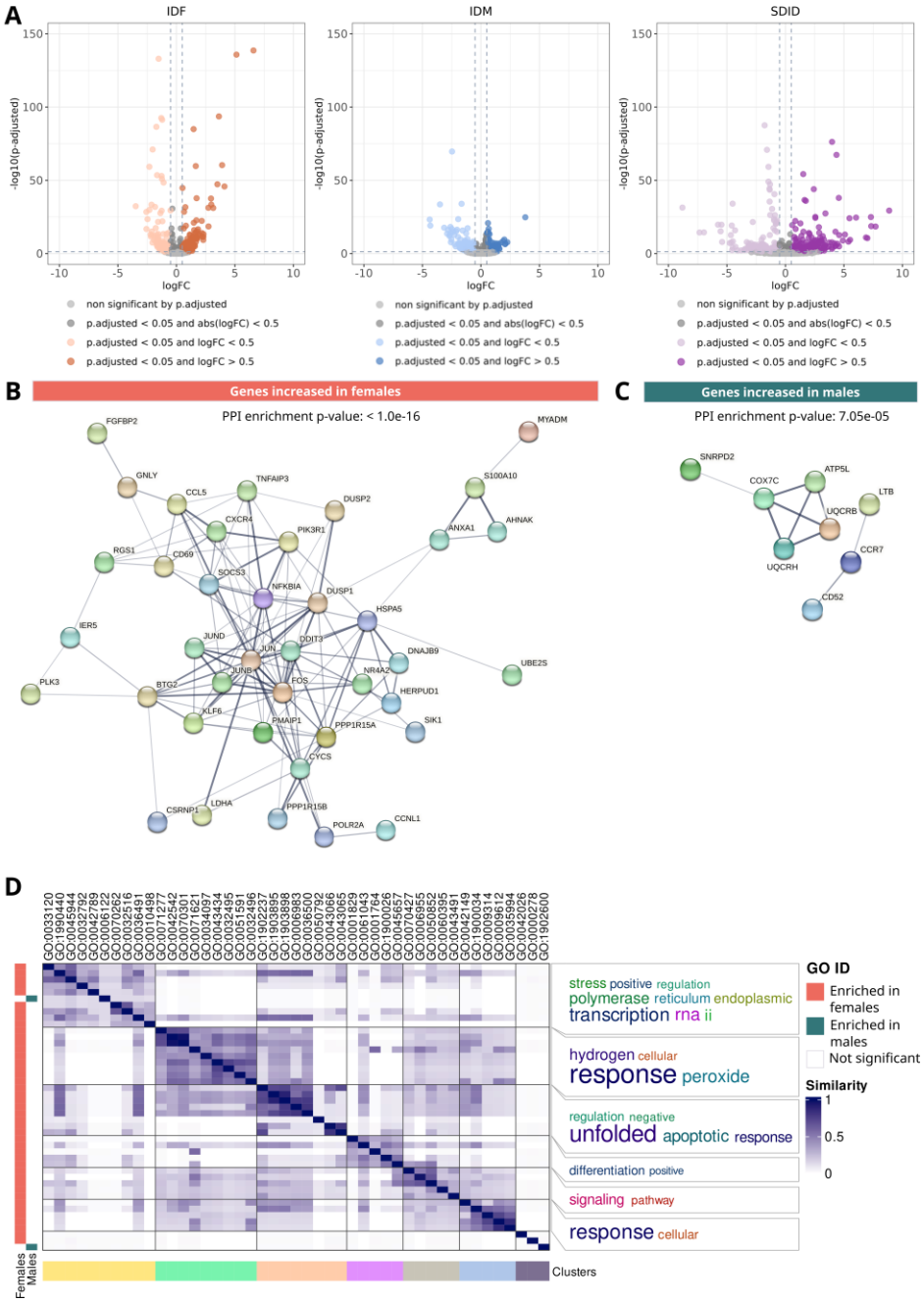
Supplementary Figure 3.S16. Sex differential signaling pathways of (A) cholinergic, (B) serotonergic and (C) dopaminergic synapses in neurons from secondary progressive multiple sclerosis. Nodes represent proteins of the signaling pathway and edges the interactions between nodes. Effector proteins are the last node in each subpathway (arrow point). The effector nodes point to the biological functions they exerted. Orange edges: subpathways with increased activation in females; green edges: subpathways with increased activation in males.



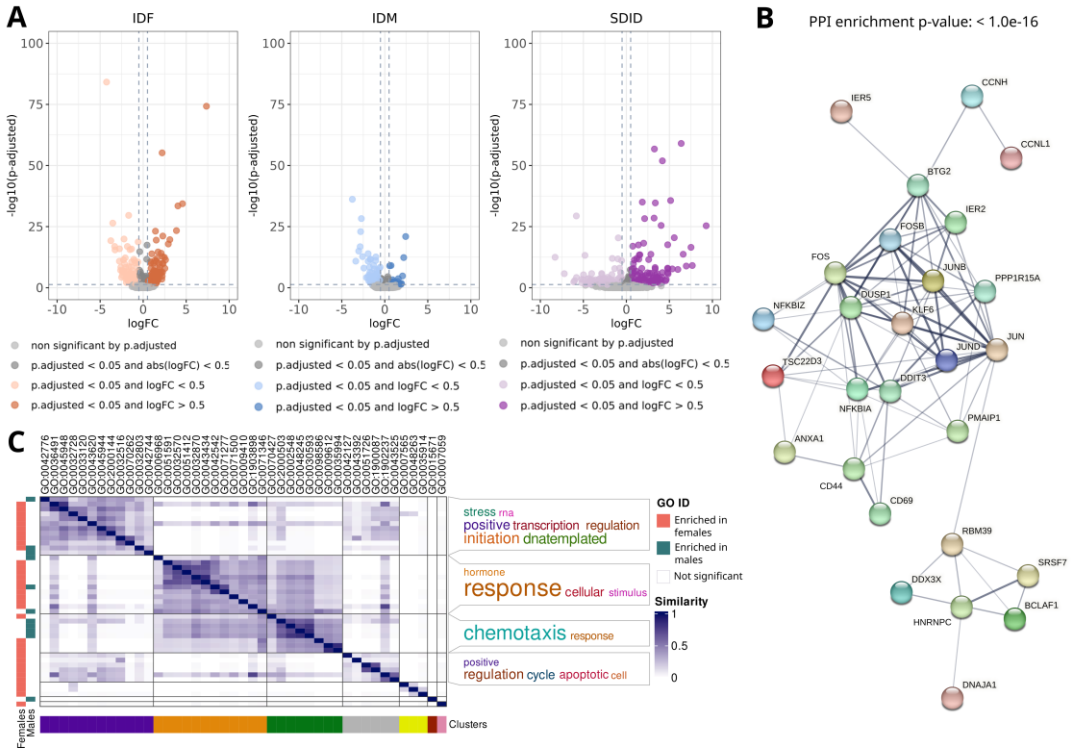
Supplementary Figure 3.S17. Summary of significant sex differences in relapsing-remitting multiple sclerosis. (A) Expression pattern of marker genes. (B) Number of cells identified for each cell type based on condition and sex. *MS: multiple sclerosis.* (C) Number of significant features by cell type, analysis and comparison tested. (D) Upset map of significant features for each cell type separated by comparison and direction of change (logFC). (E) Number of significant cell-cell interactions by group. Color indicates the cell type providing the ligand protein. The thickness of the interaction corresponds to the magnitude of interactions.



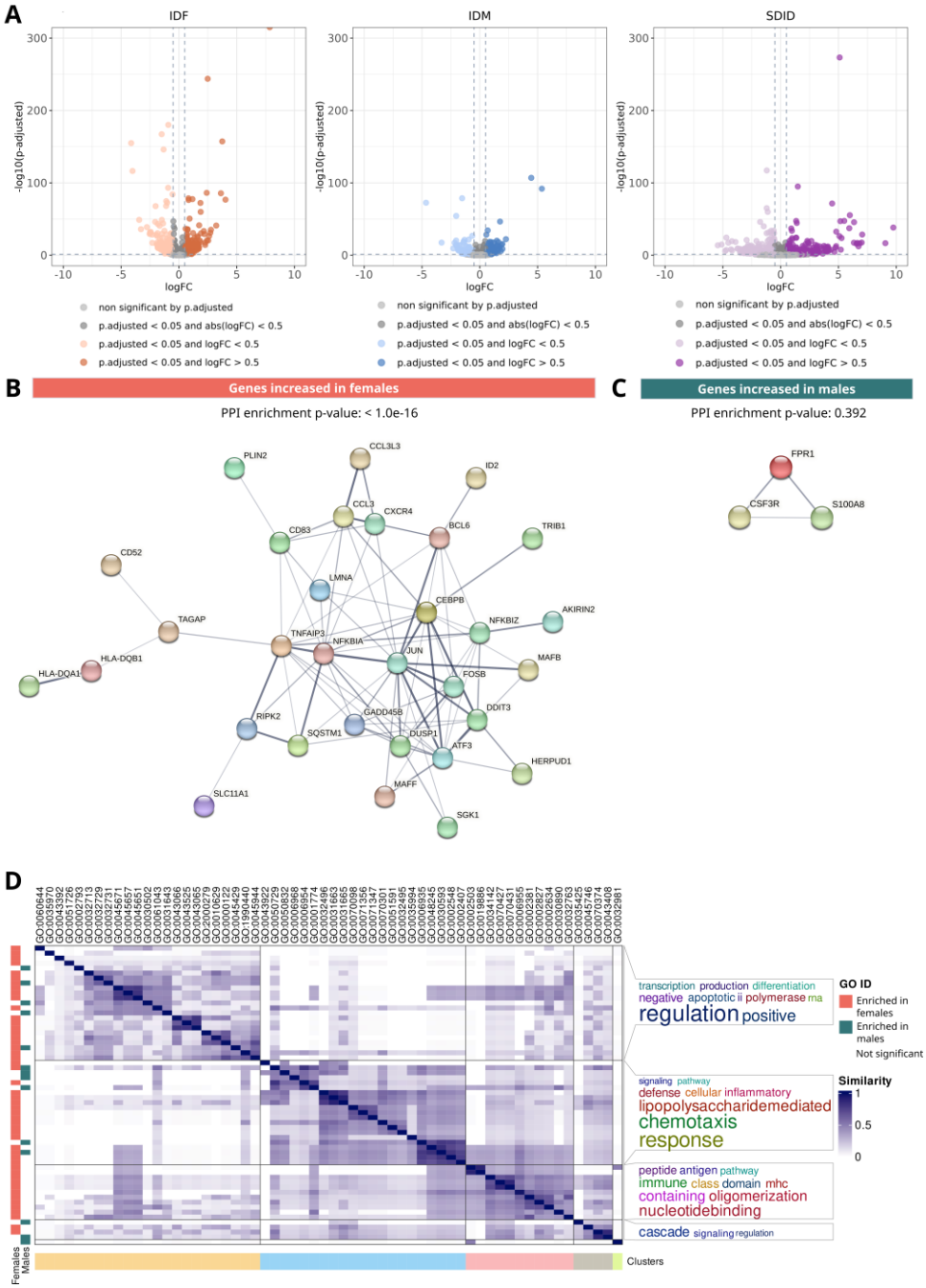
Supplementary Figure 3.S18. CD4+ T cells sex-differential atlas in relapsing-remitting multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



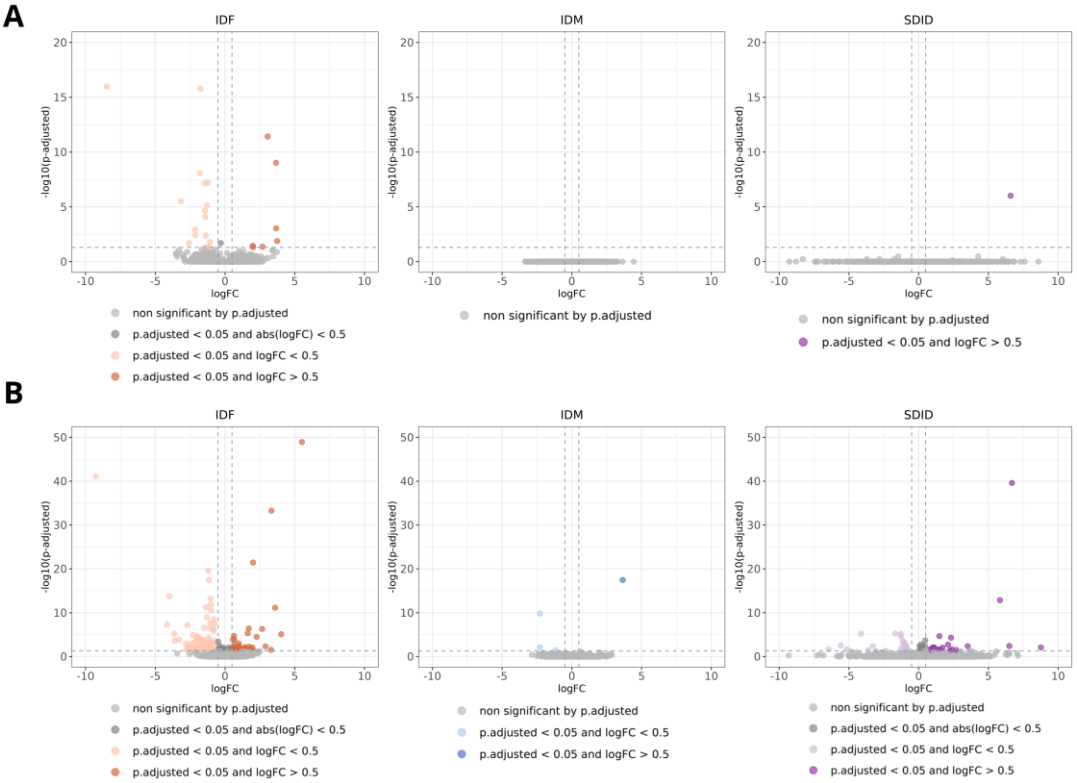
Supplementary Figure 3.S19. CD8+ T cells sex-differential atlas in relapsing-remitting multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



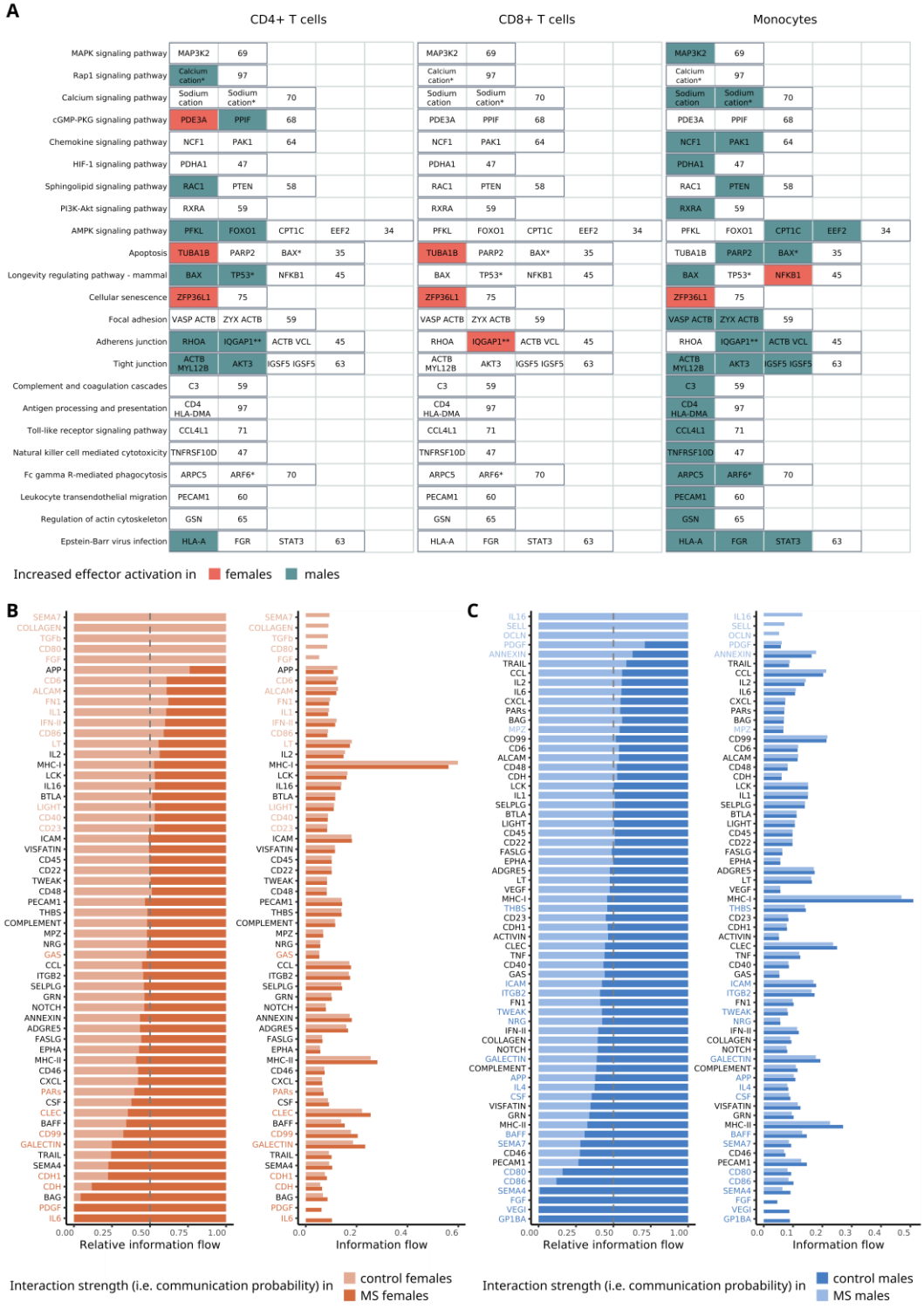
Supplementary Figure 3.S20. NK cells sex-differential atlas in relapsing-remitting multiple sclerosis. (A) Volcano plots of differential gene expression results. (B) Protein-protein interaction network of genes increased in females from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* Males only have increased CCL4 and HBA2 genes. (C) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity. *NK: natural killer.*



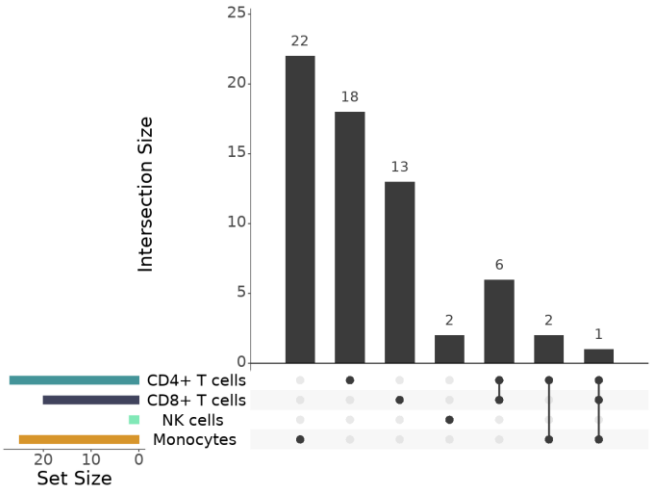
Supplementary Figure 3.S21. Monocytes sex-differential atlas in relapsing-remitting multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



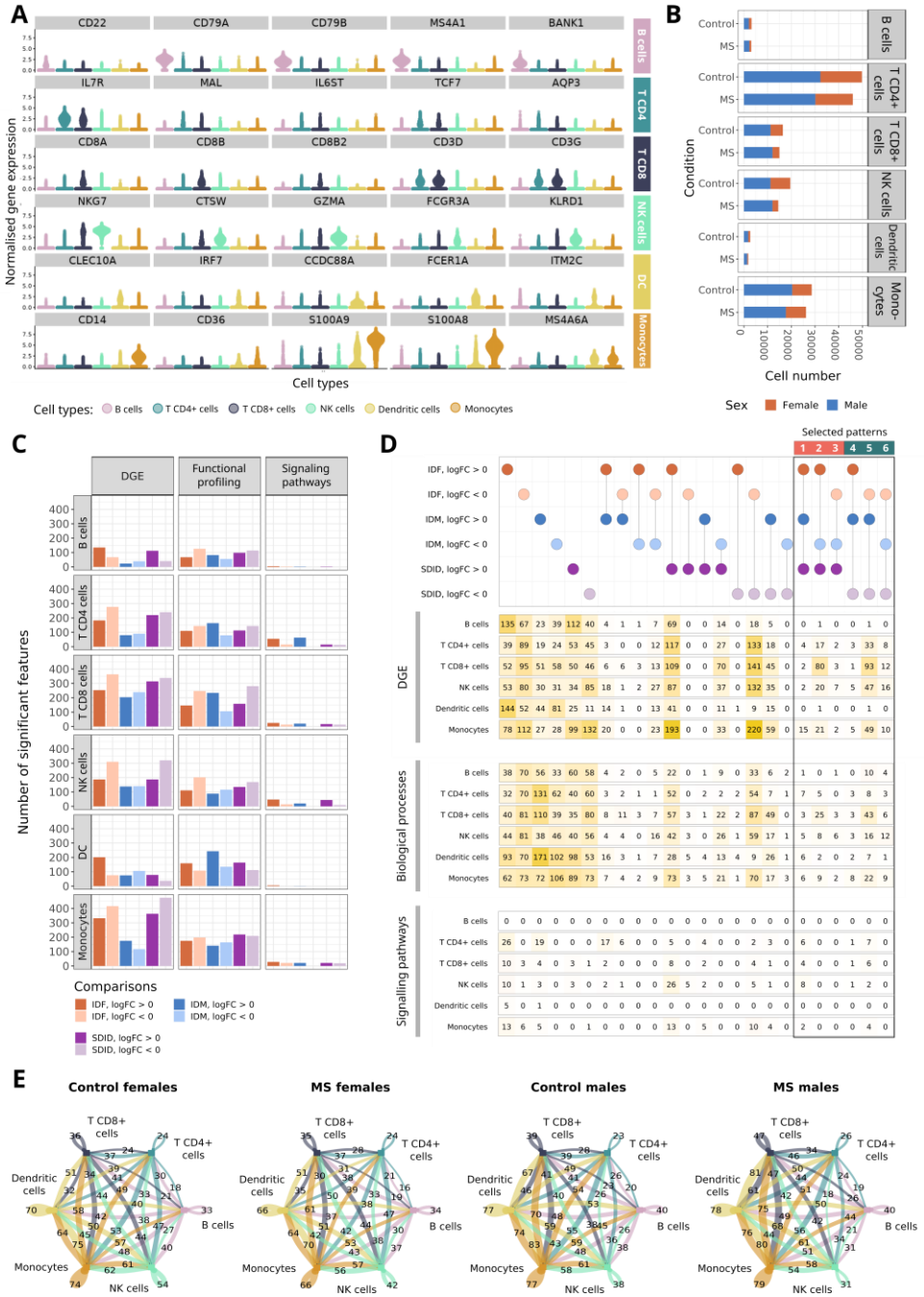
Supplementary Figure 3.S22. (A) B cells and (B) dendritic cells sex-differential atlas in relapsing-remitting multiple sclerosis. Volcano plots of differential gene expression results. Each point represents a gene distributed by its magnitude of change (logFC, X-axis) and its significance (-log10(adjusted p-value), X-axis) for each comparison. IDF: impact of disease in females (MS females vs control females); IDM: impact of disease in males (MS males vs control males); SDID: sex differential impact of disease ((MS females vs control females) - (MS males vs control males)).


A
CD4+ T cells
CD8+ T cells
Monocytes
MAPK signaling pathway
Rap1 signaling pathway
Calcium signaling pathway
cGMP-PKG signaling pathway
Chemokine signaling pathway
HIF-1 signaling pathway
Sphingolipid signaling pathway
PI3K-Akt signaling pathway
AMPK signaling pathway
Apoptosis
Longevity regulating pathway - mammal
Cellular senescence
Focal adhesion
Adherens junction
Tight junction
Complement and coagulation cascades
Antigen processing and presentation
Toll-like receptor signaling pathway
Natural killer cell mediated cytotoxicity
Fc gamma R-mediated phagocytosis
Leukocyte transendothelial migration
Regulation of actin cytoskeleton
Epstein-Barr virus infection
Increased effector activation in females males
B
Relative information flow
Information flow
Interaction strength (i.e. communication probability) in control females MS females
C
Relative information flow
Information flow
Interaction strength (i.e. communication probability) in control males MS males

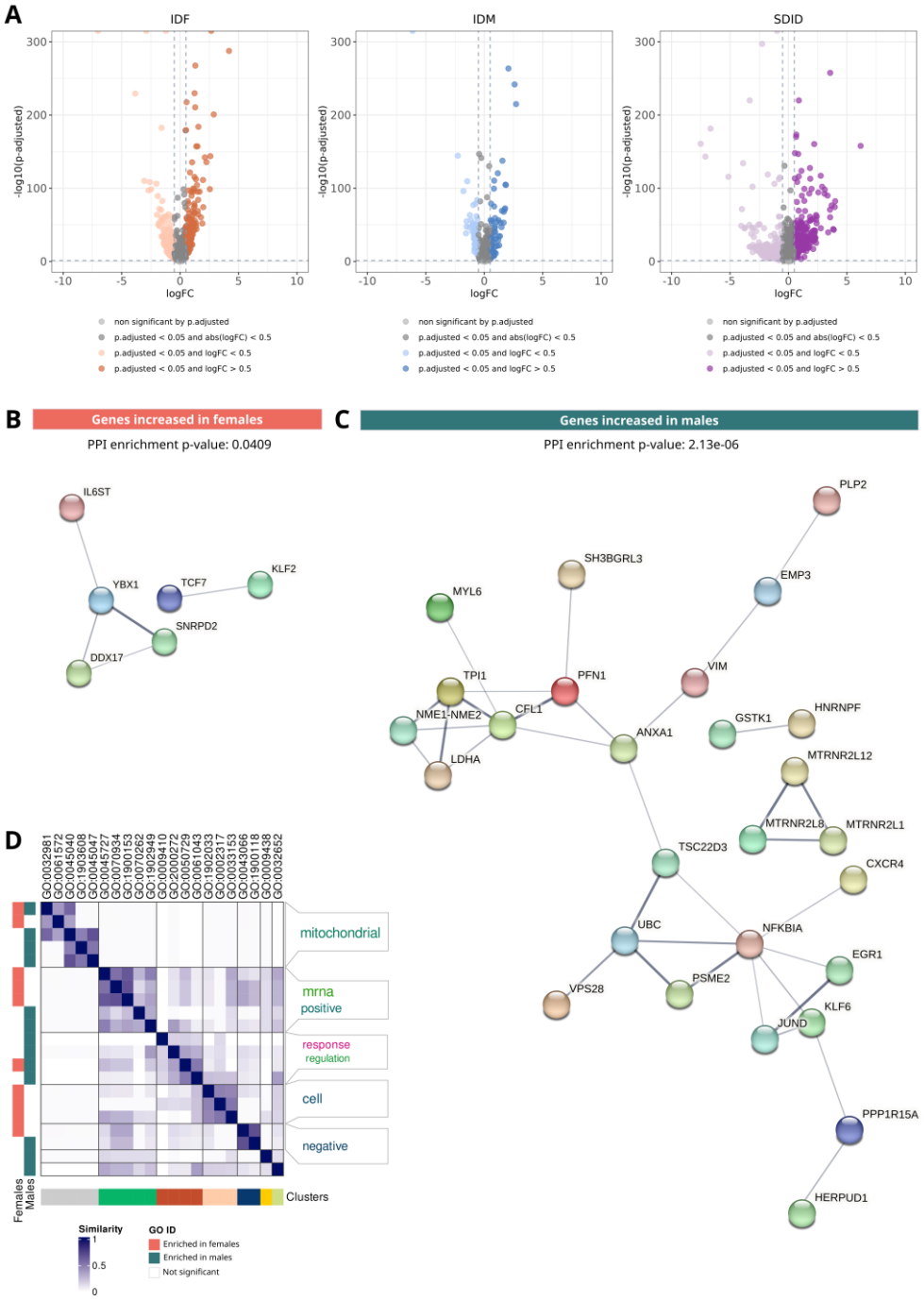
Supplementary Figure 3.S23. **Detailed sex-differential significant changes in the activation of signaling pathways and interaction strengths in cell-cell communication for relapsing-remitting multiple sclerosis.** (*Previous page*) (A) Effectors with increased activation in females (orange boxes) and males (green boxes). Each column represents a different cell type, and each row corresponds to a signaling pathway. Within each pathway, significant effectors for at least one cell type are highlighted. The numeric value at the end of each row denotes the count of subpathways within the pathway that lack significance in any cell type. (B-C) Relative (right) and absolute (left) information flow of pathways in females (B) and males (C) when comparing multiple sclerosis to control samples. Cell types evaluated: CD4+ T cells, CD8+ T cells, monocytes and NK cells. Information flow was calculated as the sum of communication probability (i.e. interaction strengths) among all ligand-receptor pairs in each group. A paired Wilcoxon test was implemented to determine significant differences, which are highlighted by coloring their names. *MS: multiple sclerosis.*



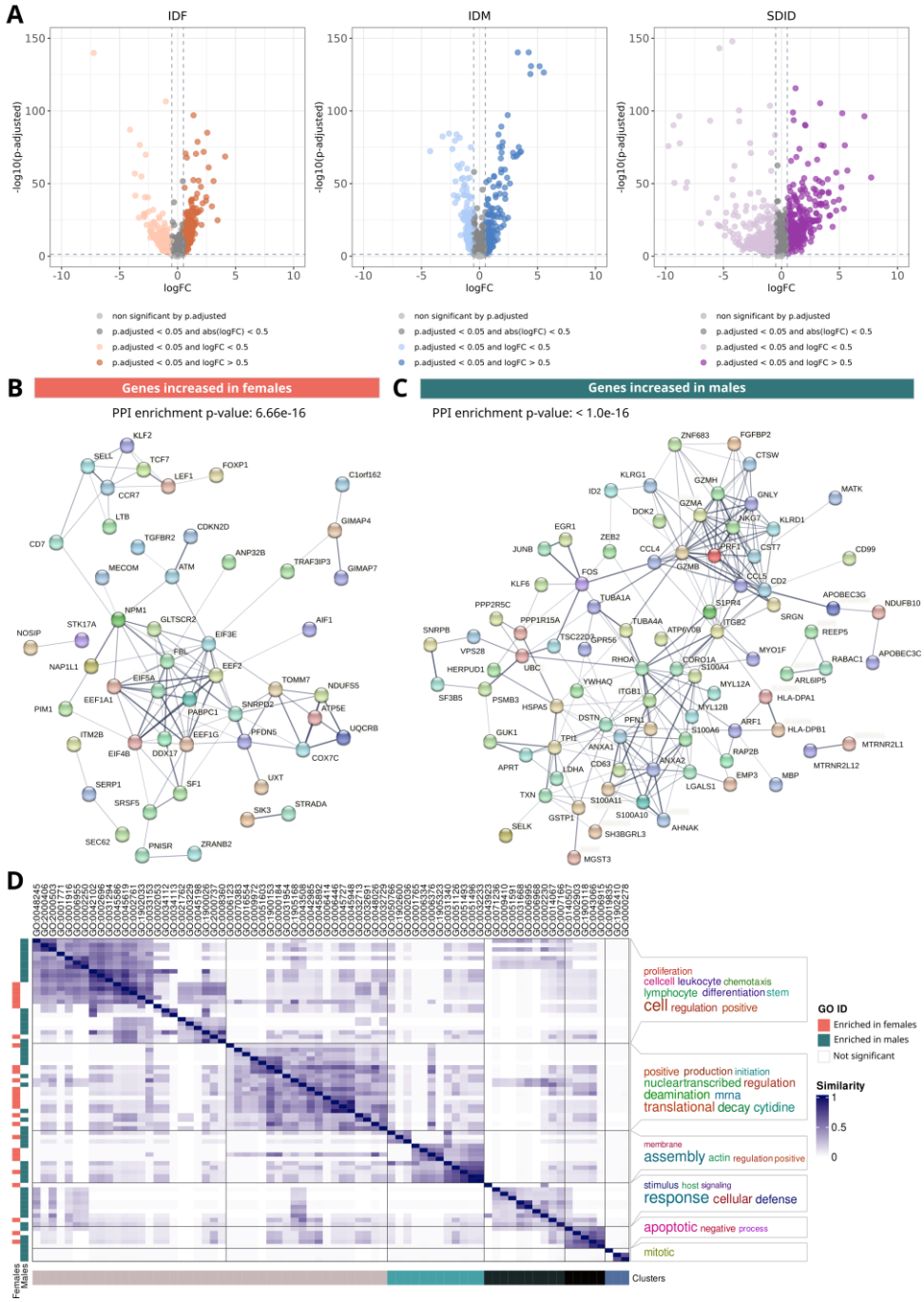
Supplementary Figure 3.S24. Upset plot for genes more expressed in males than females by cell type in relapsing-remitting multiple sclerosis. Horizontal bars represent the number of significant genes in each cell type. Dots indicate the combinations of intersections tested, with the top vertical bars denoting the number of significant genes of the corresponding intersection. *NK: natural killer.*



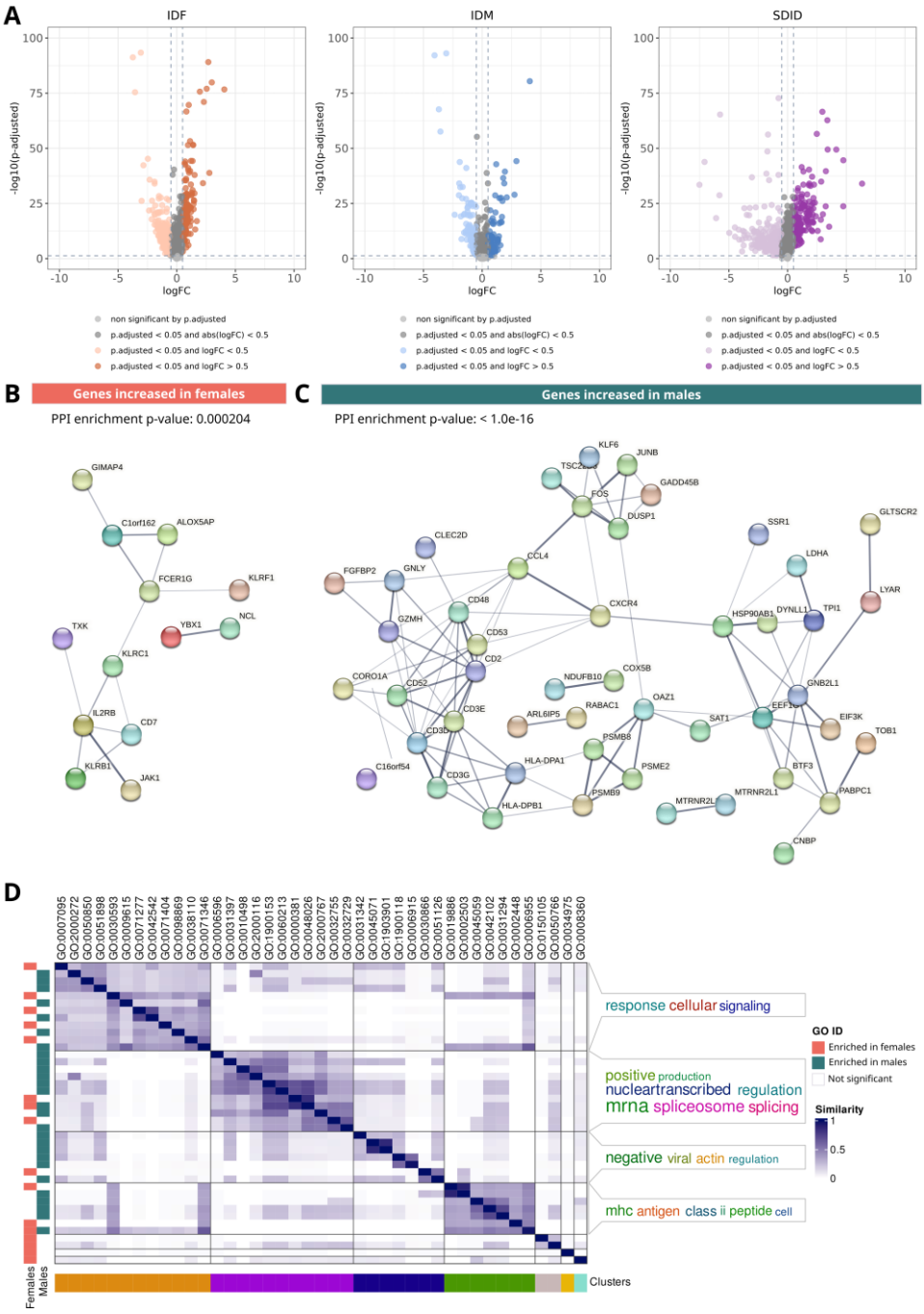
Supplementary Figure 3.S25. Summary of significant sex differences in primary progressive multiple sclerosis. (A) Expression pattern of marker genes. (B) Number of cells identified for each cell type based on condition and sex. *MS: multiple sclerosis.* (C) Number of significant features by cell type, analysis and comparison tested. (D) Upset map of significant features for each cell type separated by comparison and direction of change (logFC). (E) Number of significant cell-cell interactions by group. Color indicates the cell type providing the ligand protein. The thickness of the interaction corresponds to the magnitude of interactions.



Supplementary Figure 3.S26. CD4+ T cells sex-differential atlas in primary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.

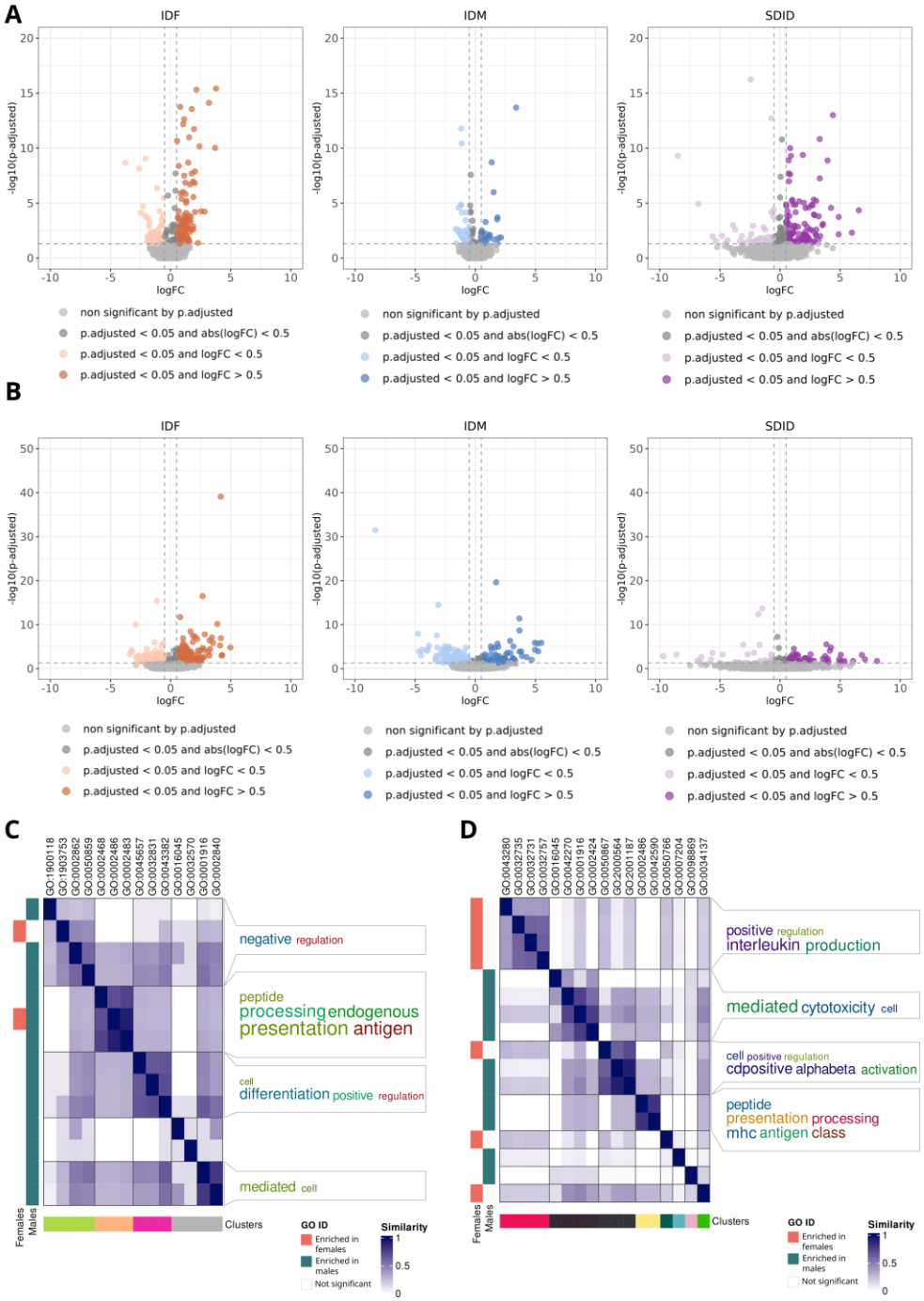


Supplementary Figure 3.S27. CD8+ T cells sex-differential atlas in primary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.



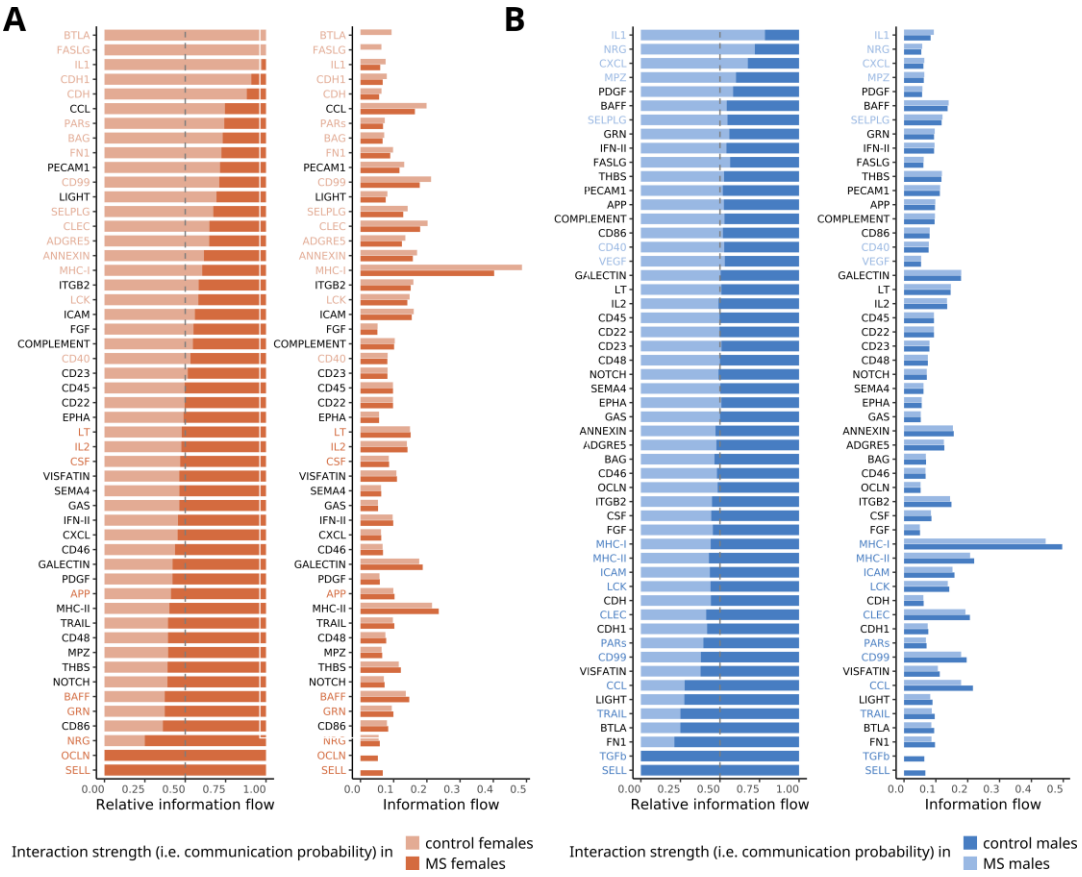
Supplementary Figure 3.S28. NK cells sex-differential atlas in primary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.

Supplementary Figure 3.S29. Monocytes sex-differential atlas in primary progressive multiple sclerosis. (A) Volcano plots of differential gene expression results. Protein-protein interaction network of genes increased in females (B, orange) and males (C, green) from those significant in all three comparisons. Edge thickness is directly proportional to the structural and functional confidence of the interaction. *PPI: protein-protein interaction.* (D) Degree of similarity between sex-differential enriched functions. Each row and column correspond with a significant function. Blue intensity indicates the degree of similarity.

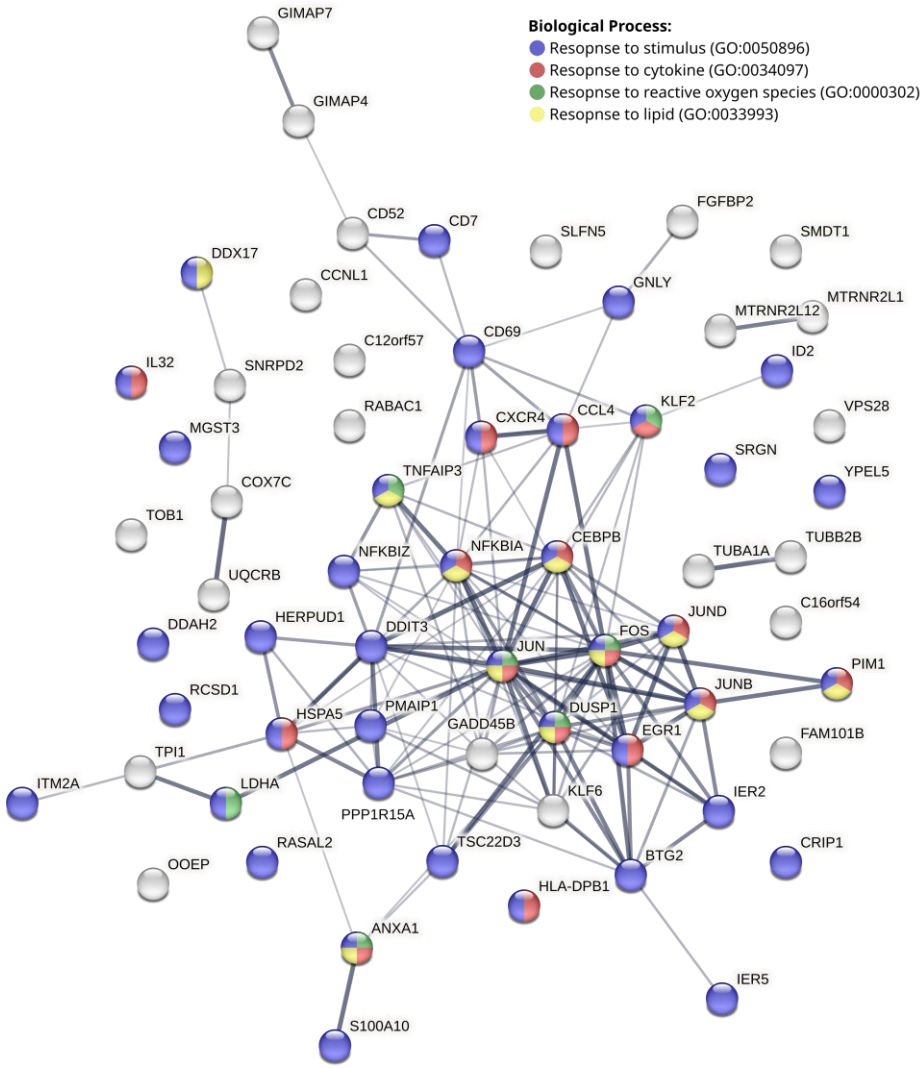


Supplementary Figure 3.S30. (A, C) B cells and (B, D) dendritic cells sex-differential atlas in primary progressive multiple sclerosis. (A,B) Volcano plots of differential gene expression results. (C,D) Degree of similarity between sex-differential enriched functions in females (left horizontal orange bars) and males (left horizontal green bars). Each row and column corresponds with a significant function. Blue intensity indicates the degree of similarity. Clusters are shown at the bottom, with the associated word cloud on the right of the plot.

Supplementary Figure 3.S31. Detailed sex-differential significant changes in the activation of signaling pathways for primary progressive multiple sclerosis. Effectors with increased activation in females (orange boxes) and males (green boxes). Each column represents a different cell type, and each row corresponds to a signaling pathway. Within each pathway, significant effectors for at least one cell type are highlighted. The numeric value at the end of each row denotes the count of subpathways within the pathway that lack significance in any cell type. *NK: natural killer.* Triangle: Nicotinic acid adenine dinucleotide phosphate. Star: D-myo-Inositol 1,4,5-triphosphate.



Supplementary Figure 3.S32. Detailed sex-differential significant changes in the cell-cell communication interaction strengths for primary progressive multiple sclerosis. Relative (right) and absolute (left) information flow of pathways in females (A) and males (B) when comparing multiple sclerosis to control samples. Cell types evaluated: CD4+ T cells, CD8+ T cells, monocytes and NK cells. Information flow was calculated as the sum of communication probability (i.e. interaction strengths) among all ligand-receptor pairs in each group. A paired Wilcoxon test was implemented to determine significant differences, which are highlighted by coloring their names. *MS: multiple sclerosis; NK: natural killer.*



Supplementary Figure 3.S33. Protein-protein interaction network of the genes that provide clustering of RRMS and PPMS immune system cell types according to their differential sex profile. Network showing the connection of CD4+ T cells, CD8+ T cells, NK cells and monocytes in RRMS and PPMS subtypes from genes with significant expression by sex in at least three of the eight genes evaluated. Edge thickness indicates the structural and functional confidence of the interaction. Color established based on the functions the genes are involved in. *GO: gene ontology; NK: natural killer; PPMS: primary progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis*.

Package	Version	Package	Version
SingleCellExperiment	1.16.0	dplyr	1.0.8
scuttle	1.4.0	Matrix	1.5.1
scdblfinder	1.8.0	BiocParallel	1.28.3
scran	1.22.1	pheatmap	1.0.12
scater	1.22.0	reshape2	1.4.4
PCAtools	2.6.0	ggpubr	0.4.0
igraph	1.2.11	simplifyEnrichment	1.4.0
bluster	1.4.0	gridExtra	2.3.0
BRETIGEA	1.0.3	ComplexHeatmap	2.13.4
singler	1.8.1	circlize	0.4.14
celldex	1.4.0	CellChat	1.6.1
MAST	1.20.0	rrvgo	1.6.0
STRINGdb	2.6.5	shiny	1.9.1
org.Hs.eg.db	3.14.0	shinythemes	1.2.0
hipathia	2.10.0	rsconnect	1.5.0
ggplot2	3.4.0		

Supplementary Table 3.S1. List of R packages and versions implemented in the bioinformatic workflow.

Sample type	KEGG ID pathways
Nervous tissue	hsa03320, hsa04010, hsa04012, hsa04014, hsa04015, hsa04020, hsa04022, hsa04024, hsa04062, hsa04064, hsa04066, hsa04068, hsa04071, hsa04072, hsa04110, hsa04115, hsa04150, hsa04151, hsa04152, hsa04210, hsa04211, hsa04218, hsa04310, hsa04330, hsa04340, hsa04350, hsa04360, hsa04370, hsa04390, hsa04510, hsa04520, hsa04530, hsa04540, hsa04612, hsa04620, hsa04621, hsa04622, hsa04623, hsa04630, hsa04668, hsa04710, hsa04713, hsa04720, hsa04722, hsa04723, hsa04724, hsa04725, hsa04726, hsa04727, hsa04728, hsa04730, hsa04740, hsa04742, hsa04750, hsa04810, hsa04915, hsa04917, hsa04921, hsa04922, hsa05010, hsa05012, hsa05014, hsa05016, hsa05020, hsa05030, hsa05031, hsa05032, hsa05169
Blood	hsa03320, hsa04010, hsa04012, hsa04014, hsa04015, hsa04020, hsa04022, hsa04024, hsa04062, hsa04064, hsa04066, hsa04068, hsa04071, hsa04072, hsa04115, hsa04150, hsa04151, hsa04152, hsa04210, hsa04211, hsa04218, hsa04310, hsa04330, hsa04340, hsa04350, hsa04370, hsa04390, hsa04510, hsa04520, hsa04530, hsa04540, hsa04610, hsa04612, hsa04620, hsa04621, hsa04622, hsa04623, hsa04630, hsa04650, hsa04660, hsa04662, hsa04666, hsa04668, hsa04670, hsa04710, hsa04713, hsa04750, hsa04810, hsa04915, hsa04917, hsa05169

Supplementary Table 3.S2. Signaling pathways analyzed based on the type of sample examined: nervous tissue (SPMS-CNS dataset) and blood (RRMS-PBMCs and PPMS-PBMCs datasets). KEGG: Kyoto Encyclopedia of Genes and Genomes; *ID: identifier.*

	Cohort 1	Cohort 2	Cohort 3
MS subtype	RRMS	RRMS	PPMS
Control samples	healthy individuals	None	healthy individuals
Technology	10x Genomics	10x Genomics	10x Genomics
N MS females	5 paired samples (5 -nat, 5 +nat)	13	3
N control females	5	13	6
N MS males	4 paired samples (4 -nat, 5 +nat)	0	3
N control males	5	0	6

Supplementary Table 3.S3. Characteristics of GSE144744 cohorts. * For one MS male sample in GSE144744c1 the untreated time point was not available. N: number of samples; -nat: not treated with natalizumab; +nat: treated with natalizumab. *MS: multiple sclerosis; RRMS relapsing-remitting MS; PPMS: primary progressive MS.*

Parent ID	Description parent	GO ID	Description GO	Parent ID	Description parent	GO ID	Description GO
GO:0006491	N-glycan processing	GO:0006491	N-glycan processing	GO:0006511	ubiquitin-dependent protein catabolic process	GO:0006511	ubiquitin-dependent protein catabolic process
GO:1902430	negative regulation of amyloid-beta formation	GO:1902430	negative regulation of amyloid-beta formation			GO:0018105	peptidyl-serine phosphorylation
		GO:0034205	amyloid-beta formation			GO:0000209	protein polyubiquitination
GO:0045892	negative regulation of transcription, DNA-templated	GO:0045892	negative regulation of transcription, DNA-templated			GO:0046777	protein autophosphorylation
		GO:0000381	regulation of alternative mRNA splicing, via spliceosome			GO:0035335	peptidyl-tyrosine dephosphorylation
		GO:0048024	regulation of mRNA splicing, via spliceosome			GO:0018107	peptidyl-threonine phosphorylation
						GO:0070936	protein K48-linked ubiquitination
						GO:0018401	peptidyl-proline hydroxylation to 4-hydroxy-L-proline
						GO:0033135	regulation of peptidyl-serine phosphorylation

Supplementary Table 3.S4. General classification of female enriched functions from cluster 4 of the astrocyte-microglia-neuron triad in secondary progressive multiple sclerosis (main manuscript Fig. 2A). Classification performed with the *rrvgo* R package setting medium threshold (value=0.7). *GO: Gene Ontology*.

Parent ID	Description parent	GO ID	Description GO	Parent ID	Description parent	GO ID	Description GO
GO:0043968	histone H2A acetylation	GO:0043968	histone H2A acetylation	GO:0006198	cAMP catabolic process	GO:0006198	cAMP catabolic process
		GO:0000338	protein deneddylation			GO:0019673	GDP-mannose metabolic process
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	GO:0006120	mitochondrial electron transport, NADH to ubiquinone	GO:0002181	cytoplasmic translation	GO:0002181	cytoplasmic translation
		GO:0006096	glycolytic process			GO:0045727	positive regulation of translation
		GO:0006123	mitochondrial electron transport, cytochrome c to oxygen			GO:0006446	regulation of translational initiation
		GO:0006119	oxidative phosphorylation			GO:1990440	positive regulation of transcription from RNA polymerase II promoter in response to endoplasmic reticulum stress
		GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	GO:0006511	ubiquitin-dependent protein catabolic process	GO:0006511	ubiquitin-dependent protein catabolic process
GO:0043154	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	GO:0043154	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process			GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process
		GO:0031397	negative regulation of protein ubiquitination			GO:0030433	ubiquitin-dependent ERAD pathway
		GO:0032148	activation of protein kinase B activity			GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
		GO:0045737	positive regulation of cyclin-dependent protein			GO:0016241	regulation of macroautophagy

			serine/threonine kinase activity				
		GO:0031954	positive regulation of protein autophosphorylation			GO:0032435	negative regulation of proteasomal ubiquitin-dependent protein catabolic process
		GO:0031396	regulation of protein ubiquitination			GO:0061136	regulation of proteasomal protein catabolic process
		GO:0032515	negative regulation of phosphoprotein phosphatase activity			GO:0019941	modification-dependent protein catabolic process
		GO:1905907	negative regulation of amyloid fibril formation	GO:0051092	positive regulation of NF-kappaB transcription factor activity	GO:0051092	positive regulation of NF-kappaB transcription factor activity
		GO:1904667	negative regulation of ubiquitin protein ligase activity			GO:0032092	positive regulation of protein binding
		GO:1902949	positive regulation of tau-protein kinase activity			GO:0043392	negative regulation of DNA binding
GO:0010894	negative regulation of steroid biosynthetic process	GO:0010894	negative regulation of steroid biosynthetic process	GO:0008216	spermidine metabolic process	GO:0008216	spermidine metabolic process
		GO:0150172	regulation of phosphatidylcholine metabolic process				

Supplementary Table 3.S5. General classification of male enriched functions from cluster 1 of the astrocyte-microglia-neuron triad in secondary progressive multiple sclerosis (main manuscript Fig. 2B). Classification performed with the *rrvgo* R package setting medium threshold (value=0.7). *GO: Gene Ontology.*

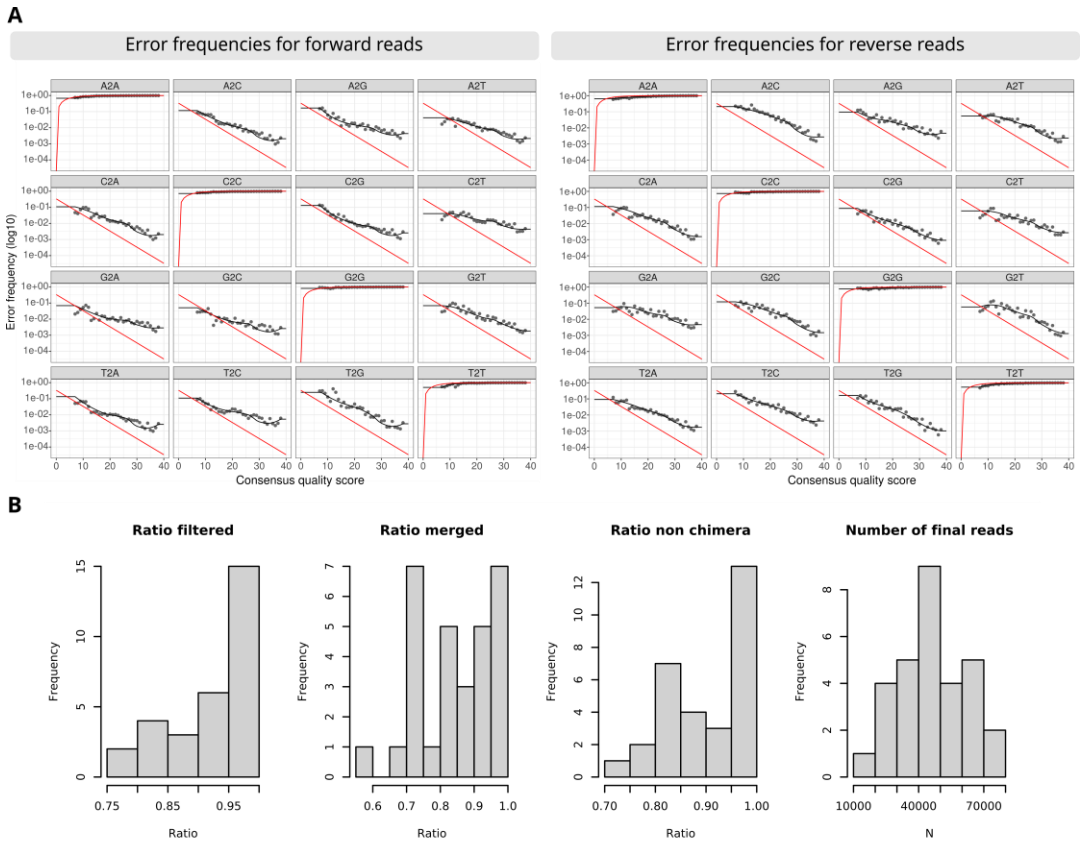
Oligodendrocytes	
Increased in females	AMER2, RASGEF1B, SLC26A3, TMEM144, ANK3, ANO4, CADM2, CLDN11, CTNND2, DYSF, FCHSD2, GPRC5B, IL1RAPL1, L3MBTL4, MALAT1, MARCH1, NCAM2, NLGN1, OLMALINC, PLXDC2, SPOCK1, LINC00685, LPHN3
Increased in males	BCYRN1, CAMK2B, CCK, CHN1, DPYSL5, ENC1, MAP1B, MIR219A2, NRGN, TUBB2B, AATK, BCAS1, CERCAM, MGAT5, MIR325HG, PDE4D, RBFOX1, SLC5A11
OPCs	
Increased in females	HIF3A, GPNMB, KAZN, PARD3, QKI, SMOC1, TNR, VCAN
Increased in males	MAP1B, NRGN, SYT1

Supplementary Table 3.S6. Sex differentially expressed genes in oligodendrocytes and OPCs. Listed genes are significant in all 3 comparisons: IDF: impact of disease in females (MS females vs control females); IDM: impact of disease in males (MS males vs control males); SDID: sex differential impact of disease ((MS females vs control females) - (MS males vs control males)). *OPCs: oligodendrocyte precursor cells*.

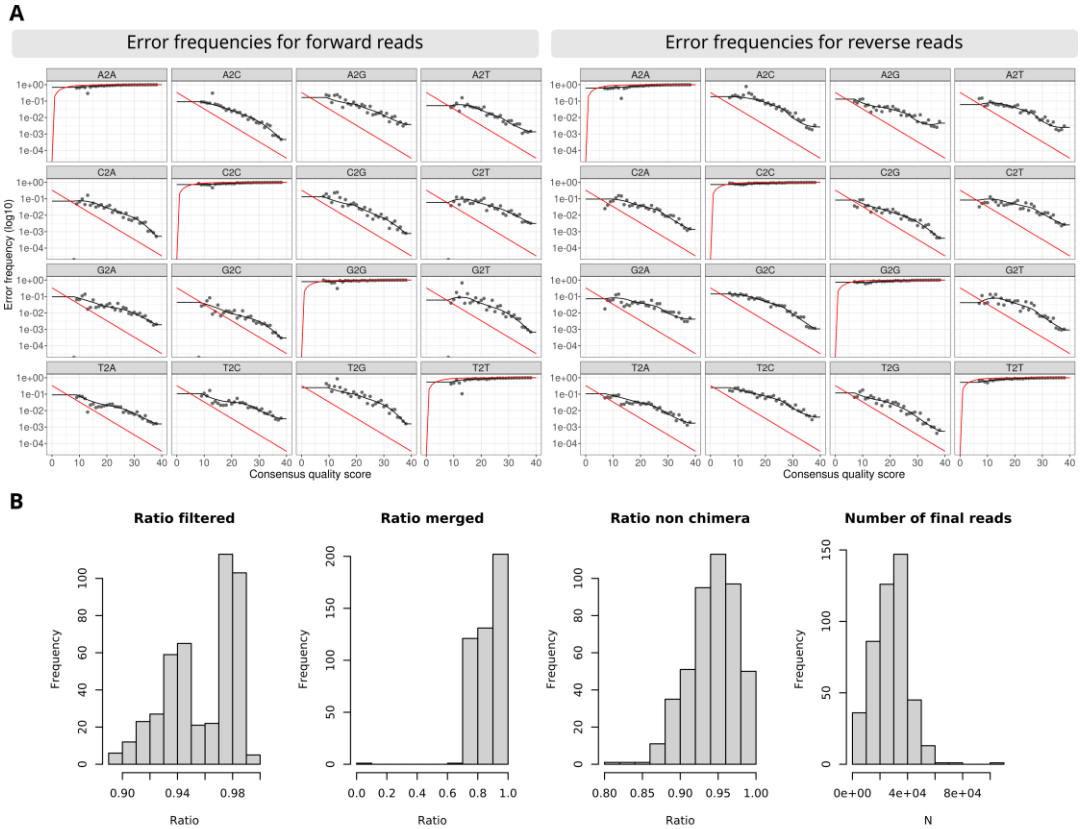
	RRMS		**PPMS**	
	Females	**Males**	**Females**	**Males**
CD4+ T cells				
CD8+ T cells				HLA-DPA1, HLA-DPB1
NK cells				HLA-DPA1, HLA-DPB1
Monocytes	HLA-DQB1, HLA-DQA1	HLA-DRB5		HLA-A, HLA-DRA, HLA-DRB1, HLA-DPB1

Supplementary Table 3.S7. Human leukocyte antigens dysregulated in multiple sclerosis. HLA genes have been filtered out among those significant in all 3 comparisons: IDF: impact of disease in females (MS females vs control females); IDM: impact of disease in males (MS males vs control males); SDID: sex differential impact of disease ((MS females vs control females) - (MS males vs control males)). *NK: natural killer; PPMS: primary progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis.*

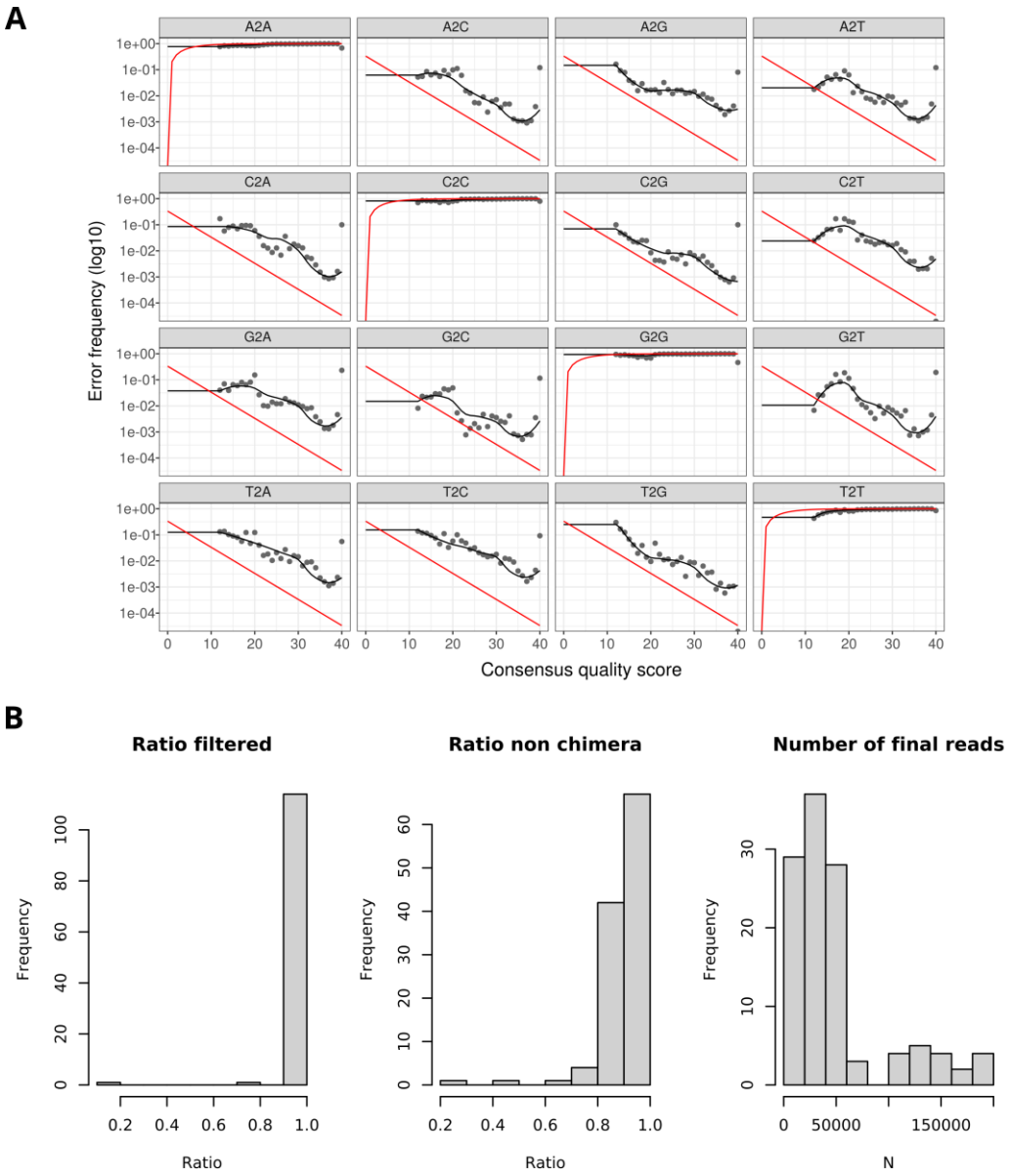
10.2. SUPPLEMENTARY MATERIAL STUDY II



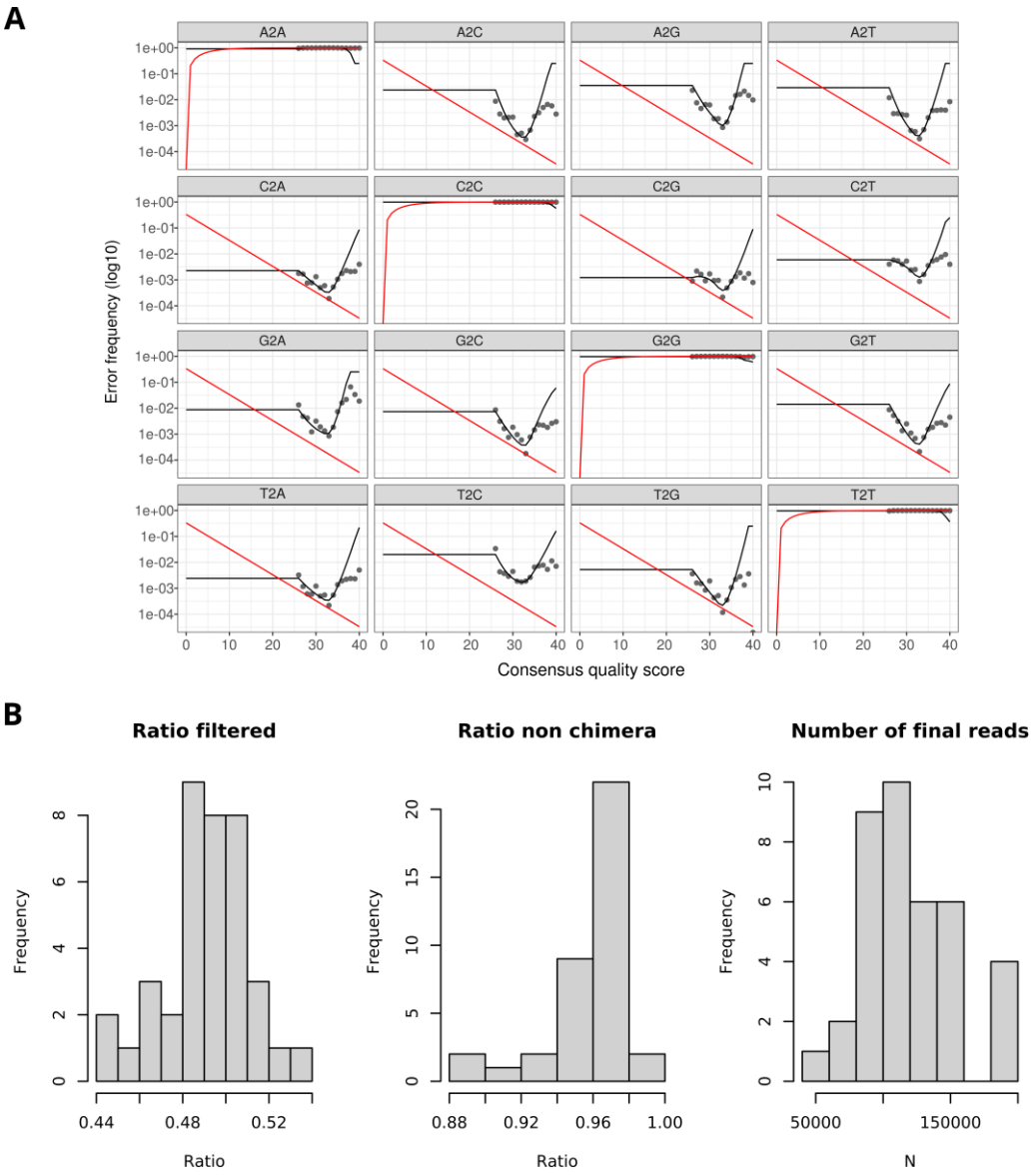
Supplementary Figure 4.S1. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA684124 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



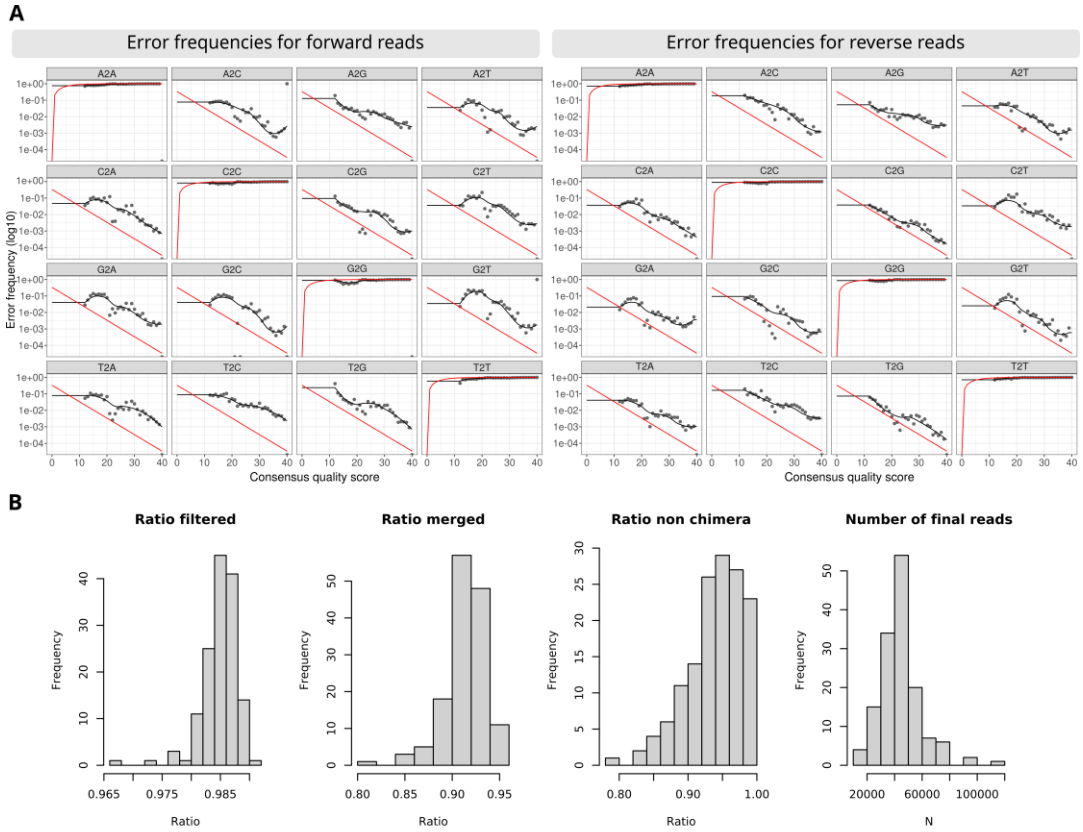
Supplementary Figure 4.S2. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA721421 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



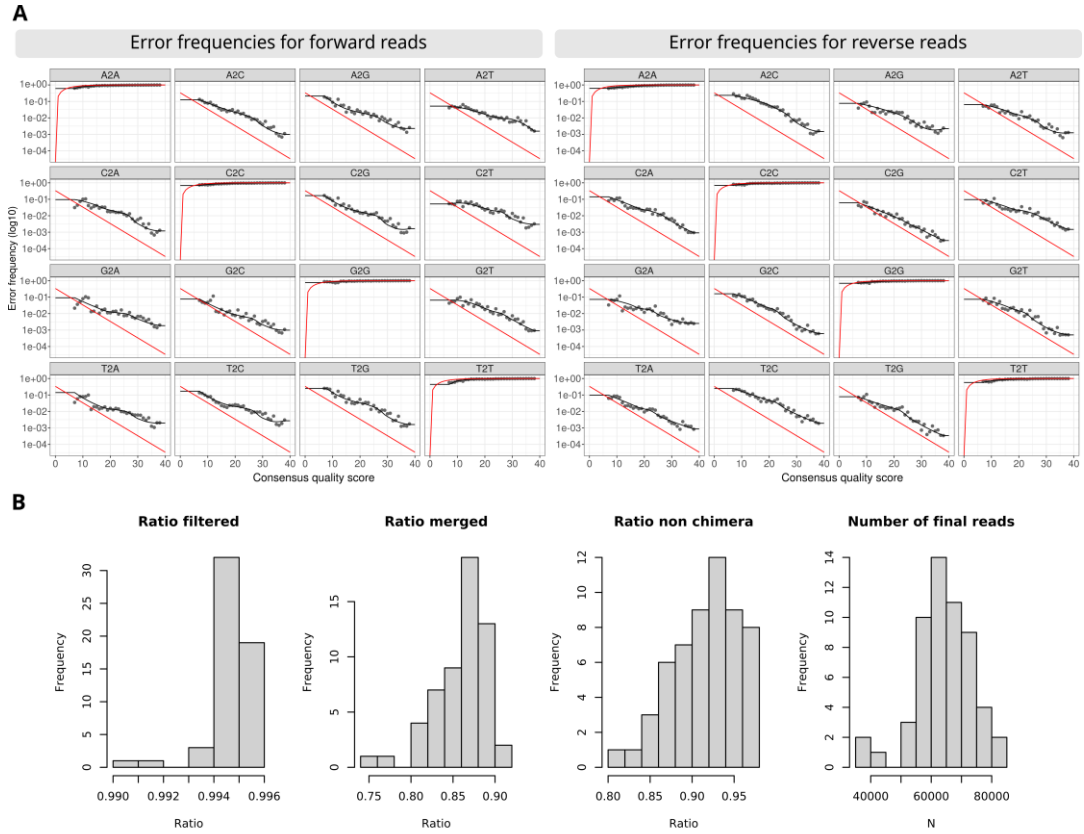
Supplementary Figure 4.S3. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJEB99111 dataset. (A) Error rate models. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



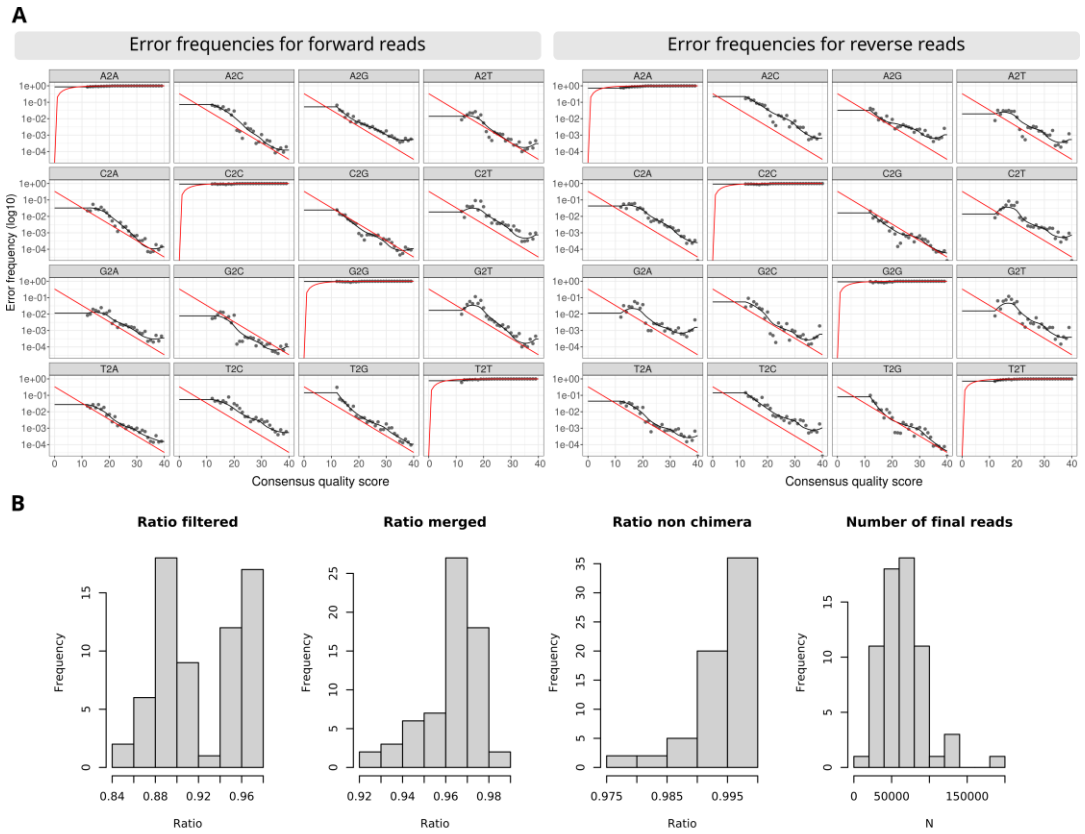
Supplementary Figure 4.S4. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA748865 dataset. (A) Error rate models. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



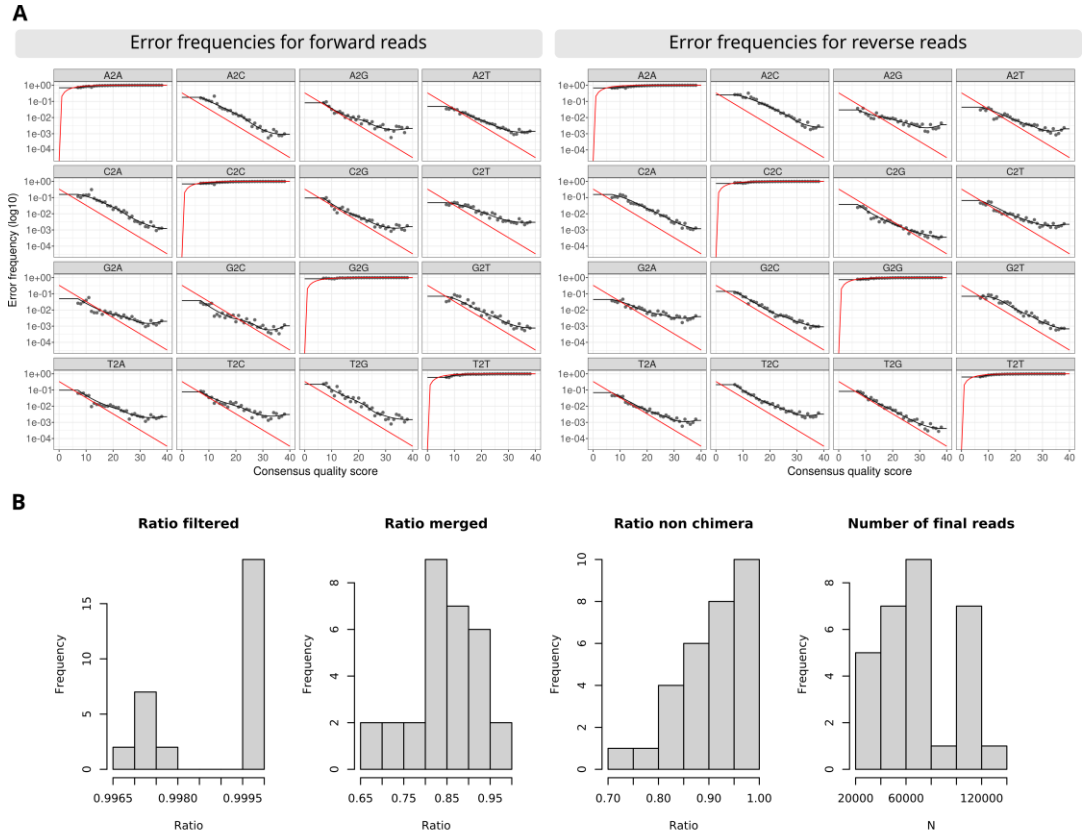
Supplementary Figure 4.S5. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA889427 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



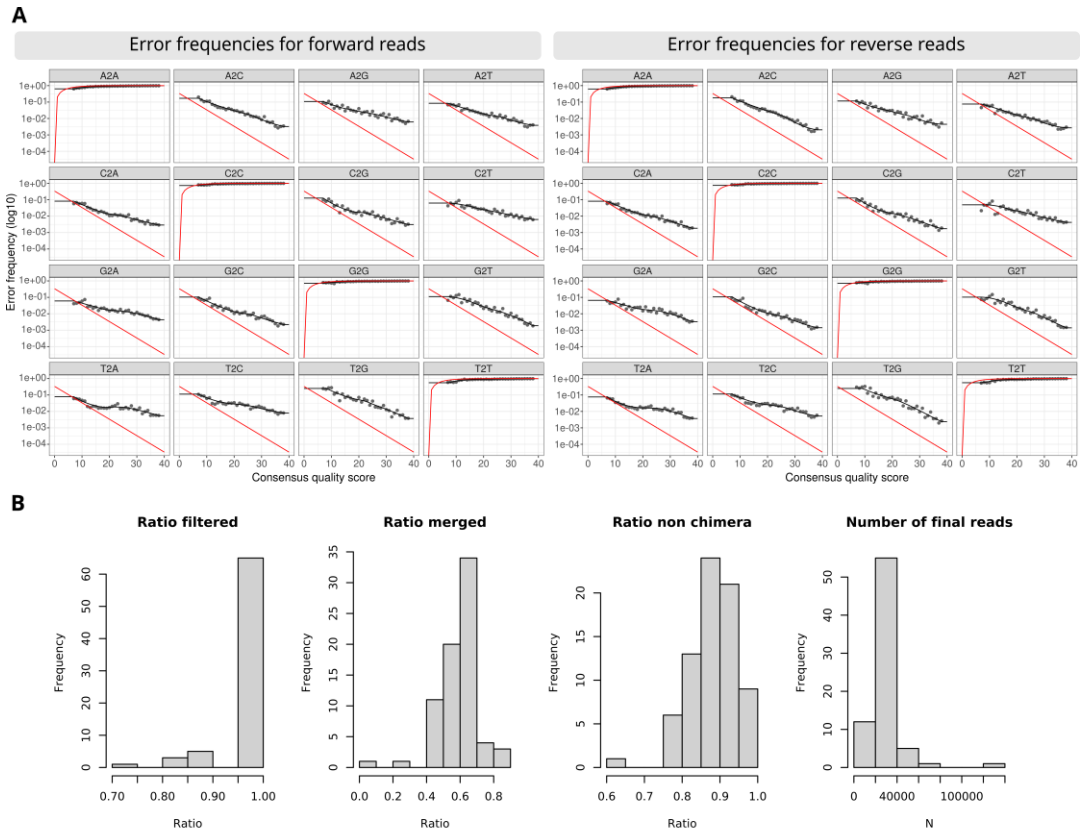
Supplementary Figure 4.S6. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA732670 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



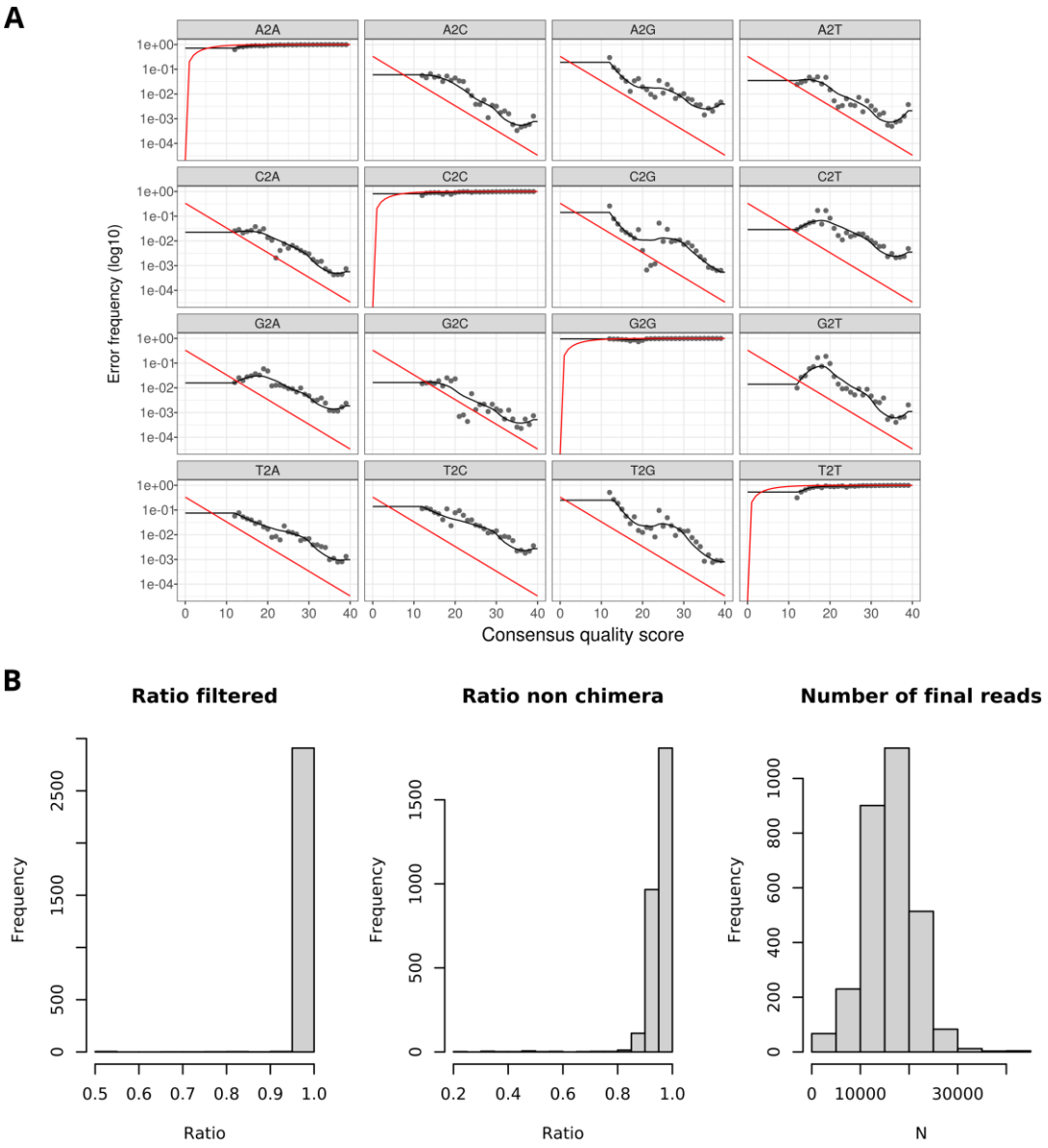
Supplementary Figure 4.S7. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJEB34168 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



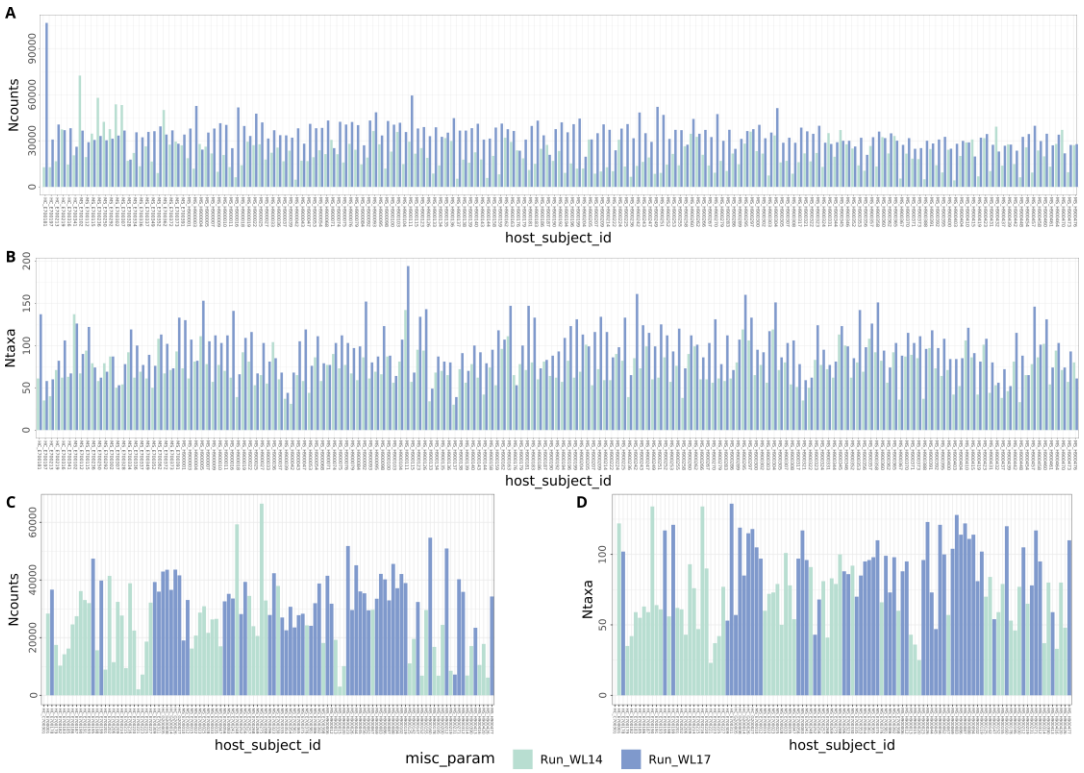
Supplementary Figure 4.S8. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJNA565173 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



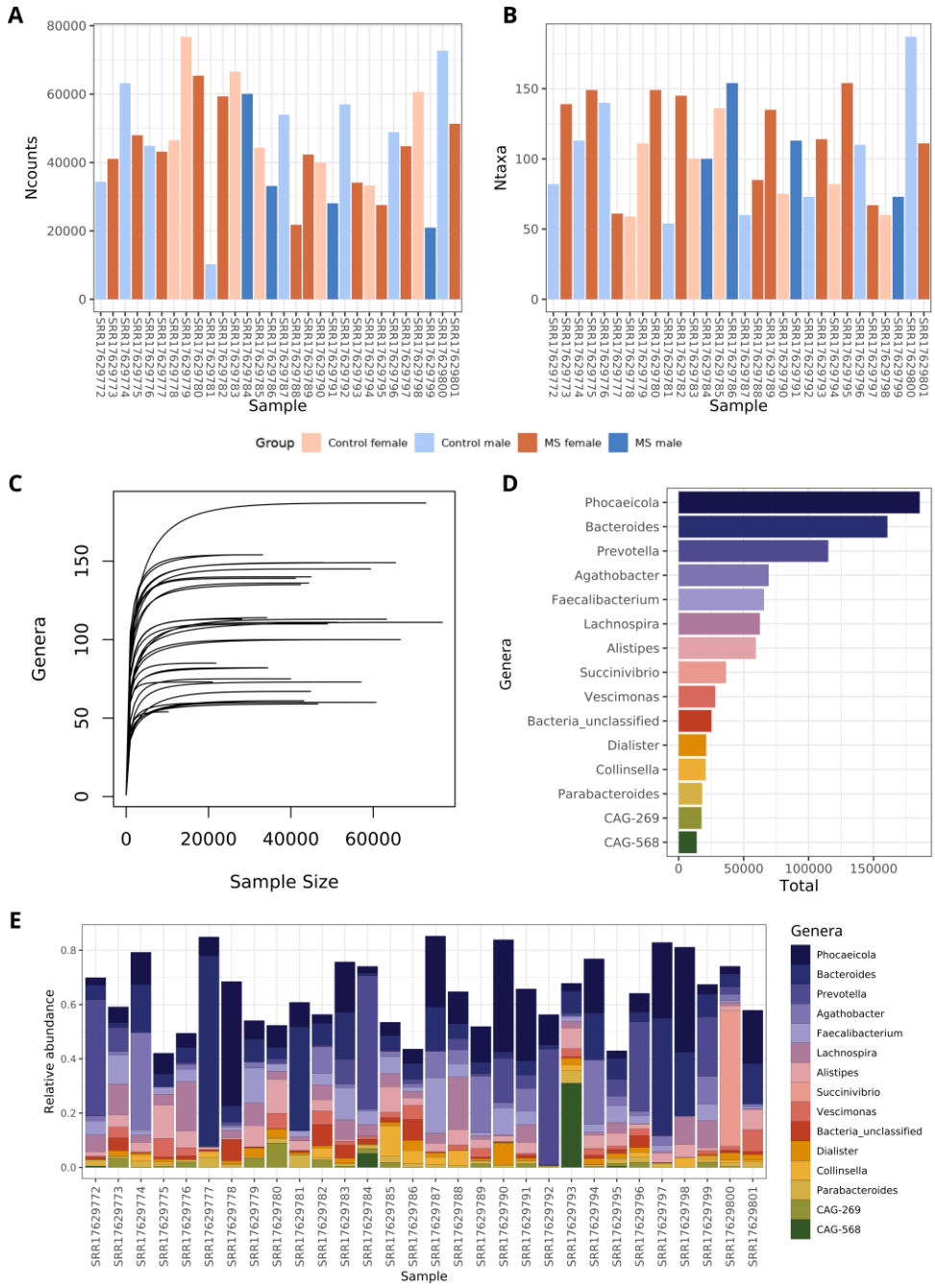
Supplementary Figure 4.S9. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJEB67783 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



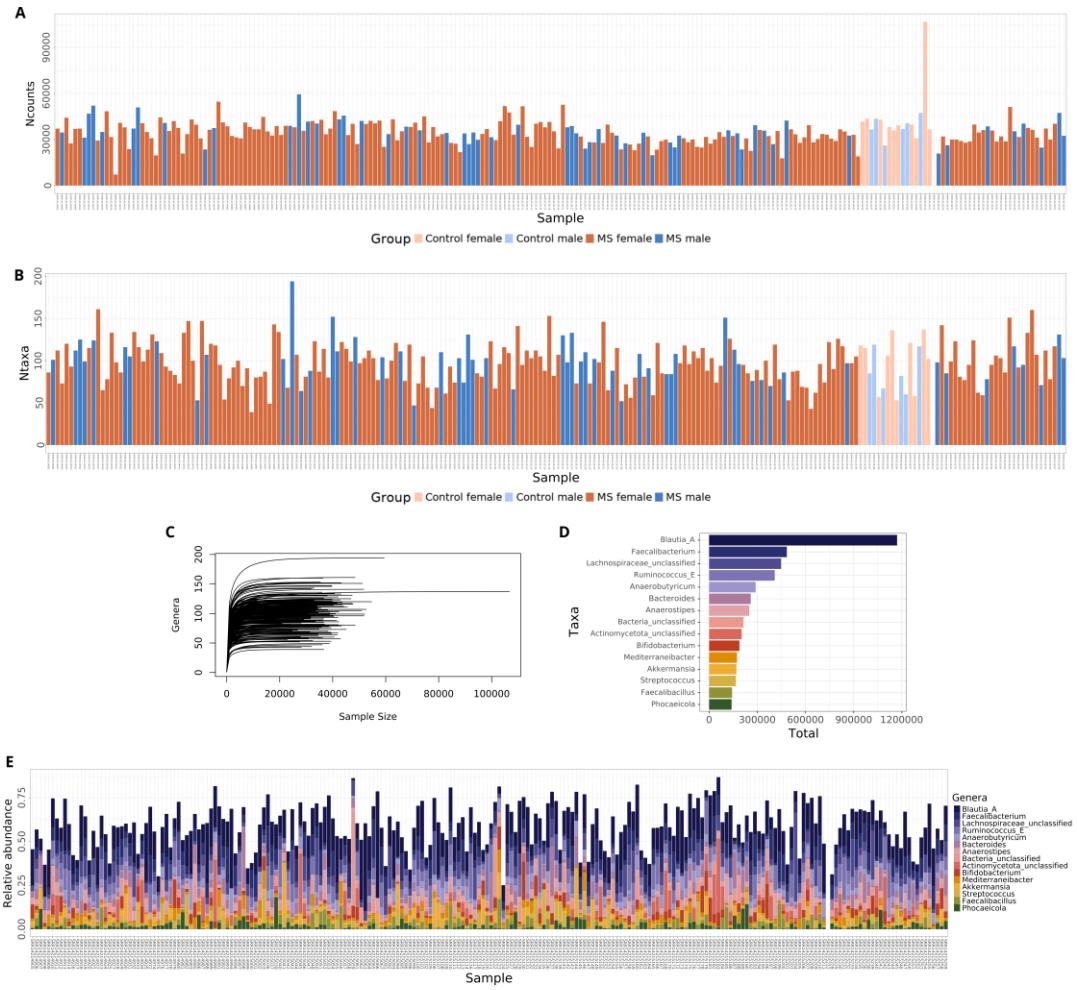
Supplementary Figure 4.S10. Intermediate results of ASVs inference with the DADA2 algorithm for the PRJEB32762 dataset. (A) Error rate models for forward (left) and reverse (right) reads. (B) Distribution of read proportions retained per sample at each step of the DADA2 pipeline: post-filtering, post-merging, post-chimera removal, and final read count. *ASV: amplicon sequence variant.*



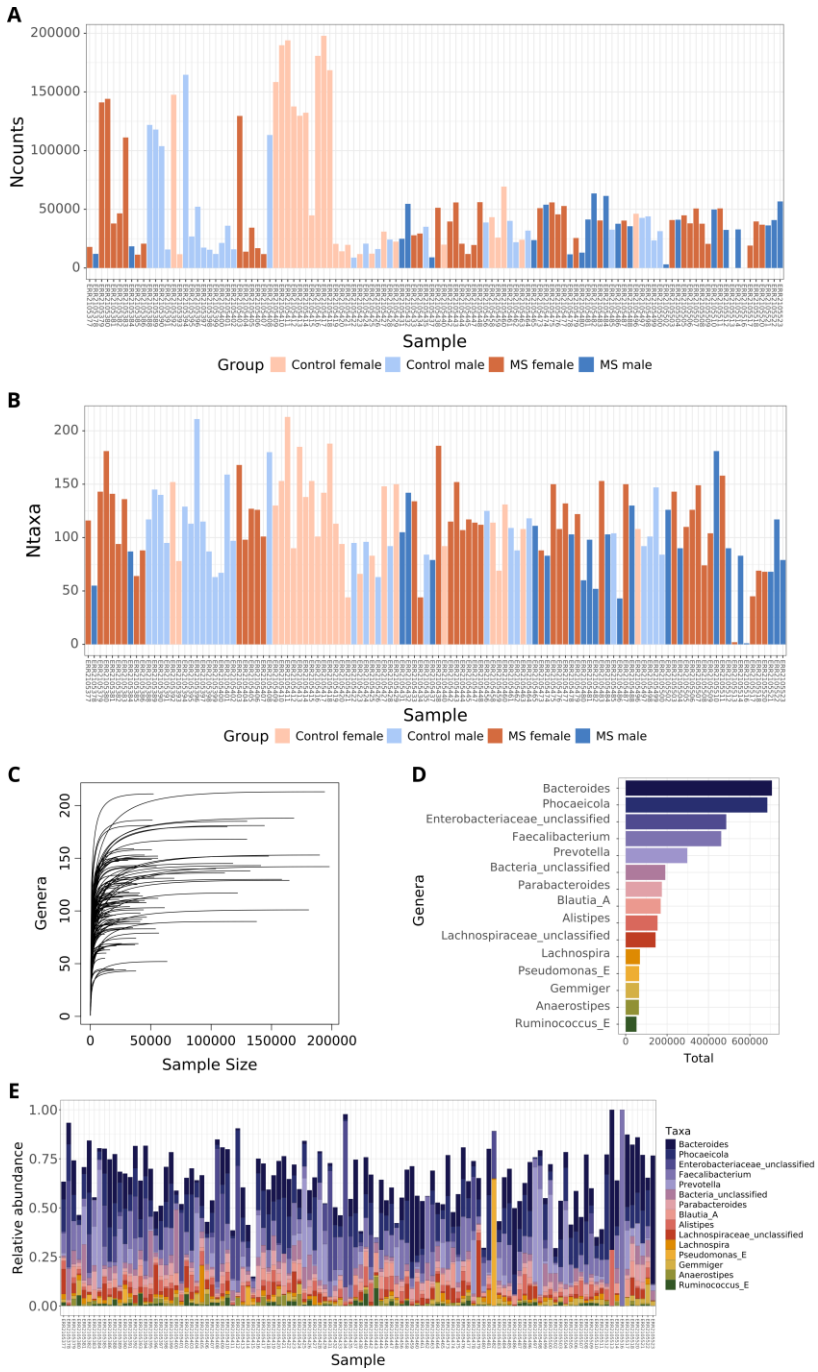
Supplementary Figure 4.S11. Distribution of the number of counts and number of identified taxa for the PRJNA721421 dataset. (A-B) Distribution of the number of counts (A) and identified taxa (B) for samples sequenced in both WL14 and WL17 runs. (C-D) Distribution of the number of counts (C) and identified taxa (D) for samples sequenced only in one run, WL14 or WL17.



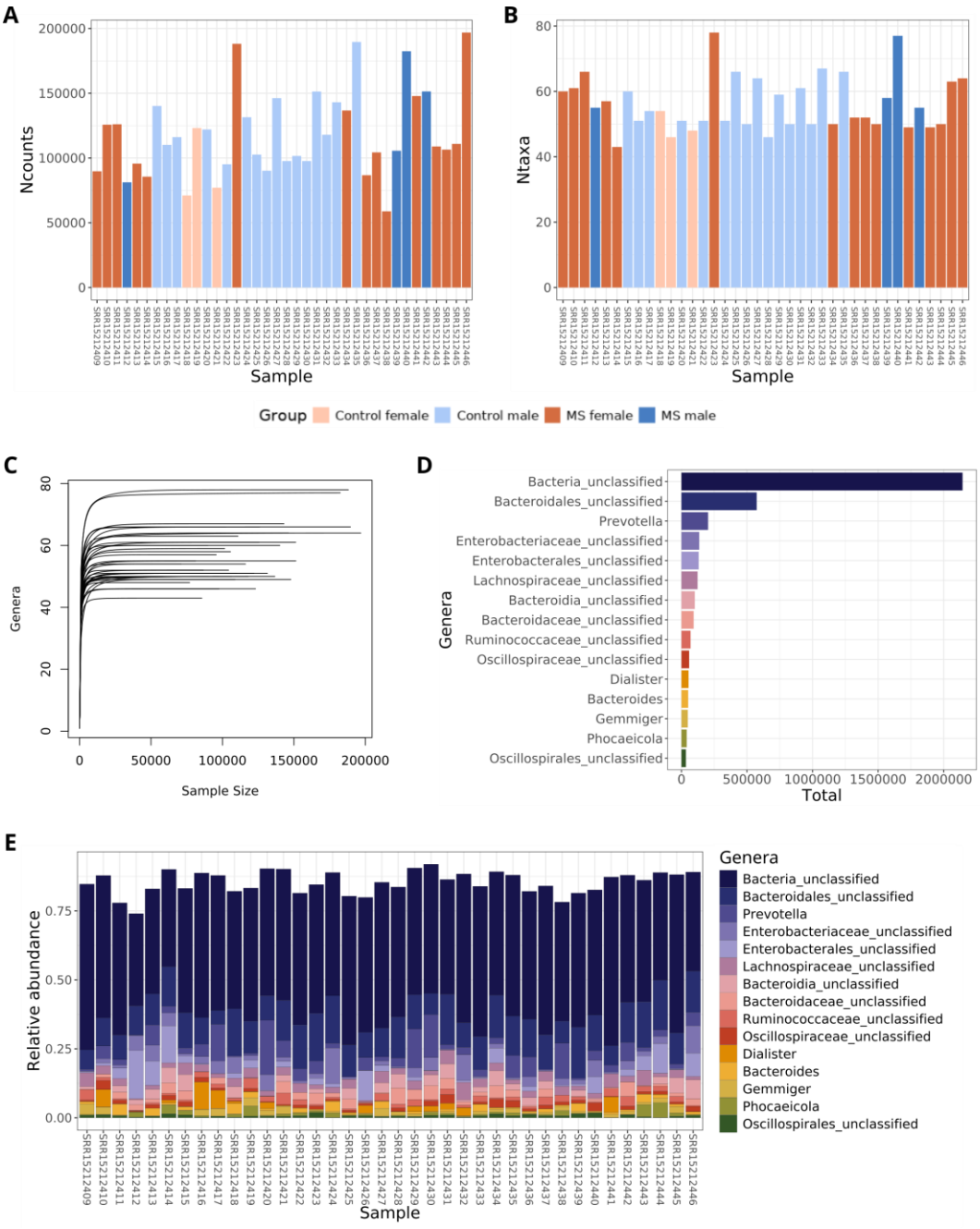
Supplementary Figure 4.S12. Quality control and taxonomic composition metrics for dataset PRJNA684124. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



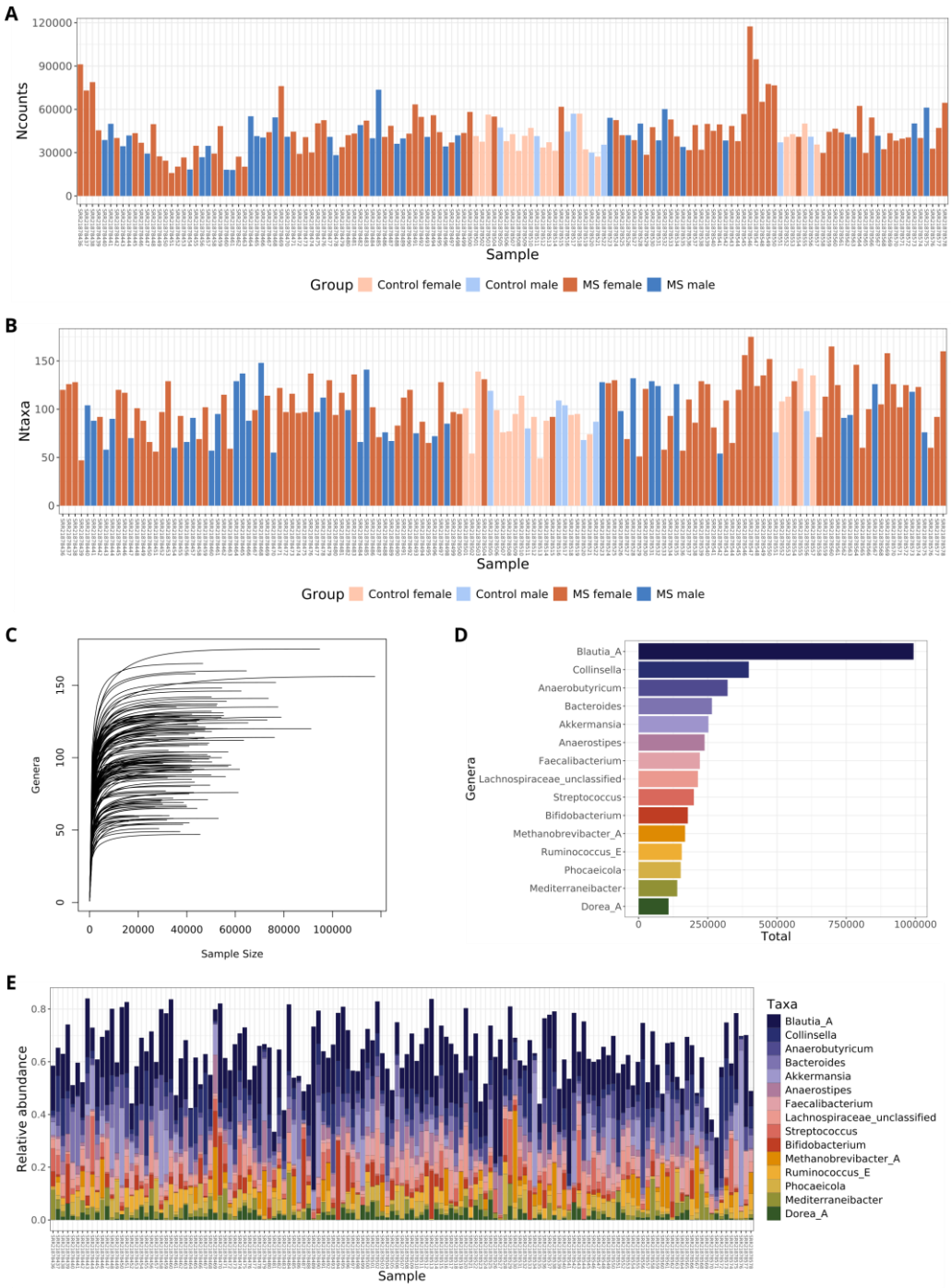
Supplementary Figure 4.S13. Quality control and taxonomic composition metrics for dataset PRJNA721421. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



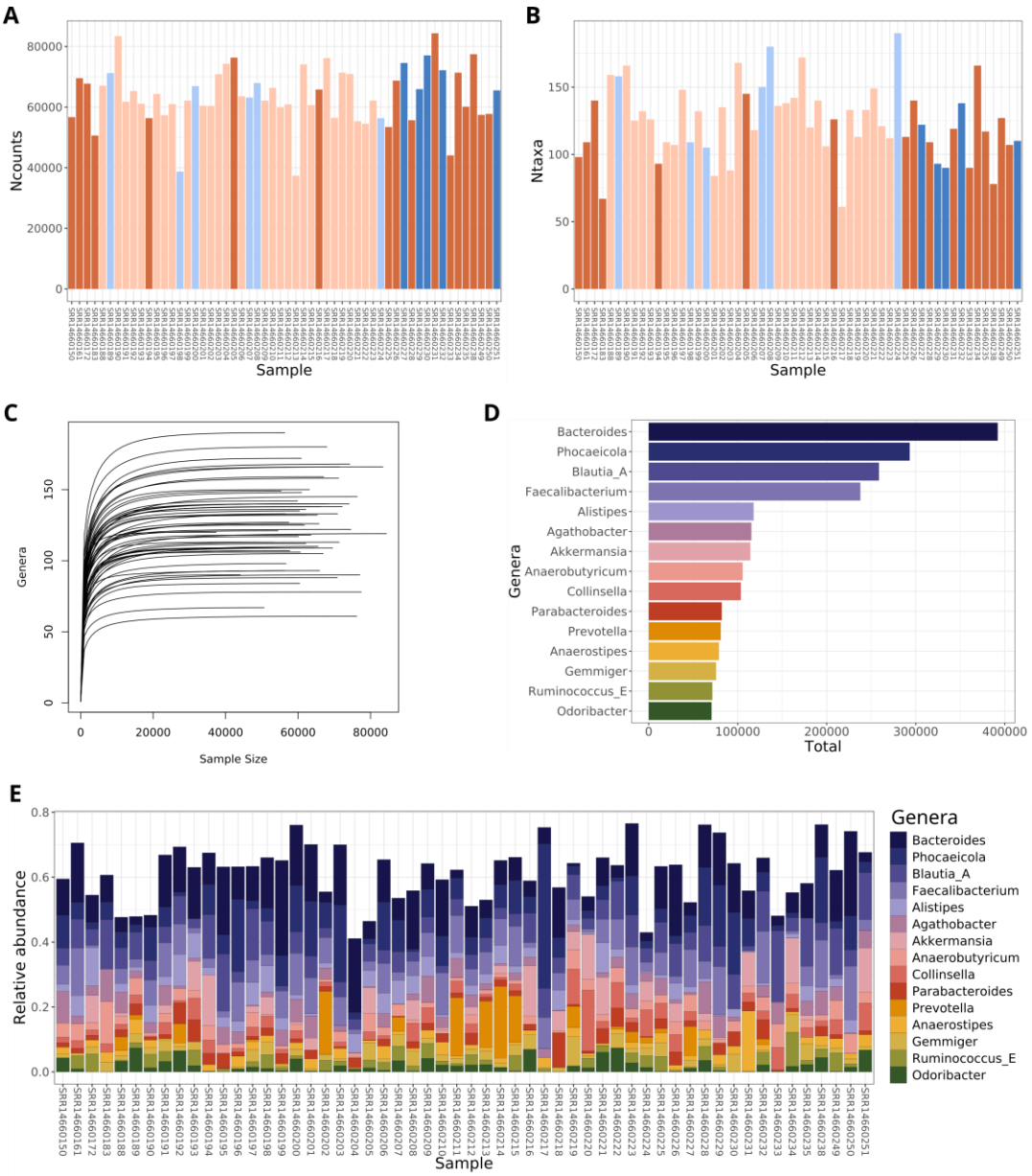
Supplementary Figure 4.S14. Quality control and taxonomic composition metrics for dataset PRJEB99111. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



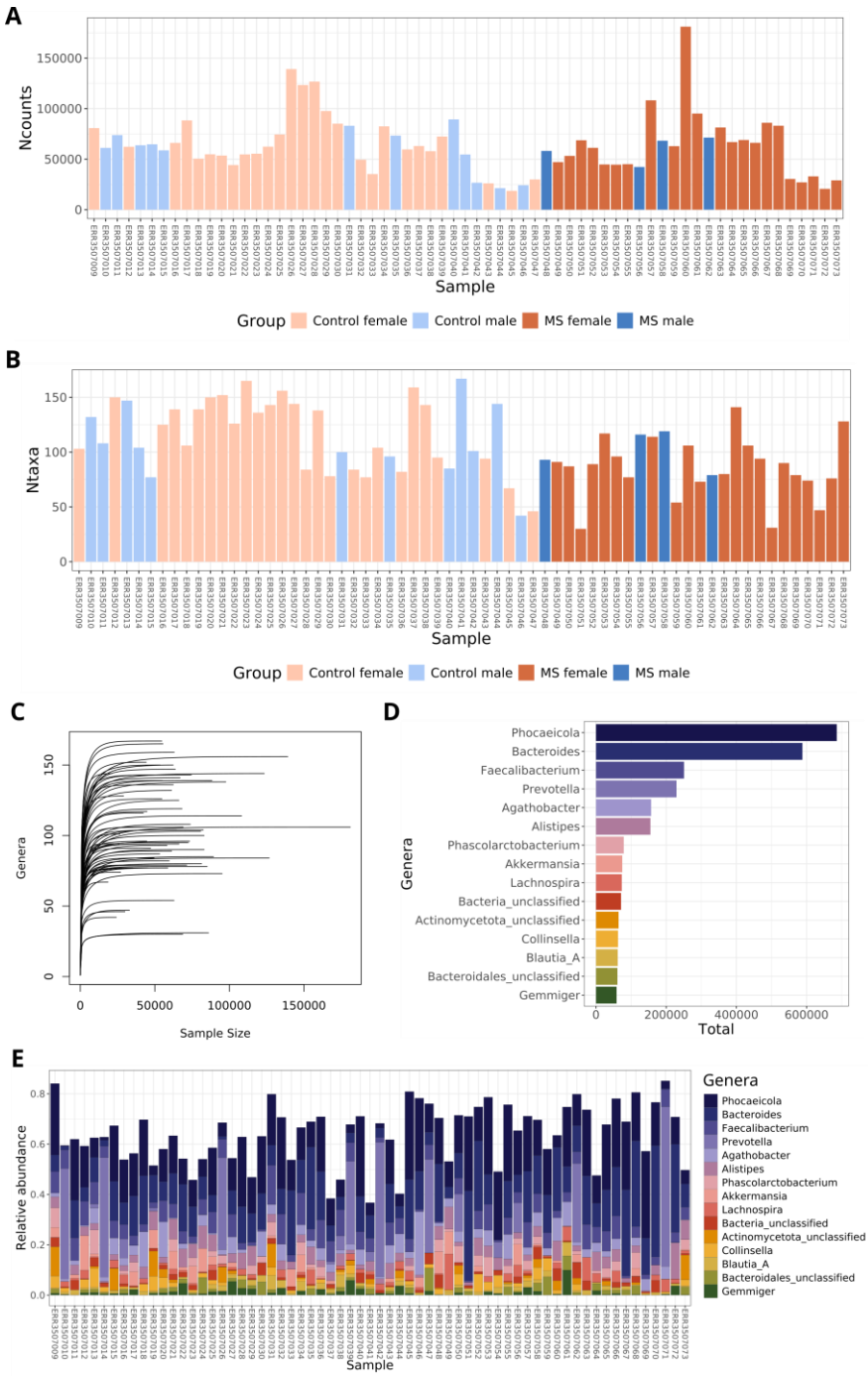
Supplementary Figure 4.S15. Quality control and taxonomic composition metrics for dataset PRJNA748865. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



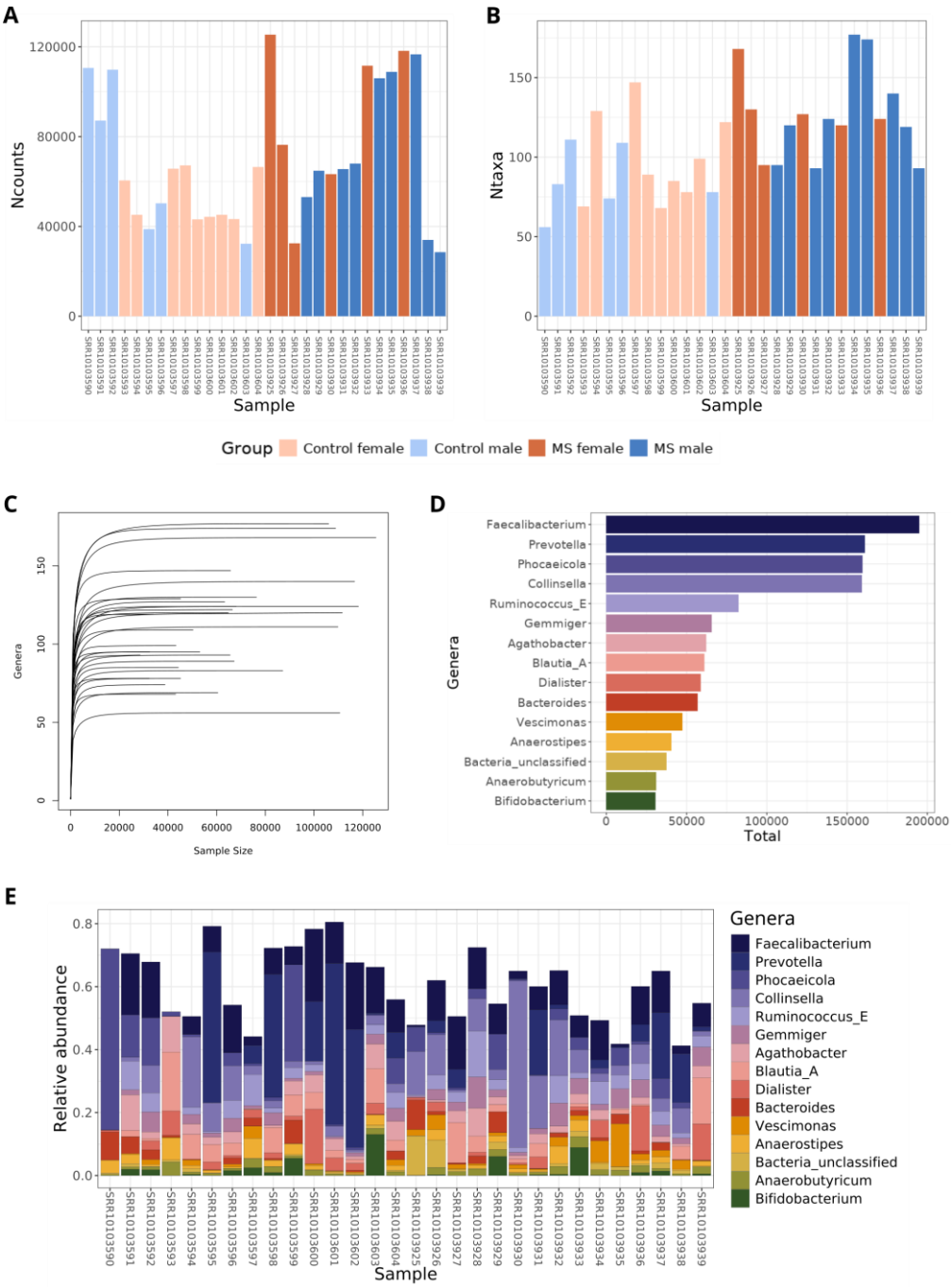
Supplementary Figure 4.S16. Quality control and taxonomic composition metrics for dataset PRJNA889427. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



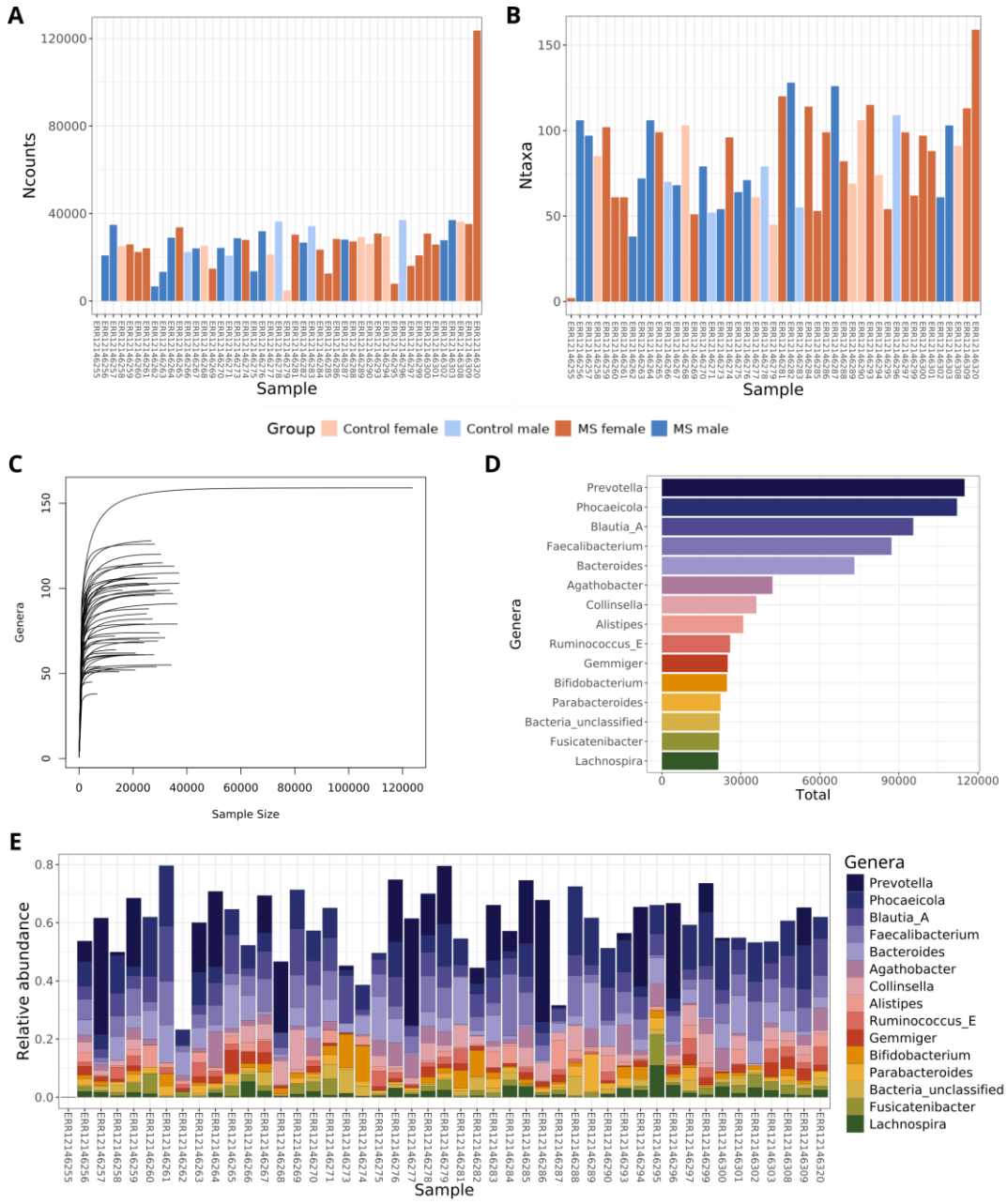
Supplementary Figure 4.S17. Quality control and taxonomic composition metrics for dataset PRJNA732670. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



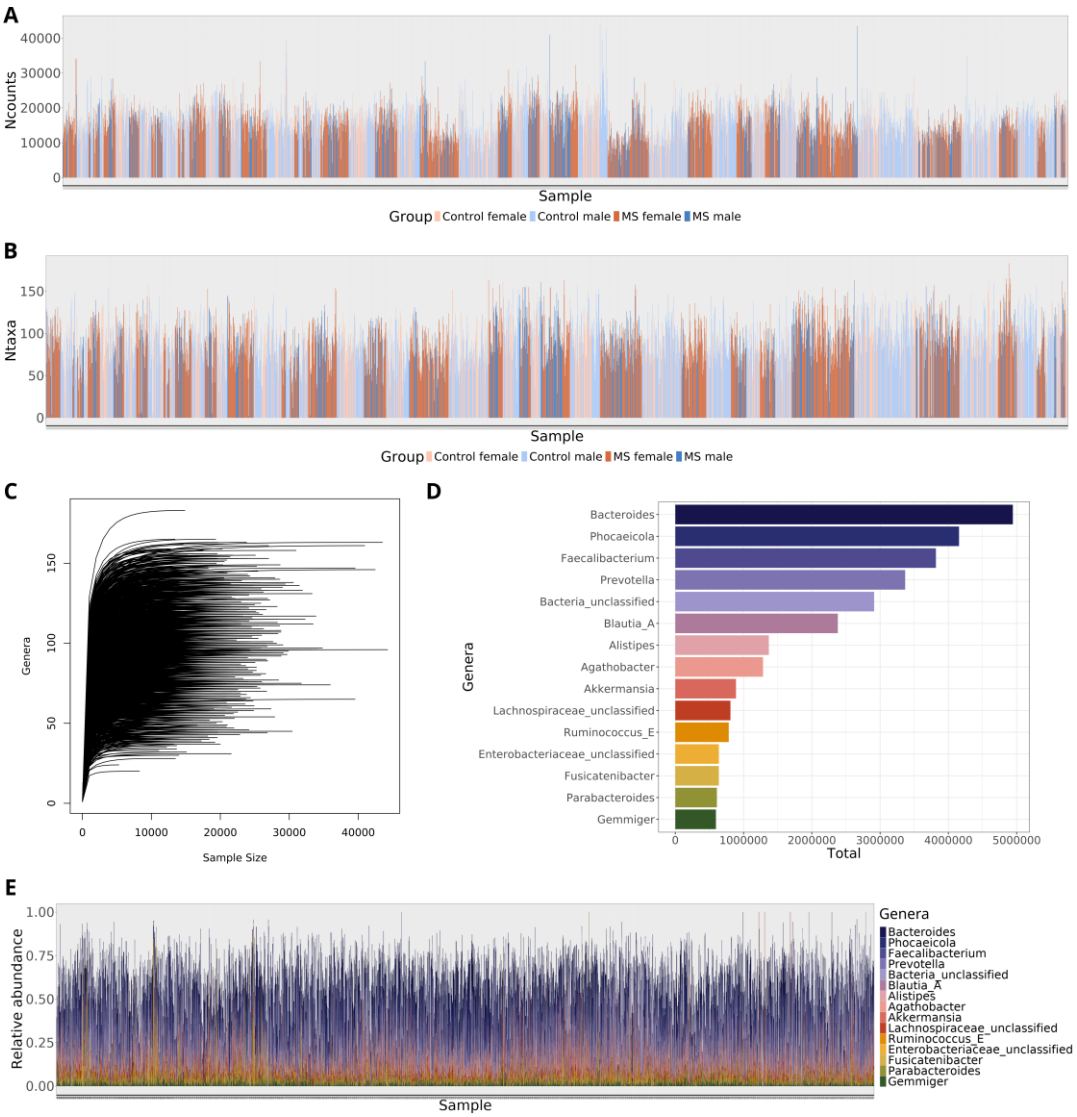
Supplementary Figure 4.S18. Quality control and taxonomic composition metrics for dataset PRJEB34168. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



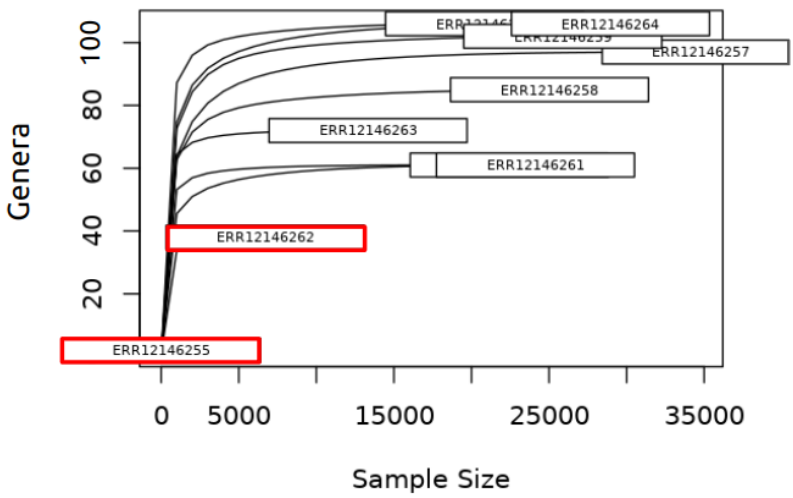
Supplementary Figure 4.S19. Quality control and taxonomic composition metrics for dataset PRJNA565173. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



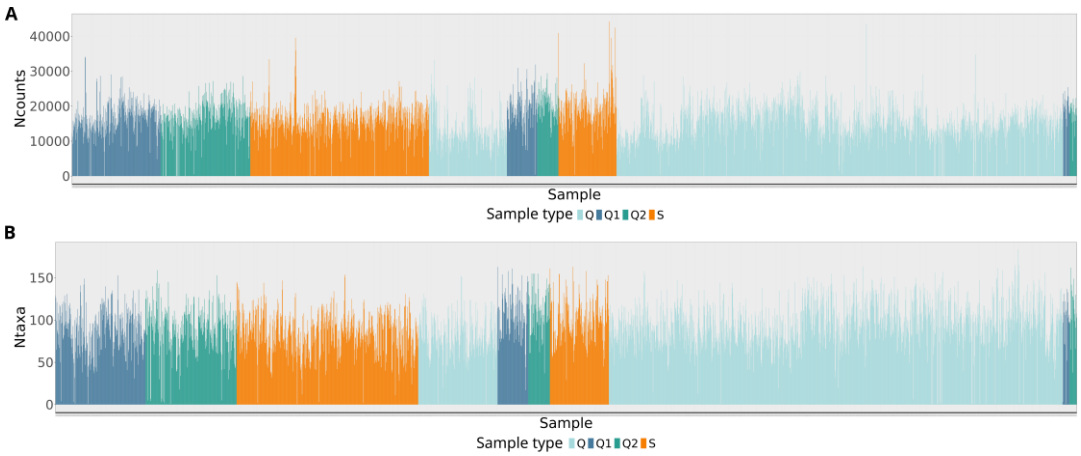
Supplementary Figure 4.S20. Quality control and taxonomic composition metrics for dataset PRJEB67783. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



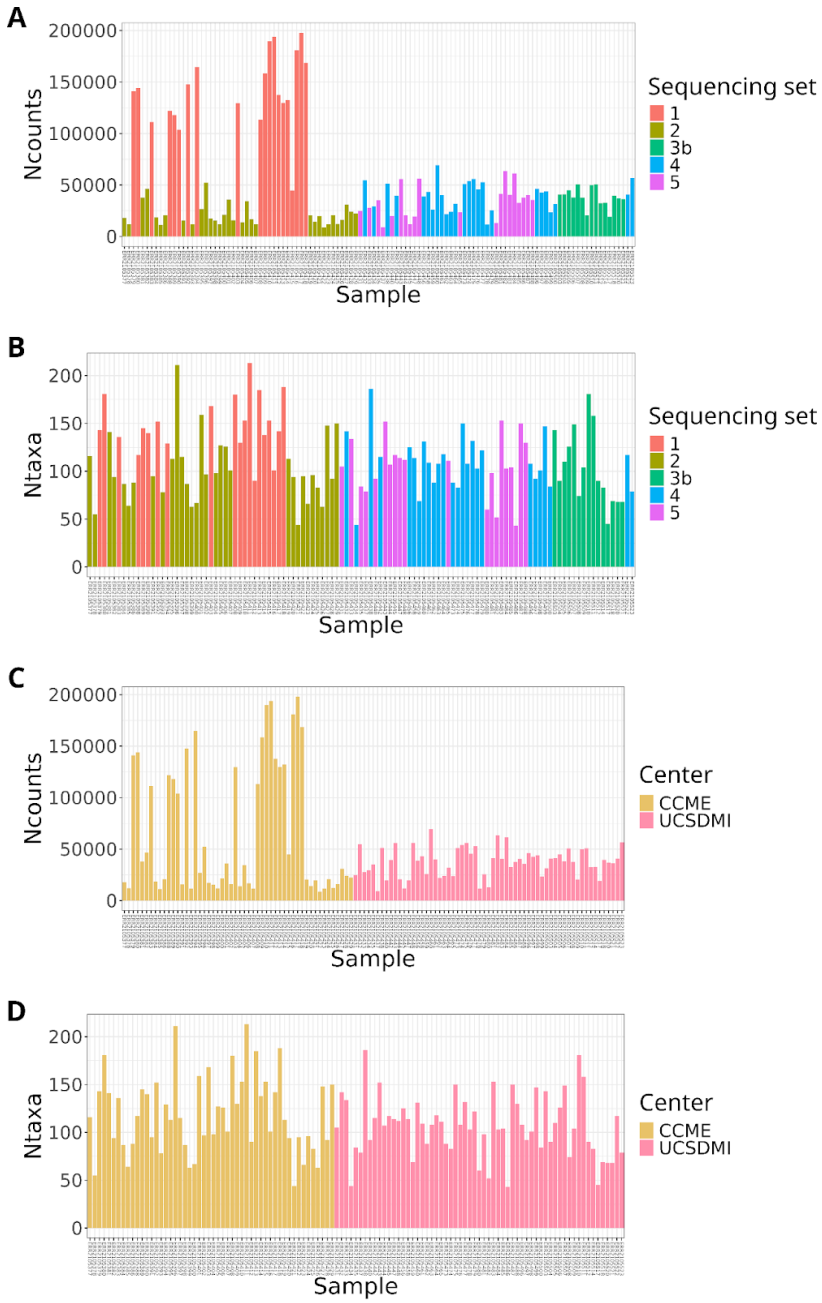
Supplementary Figure 4.S21. Quality control and taxonomic composition metrics for dataset PRJEB32762. (A) Total counts per sample. (B) Number of distinct taxa identified per sample. (C) Rarefaction curves showing the relationship between sequencing depth and taxa detection. (D) Total abundance of the 15 most abundant genera across all samples. (E) Relative abundance of the top 15 genera per sample. *MS: multiple sclerosis.*



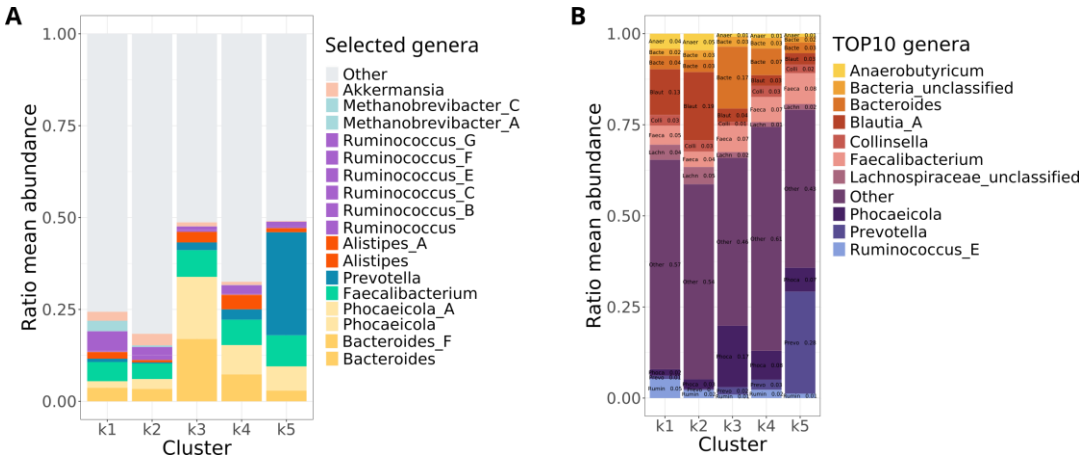
Supplementary Figure 4.S22. Rarefaction curves for a subset of 10 samples from the PRJEB67783 dataset. Red box frame: samples excluded from further analysis as they did not reach the *plateau* phase.



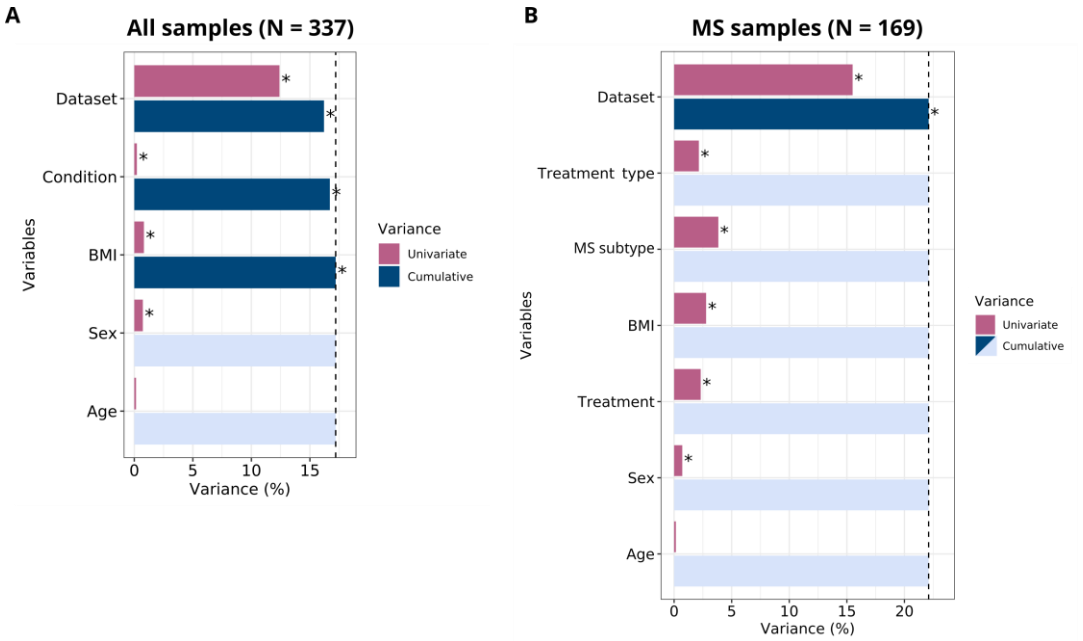
Supplementary Figure 4.S23. Distribution of the (A) total counts and (B) number of distinct taxa identified per sample for PRJEB32762 dataset. Color: sample collection method: dry obtained by Q-tip (Q, Q1, Q2) or wet obtained by frozen the samples (S).



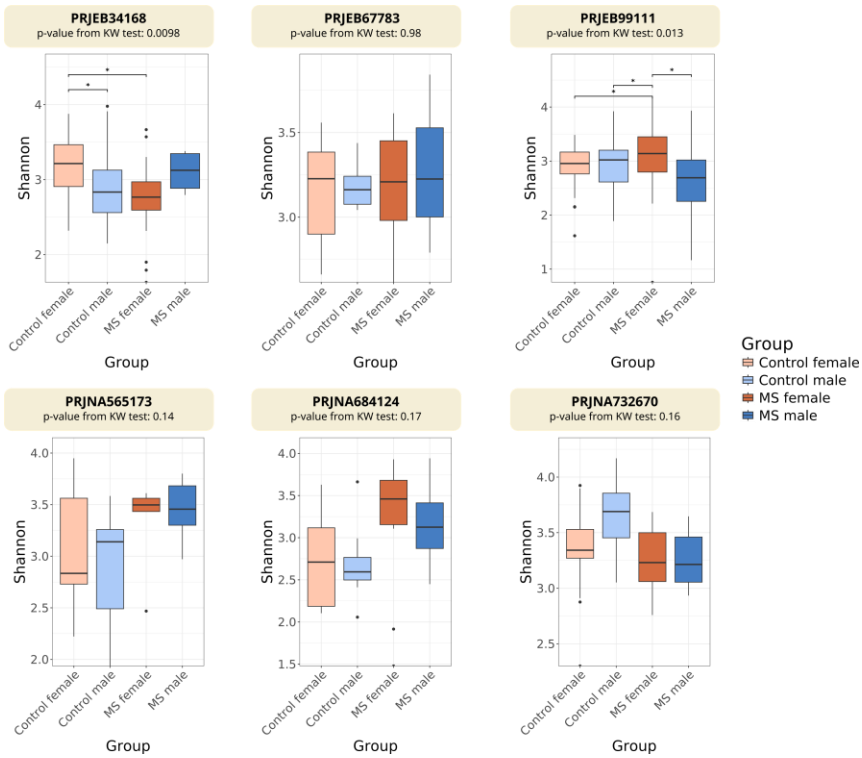
Supplementary Figure 4.S24. Distribution of the total counts and number of distinct taxa identified per sample for PRJEB99111 dataset colored by technical variables. (A-B) Sequencing set and (C-D) center where the sample has been processed.



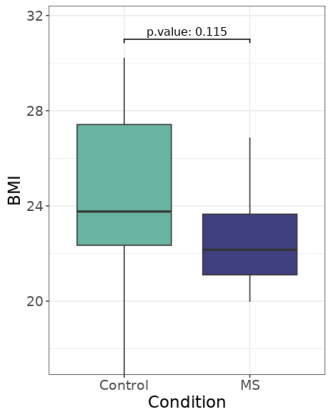
Supplementary Figure 4.S25. Relative genera distribution in the clusters identified with Dirichlet Multinomial Mixture modelling in the combined dataset conformed by the nine selected studies. (A) Selected genera to identify enterotypes. (B) Top 10 most abundant genera.



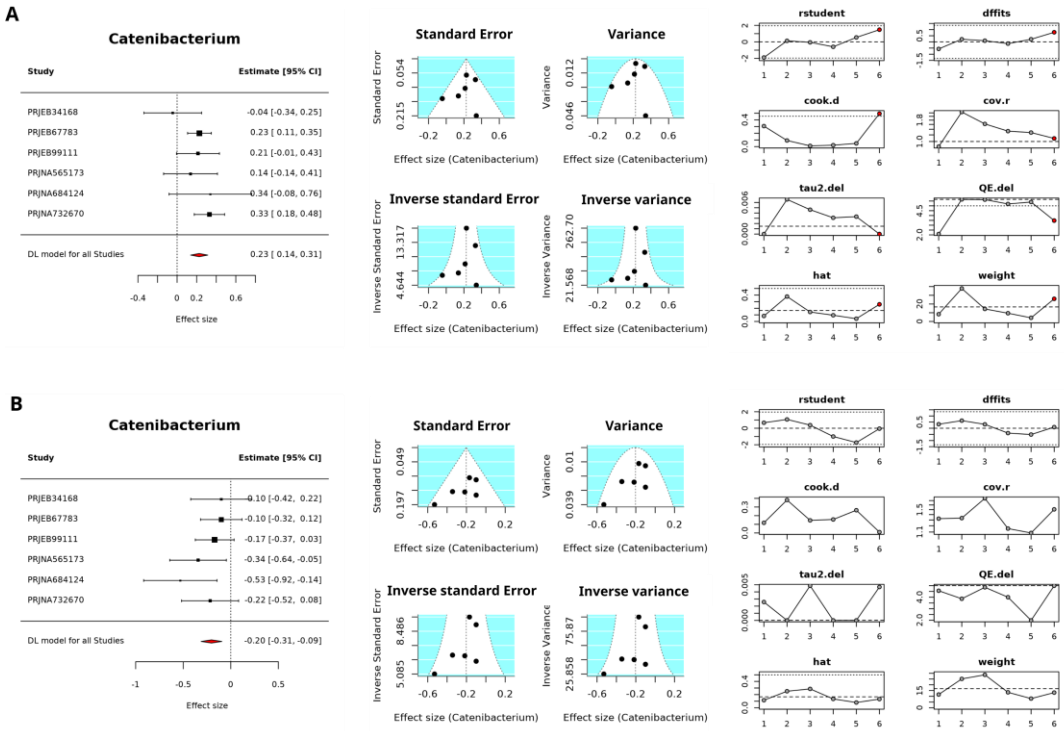
Supplementary Figure 4.S26. Variable contribution to microbiome compositional variation considering all samples from the six datasets included in the meta-analysis. Results considering each variable independently (univariate variance) or in the multivariate model (cumulative variance). The asterisk (*) indicates if the contribution is significant; the black dashed line represents the cut-off for significant non-redundant contribution to the multivariate model considering p-value < 0.05. For multivariate models, samples with missing data are excluded. (A) Number of missing values: age (N = 5), BMI (N = 157). (B) Number of missing values: BMI (N = 75), MS subtype (N = 3). *BMI: body max index; MS: multiple sclerosis.*



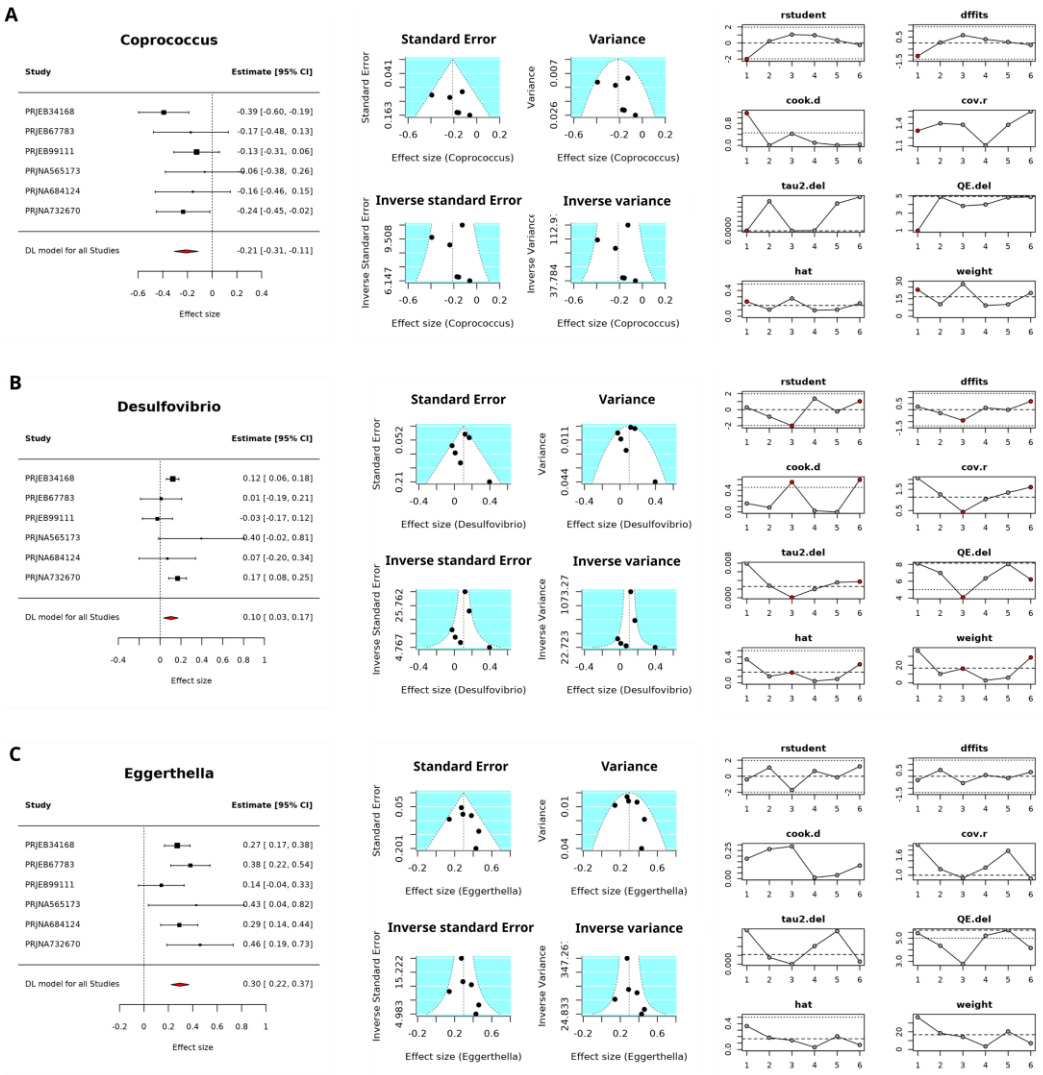
Supplementary Figure 4.S27. Alpha diversity measured by Shannon index across the six datasets included in the meta-analysis. Individual test by dataset comparing the groups defined by condition and sex. Asterisks (*) indicate pairwise significant differences with adjusted p-values < 0.05 from Dunn's post hoc test (after p-value < 0.05 in Kruskal–Wallis test). *KW: Kruskall-Wallis; MS: multiple sclerosis.*



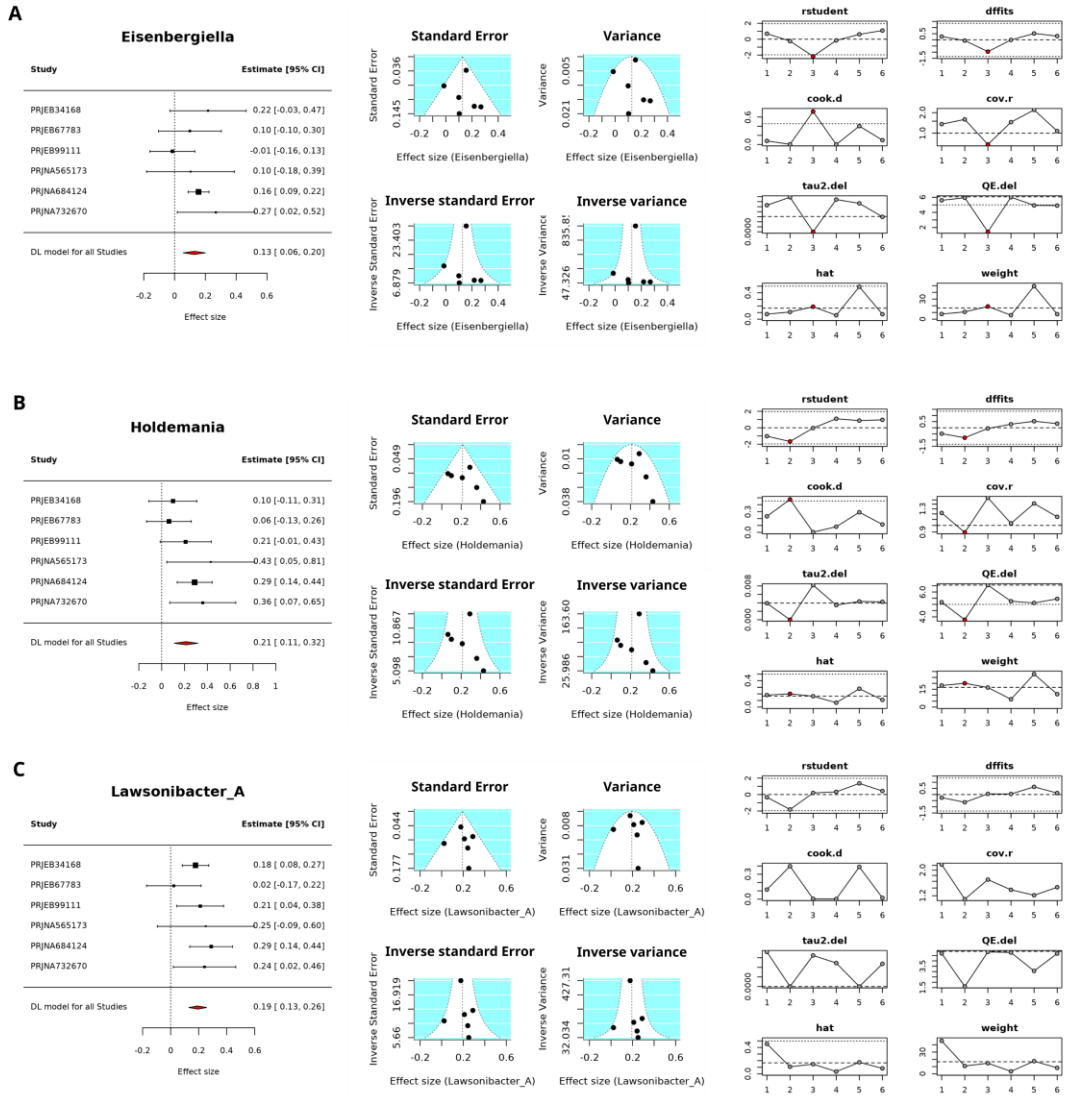
Supplementary Figure 4.S28. Distribution of body mass index by condition for PRJNA565173 dataset. P-value obtained from Wilcoxon rank-sum test. *BMI: body mass index; MS: multiple sclerosis.*



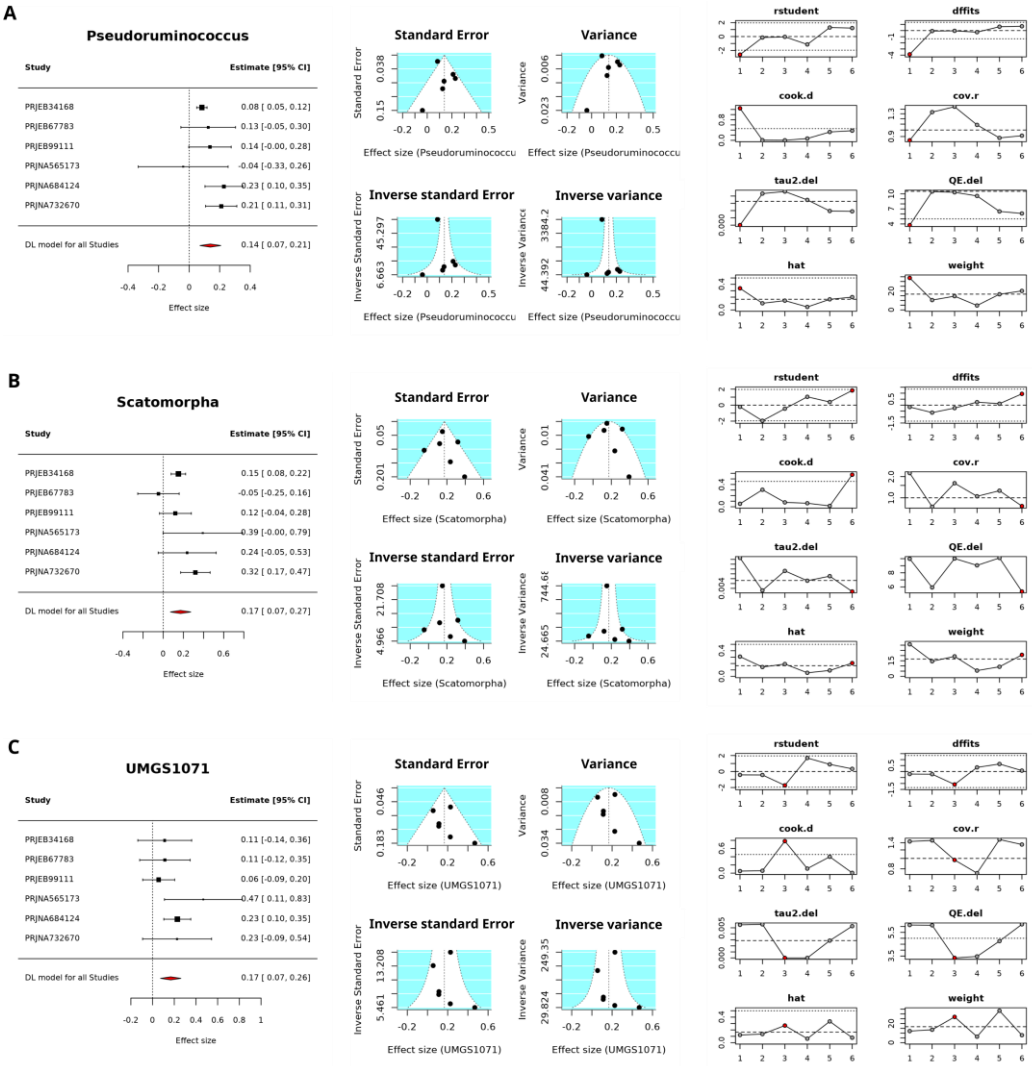
Supplementary Figure 4.S29. Meta-analysis results for *Catenibacterium* when comparing (A) MS males *versus* control males (Male comparison) and (B) MS females *versus* MS males (MS comparison). From left to right: (i) Forest plots: Each point represents the effect size from an individual dataset, with horizontal lines indicating the 95% confidence intervals. The diamond at the bottom represents the combined effect size and its confidence interval. Positive values indicate higher abundance in MS females; negative values indicate higher abundance in MS males. (ii) Funnel plots: effect sizes (X-axis) compared to precision estimates (Y-axis). Each point represents an individual dataset effect size plotted against its standard error (top left), sample variance (top right), inverse standard error (bottom left), and inverse sampling variance (bottom right). The white triangle shows the 95% confidence region under the null hypothesis of no bias. (iii) Influence plots: assessment of each study's influence on the combined effect size using multiple diagnostic metrics. Study ID numbers are ordered based on their position in the forest plot (e.g., PRJEB34168 corresponds to study 1). Detailed description of the metric is defined in the Materials and Methods section. Red dots: influential studies; grey dots: non influential studies. *CI: Confidence Interval; DL: DerSimonian-Laird; ID: Identifier.*



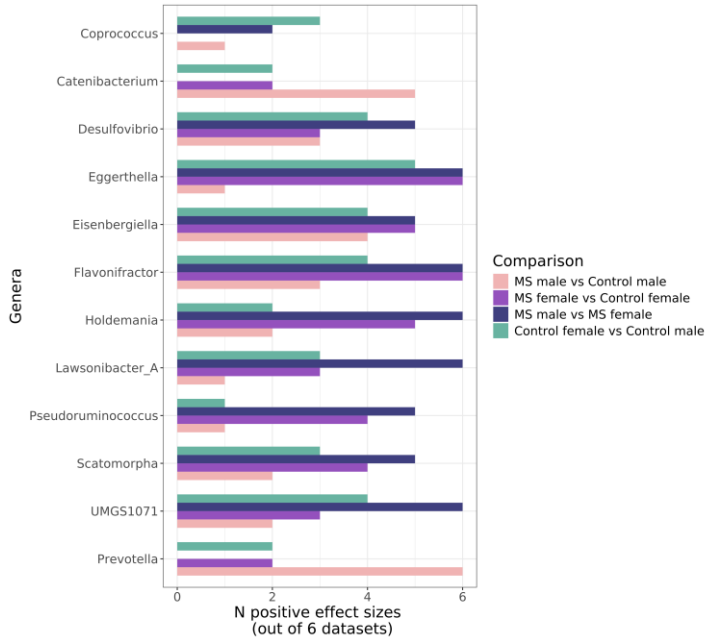
Supplementary Figure 4.S30. Meta-analysis results for (A) *Coprococcus* when comparing (A) MS females *versus* control females (Female comparison) (B) *Desulfovibrio* and (C) *Eggerthella* when comparing MS females *versus* MS males (MS comparison). From left to right: (i) Forest plots: Each point represents the effect size from an individual dataset, with horizontal lines indicating the 95% confidence intervals. The diamond at the bottom represents the combined effect size and its confidence interval. Positive values indicate higher abundance in MS females; negative values indicate higher abundance in MS males. (ii) Funnel plots: effect sizes (X-axis) compared to precision estimates (Y-axis). Each point represents an individual dataset effect size plotted against its standard error (top left), sample variance (top right), inverse standard error (bottom left), and inverse sampling variance (bottom right). The white triangle shows the 95% confidence region under the null hypothesis of no bias. (iii) Influence plots: assessment of each study's influence on the combined effect size using multiple diagnostic metrics. Study ID numbers are ordered based on their position in the forest plot (e.g., PRJEB34168 corresponds to study 1). Detailed description of the metric is defined in the Materials and Methods section. Red dots: influential studies; grey dots: non influential studies. *CI: Confidence Interval; DL: DerSimonian-Laird; ID: Identifier.*



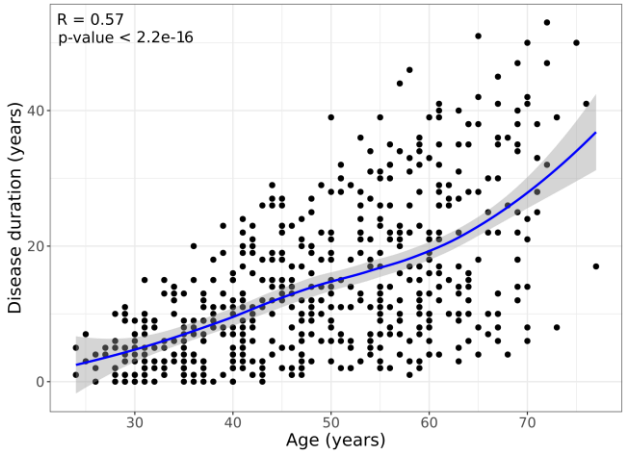
Supplementary Figure 4.S31. Meta-analysis results for (A) *Eisenbergiella* (B) *Holdemania* and (C) *Lawsonibacter_A* when comparing MS females *versus* MS males (MS comparison). From left to right: (i) Forest plots: Each point represents the effect size from an individual dataset, with horizontal lines indicating the 95% confidence intervals. The diamond at the bottom represents the combined effect size and its confidence interval. Positive values indicate higher abundance in MS females; negative values indicate higher abundance in MS males. (ii) Funnel plots: effect sizes (X-axis) compared to precision estimates (Y-axis). Each point represents an individual dataset effect size plotted against its standard error (top left), sample variance (top right), inverse standard error (bottom left), and inverse sampling variance (bottom right). The white triangle shows the 95% confidence region under the null hypothesis of no bias. (iii) Influence plots: assessment of each study's influence on the combined effect size using multiple diagnostic metrics. Study ID numbers are ordered based on their position in the forest plot (e.g., PRJEB34168 corresponds to study 1). Detailed description of the metric is defined in the Materials and Methods section. Red dots: influential studies; grey dots: non influential studies. *CI: Confidence Interval; DL: DerSimonian-Laird; ID: Identifier.*



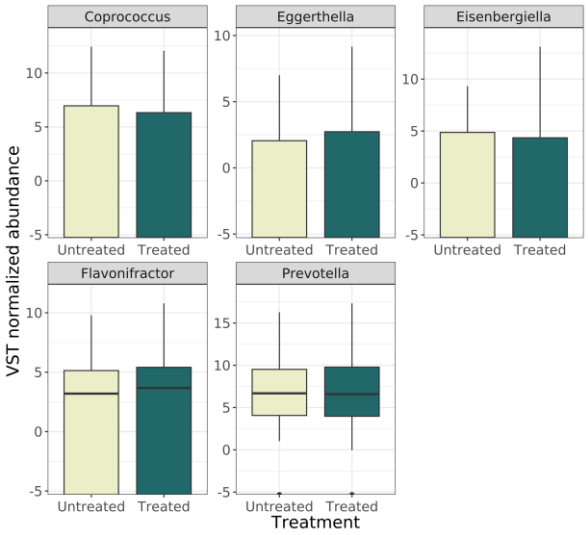
Supplementary Figure 4.S32. Meta-analysis results for (A) *Pseudoruminococcus* (B) *Scatomorpha* and (C) *UMGS1071* when comparing MS females *versus* MS males (MS comparison). From left to right: (i) Forest plots: Each point represents the effect size from an individual dataset, with horizontal lines indicating the 95% confidence intervals. The diamond at the bottom represents the combined effect size and its confidence interval. Positive values indicate higher abundance in MS females; negative values indicate higher abundance in MS males. (ii) Funnel plots: effect sizes (X-axis) compared to precision estimates (Y-axis). Each point represents an individual dataset effect size plotted against its standard error (top left), sample variance (top right), inverse standard error (bottom left), and inverse sampling variance (bottom right). The white triangle shows the 95% confidence region under the null hypothesis of no bias. (iii) Influence plots: assessment of each study's influence on the combined effect size using multiple diagnostic metrics. Study ID numbers are ordered based on their position in the forest plot (e.g., PRJEB34168 corresponds to study 1). Detailed description of the metric is defined in the Materials and Methods section. Red dots: influential studies; grey dots: non influential studies. *CI: Confidence Interval; DL: DerSimonian-Laird; ID: Identifier.*



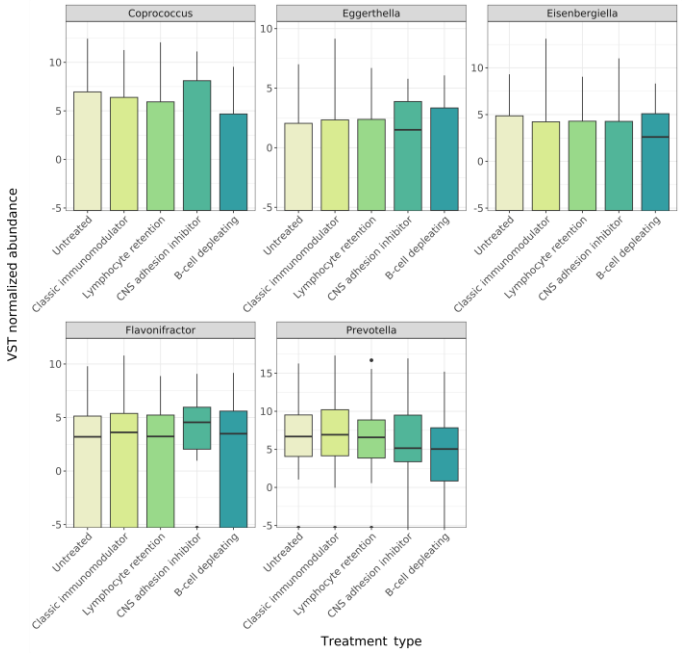
Supplementary Figure 4.S33. Number of individual datasets supporting a positive effect size for each genus across comparisons. Each bar represents the number of individual studies (out of six) reporting a positive effect size (i.e., greater abundance in the first group of the comparison) for a given genus. The x-axis corresponds to genera with a significant meta-analysis result (adjusted p-value < 0.05), grouped by comparison: Control (control females vs. control males), Female (MS females vs. control females), Male (MS males vs. control males), and MS (MS females vs. MS males). *MS: multiple sclerosis.*



Supplementary Figure 4.S34. Age and disease duration correlation in multiple sclerosis samples. Each dot represents an individual sample. The blue line represents the fitted linear regression trend. The shaded gray area around the blue line corresponds to the 95% confidence interval of the regression. R: rho (Spearman coefficient correlation).



Supplementary Figure 4.S35. Normalized abundance of the validated genera across treated and untreated patients within the multiple sclerosis cohort of the validation dataset. Kruskal–Wallis test adjusted p-values: *Coprococcus* (0.55), *Eisenbergiella* (0.91), *Eggerthella* (0.75), *Flavonifractor* (0.55), and *Prevotella* (0.91). *VST: variance stabilizing transformation.*



Supplementary Figure 4.S36. Normalized abundance of the validated genera across treatment status within the multiple sclerosis cohort of the validation dataset. Kruskal–Wallis test adjusted p-values: *Coprococcus* (0.66), *Eisenbergiella* (0.68), *Eggerthella* (0.49), *Flavonifractor* (0.49), and *Prevotella* (0.49). *CNS: central nervous system; VST: variance stabilizing transformation.*

Package	Version	Package	Version
FastQC	0.11.8	microbiome	1.24.0
MultiQC	1.12.0	DT	0.33
Cutadapt	4.6	rstatix	0.7.2
dada2	1.30.0	boot	1.3-28.1
dplyr	1.1.4	metafor	4.8.0
stringr	1.5.1	ggsignif	0.6.4
ggplot2	3.5.1	gridExtra	2.3.0
phyloseq	1.46.0	dunn.test	1.3.6
vegan	2.6.8	coin	1.4.3
DirichletMultinomial	1.44.0	ComplexHeatmap	2.18.0
MetBrewer	0.2.0	ggpubr	0.6.0

Supplementary Table 4.S1. List of command-line tools and R packages with the corresponding versions implemented in the bioinformatic workflow.

Study	Original name	Homogenized name	Treatment type	Study	Original name	Homogenized name	Treatment type
PRJNA684124	Calidibrine	Cladribine	Lymphocyte depleating	**PRJNA721421**	TYSABRI	Natalizumab	CNS adhesion inhibitor
PRJNA684124	Copaxone- Deltacortene-Chlorphenamine	Glatiramer acetate	Classic immunomodulator	**PRJNA721421**	Washout	Untreated	Untreated
PRJNA684124	Copaxone-Etzolam- Naloxone-Tamsulosin	Glatiramer acetate	Classic immunomodulator	**PRJNA748865**	Avonex	Interferon-beta	Classic immunomodulator
PRJNA684124	Deltacortene-Chlorphenamine	Corticosteroid	Corticosteroid	**PRJNA748865**	Betaferon	Interferon-beta	Classic immunomodulator
PRJNA684124	Efexor- Norvasc- antihypertensive-Atorvastatin- cardiospirina-Vitamin D	Untreated	Untreated	**PRJNA748865**	Gilenya	Fingolimod	Lymphocyte retention
PRJNA684124	Euthyrox- Copaxone	Glatiramer acetate	Classic immunomodulator	**PRJNA748865**	Rebif 44	Interferon-beta	Classic immunomodulator
PRJNA684124	Gilenya – Vitamin D- Efexor 37,5mg-Mayafer-Deursil	Fingolimod	Lymphocyte retention	**PRJNA748865**	Tecfidera	Dimethyl fumarate	Classic immunomodulator
PRJNA684124	Lioresal 25mg- Entact 10 mg - Urogyn-Oxybutynin 5mg	Untreated	Untreated	**PRJNA732670**	Avonex	Interferon-beta	Classic immunomodulator
PRJNA684124	Sativix (cannabis) - Lioresal - Tiobec	Untreated	Untreated	**PRJNA732670**	Copaxone	Glatiramer acetate	Classic immunomodulator
PRJNA684124	Teriflunomide-Simvastatin-Dicloreum antiinflammatory- Vitamin D	Teriflunomide	Lymphocytes depleating	**PRJNA732670**	Ocrevus	Ocrelizumab	B-cell depleating
PRJNA684124	Untreated	Untreated	Untreated	**PRJNA732670**	Rebif	Interferon-beta	Classic immunomodulator
PRJNA684124	Vitamin D	Untreated	Untreated	**PRJNA732670**	Tecfidera	Dimethyl fumarate	Classic immunomodulator
PRJNA684124	Vitamin D- Atorvastatin	Untreated	Untreated	**PRJEB67783**	AUBAGIO	Teriflunomide	Lymphocyte depleating
PRJNA721421	anti-CD20	anti-CD20	B-cell depleating	**PRJEB67783**	AVONAX+AZATIOPRINA	Interferon-beta + Azathioprine	Classic immunomodulator + Lymphocyte depleating
PRJNA721421	AUBAGIO	Teriflunomide	Lymphocyte depleating	**PRJEB67783**	AVONEX	Interferon-beta	Classic immunomodulator
PRJNA721421	BIOTIN	Untreated	Untreated	**PRJEB67783**	CLADRIBINA	Cladribine	Lymphocyte depleating
PRJNA721421	CELLCEPT	Mycophenolate mofetil	Lymphocyte depleating	**PRJEB67783**	COPAXONE	Glatiramer acetate	Classic immunomodulator

PRJNA721421	COPAXONE	Glatiramer acetate	Classic immunomodulator	**PRJEB67783**	DIMETILFUMARATO	Dimethyl fumarate	Classic immunomodulator
PRJNA721421	GILENYA	Fingolimod	Lymphocyte retention	**PRJEB67783**	FINGOLIMOD	Fingolimod	Lymphocyte retention
PRJNA721421	Interferon-beta	Interferon-beta	Classic immunomodulator	**PRJEB67783**	GILENUA	Fingolimod	Lymphocyte retention
PRJNA721421	METHOTREXATE_NOTRADENAME_	Methotrexate	Lymphocyte depleating	**PRJEB67783**	IFN-BETA	Interferon-beta	Classic immunomodulator
PRJNA721421	RILUTEK	Riluzol	Benzothiazoles	**PRJEB67783**	NATALIZUMAB	Natalizumab	CNS adhesion inhibitor
PRJNA721421	RILUTEK_RITUXAN	Rituximab	B-cell depleating	**PRJEB67783**	OCREVUS	Ocrelizumab	B-cell depleating
PRJNA721421	RITUXAN_COPAXONE	Rituximab and glatiramer acetate	Classic immunomodulator and B-cell depleating	**PRJEB67783**	PLEGRIDY	Interferon-beta	Classic immunomodulator
PRJNA721421	Solumedrol	Corticosteroid	Corticosteroid	**PRJEB67783**	TECFIDERA	Dimethyl fumarate	Classic immunomodulator
PRJNA721421	TECFIDERA	Dimethyl fumarate	Classic immunomodulator	**PRJEB67783**	TISABRY	Natalizumab	CNS adhesion inhibitor
PRJNA721421	TECFIDERA_AUBAGIO	Dimethyl fumarate and Teriflunomide	Classic immunomodulator and Lymphocyte depleating	**PRJEB67783**	TYSABRI	Natalizumab	CNS adhesion inhibitor
PRJNA721421	TECFIDERA_COPAXONE	Dimethyl fumarate and Glatiramer acetate	Classic immunomodulator				

Supplementary Table 4.S2. **Standardization of MS treatment names across studies**. Original treatment names as reported in each study, their homogenized nomenclature implemented in this doctoral thesis, and the classification of treatments into broader categories according to their mechanism of action. *CNS: central nervous system.*

Dataset	Discarded low quality samples
PRJNA684124	none
PRJNA721421	SRR14215212
PRJEB99111	ERR2105516, ERR2105513, ERR2105502
PRJNA748865	none
PRJNA889427	none
PRJNA732670	none
PRJEB34168	none
PRJNA565173	none
PRJEB67783	ERR12146255, ERR12146262, ERR12146279, ERR12146303
PRJEB32762	ERR7012222, ERR3339587, ERR3339623, ERR3339628, ERR3339667, ERR3339843, ERR3339850, ERR3339851, ERR3339885, ERR3339886, ERR3339889, ERR3339893, ERR3339912, ERR3339937, ERR3340058, ERR3340167, ERR3340282, ERR3340296, ERR3340345, ERR3340474, ERR3340568, ERR6941718, ERR6941873, ERR6941988, ERR6942109, ERR6942384, ERR6942493, ERR6942555, ERR6942627, ERR6942960, ERR6943208, ERR6943447, ERR7012054, ERR7012059, ERR7012073, ERR7012077, ERR7012082, ERR7012083, ERR7012090, ERR7012197, ERR7012198, ERR7012204, ERR7012212, ERR7012239, ERR7012267, ERR7012339, ERR7012352, ERR7012391, ERR7012469, ERR7012537, ERR7012547

Supplementary Table 4.S3. List of discarded low-quality samples by dataset. The table includes the dataset identifier and the corresponding sample identifiers that were excluded based on quality control criteria.

Dataset	Sample size	Community typing results
PRJNA684124	30	No clusters identified
PRJEB99111	113	Two clusters identified: one *Bacteroides*-driven and one *Prevotella*-driven
PRJNA732670	56	Two clusters identified: one *Bacteroides*-driven and one *Prevotella*-driven
PRJEB34168	65	No clusters identified
PRJNA565173	30	No clusters identified
PRJEB67783	43	No clusters identified

Supplementary Table 4.S4. Summary of the results obtained from unsupervised community typing based on Dirichlet Multinomial Mixture modelling applied individually to each dataset. The table includes the number of samples analysed per study and the main findings regarding the presence or absence of distinct clusters.

10.3. SUPPLEMENTARY MATERIAL STUDY III

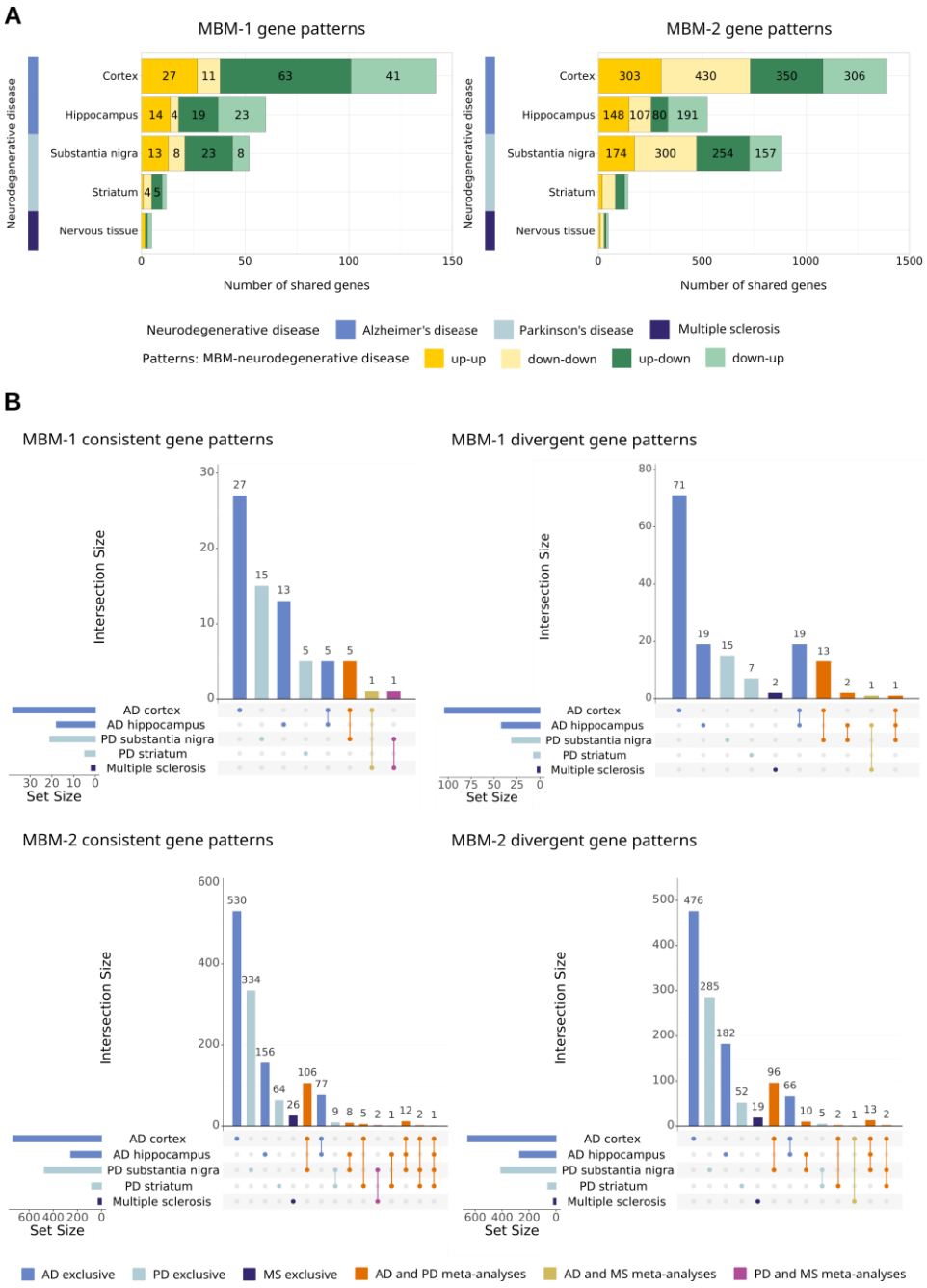


Figure 5.S1. Individual intersection analyses for MBM-1 and MBM-2. (a) Distribution of the number of dysregulated genes in both MBM and each neurodegenerative disease categorized by pattern (MBM-1 right panel, MBM-2 left panel). (b) MBM-1 (top panels) and MBM-2 (bottom

panels) upset plots for genes displaying consistent (up-up and down-down MBM-neurodegenerative patterns) and divergent (up-down and down-up MBM-neurodegenerative patterns) signatures. Representation of the total number of significant genes by neurodegenerative disease (horizontal bars), and the fraction present in the intersection of the designated groups (vertical bars), identified by colored dots beneath. *AD: Alzheimer's disease; MBM: melanoma brain metastasis; MS: multiple sclerosis; PD: Parkinson's disease.*

Package	Version	Package	Version
ggplot2	3.4.0	clusterProfiler	4.2.2
UpSetR	1.4.0	org.Hs.eg.db	3.14.0
stringr	1.4.0	AnnotationDbi	1.56.2
ggpubr	0.4.0	egg	0.4.5
ggh4x	0.2.3	reshape2	1.4.4
tidyverse	1.3.1	ComplexHeatmap	2.13.4
dplyr	1.0.8	enrichplot	1.14.2
tidyr	1.2.0		

Table 5.S1. List of R packages and versions implemented in the bioinformatic workflow.

Table 5.S2. Number of significantly differentially expressed genes in diseased vs normal tissue per study. (*Next page*) The study identifier of the public repository, the neurodegenerative disease, and the brain region assessed are indicated. The number of differentially expressed genes in cases vs. controls is reported, with upregulated genes being those more expressed in cases (logFC > 0), and downregulated genes those more expressed in controls (logFC < 0). *AD: Alzheimer's disease; CT: cortex; HP: hippocampus; ID: identifier; MS: multiple sclerosis; PD: Parkinson's disease; SN: substantia nigra; ST: striatum.*

Study ID	Neurodegenerative disease	Brain region	Upregulated genes	Downregulated genes
GSE118553	AD	CT	1858	1367
GSE125583	AD	CT	5706	5303
GSE132903	AD	CT	7631	6627
GSE15222	AD	CT	4859	3756
GSE37263	AD	CT	0	0
GSE48350	AD	CT	61	64
GSE5281	AD	CT	1200	5012
GSE84422	AD	CT	439	455
GSE1297	AD	HP	1	0
GSE29378	AD	HP	801	497
GSE48350	AD	HP	6092	1856
GSE5281	AD	HP	1132	902
GSE84422	AD	HP	0	0
E-MEXP-1416	PD	SN	0	0
GSE20295	PD	SN	42	32
GSE7621	PD	SN	0	2
GSE8397	PD	SN	2001	1381
GSE20295	PD	ST	0	0
GSE20146	PD	ST	0	0
GSE28894	PD	ST	90	144
GSE108000	MS	Brain	2038	3114
GSE111972	MS	Corpus callosum and occipital cortex	0	0
GSE131281	MS	Brain	162	188
GSE135511	MS	Motor cortex	451	401