

# Diferencias de sexo en la esquizofrenia

Estudios in Silico - Máster en Bioinformática UV

Celine García Rodríguez

2026-02-04

La esquizofrenia es un trastorno neuropsiquiátrico caracterizado por una elevada heterogeneidad clínica y molecular, en el que las diferencias asociadas al sexo juegan un papel clave en la incidencia, sintomatología y respuesta al tratamiento. En este contexto, las tecnologías de secuenciación de RNA a nivel de célula única permiten estudiar con alta resolución los mecanismos moleculares subyacentes a la enfermedad. El presente trabajo tiene como objetivo evaluar la reproducibilidad del estudio Single-Cell Transcriptional Profiling Reveals Cell Type-Specific Sex-Dependent Molecular Patterns of Schizophrenia mediante un análisis in silico de datos de single-nuclei RNA-seq de corteza prefrontal humana, divididos por sexo y condición clínica.

Para ello, se analizaron 48 muestras correspondientes a individuos control y pacientes con esquizofrenia, empleando un workflow basado en Seurat y Harmony para la integración de datos, seguido de análisis de expresión diferencial y enriquecimiento funcional mediante GSEA. Se identificaron 16 tipos celulares, incluyendo subpoblaciones neuronales excitatorias, interneuronas inhibitorias y células gliales. Los resultados revelan un marcado dimorfismo sexual en los perfiles transcripcionales, con un mayor número de genes diferencialmente expresados en hombres y una regulación opuesta dependiente del sexo especialmente en astrocitos y oligodendrocitos, lo cual se refleja en las funciones biológicas expresadas en cada uno de ellos.

## 1 Introducción

La esquizofrenia es una de las enfermedades mentales más devastadoras [1]. Las diferencias en la sintomatología, respuesta al tratamiento y asociadas al sexo tienen un gran impacto en la calidad de vida de los pacientes así como su entorno y sistema sanitario. aunque el término es relativamente reciente en comparación con otras enfermedades, fue a mediados del siglo XIX cuando comenzó a describirse de forma sistemática.

Esta enfermedad afecta a más de 27 millones de personas a nivel mundial y puede estar asociada tanto a factores genéticos, de riesgo o sociodemográficos. En España, más de 135.000 individuos fueron diagnosticados en 2023 [5]. Generalmente, el ratio hombre:mujer es de aproximadamente 1.4:1 con una mayor incidencia en hombres entre 35-54 años mientras que en mujeres rondan los 45-54 años[7]. Actualmente el diagnóstico es complejo y debe ser realizado por un especialista, quien emplea distintas escalas psicopatológicas como la PANS(*Positive and Negative Syndrome Scale*) o la MCCB (*MATRICES Consensus Cognitive Battery*).

Aunque los síntomas se han asociado a múltiples neurotransmisores, se considera que la clave de la enfermedad radica en la disfunción del sistema dopaminérgico subcortical y alteraciones del sistema GABAérgico [6]. Se ha demostrado la existencia de diferencias biológicas entre sexos. Además de la incidencia, en los hombres suelen predominar los síntomas negativos como la falta de motivación, dificultad para expresar emociones y el aislamiento social, los cuales pueden llegar a confundirse con otras enfermedades como la depresión e incluso evolucionar a estados de catatonia[2]. En cambio, las mujeres muestran generalmente síntomas positivos caracterizados por alteraciones en el pensamiento, conducta y percepción de la realidad, con manifestaciones en forma de delirios o alucinaciones. No existen diferencias con respecto a la tercera categoría de síntomas, los cognitivos que engloban la falta de atención, concentración y memoria con dificultad para mantener una conversación o adquirir nuevos conocimientos.

La hipótesis del estrógeno es una de las principales explicaciones propuestas para las diferencias asociadas al sexo, defendiendo el papel neuroprotector del estrógeno endógeno producido por las mujeres. Sin embargo, los mecanismos de regulación génica implicados aún no han sido completamente identificados lo que limita el desarrollo de nuevas terapias. Actualmente, el tratamiento se basa principalmente en la administración de antipsicóticos de primera generación ( como la clorpromazina) o de segunda generación ( como la quetiapina), combinado con cuidados especializados[4].

En este contexto, la tecnología Single-Cell RNA sequencing se ha posicionado como una poderosa herramienta transcriptómica clave para el estudio de neuropatologías complejas ya que permite obtener información precisa sobre la citoarquitectura cerebral y mecanismos moleculares asociados[3,8]. *Single-Cell Transcriptional Profiling Reveals Cell Type-Specific Sex-Dependent Molecular Patterns of Schizophrenia* (Universidad de Nankai, China)[9] analiza distintos datasets de la corteza prefrontal incorporando un cribado de fármacos y la hipótesis del estrógeno. También pudo identificar factores de transcripción asociados a la enfermedad así como mapas de conectividad enfermedad-gen-medicación.

El objetivo de este trabajo de *Estudios in silico en Bioinformática* ha sido replicar este estudio para evaluar la reproducibilidad de los resultados hasta el análisis de enriquecimiento funcional.

## 2 Metodología

### 2.1 Archivos, Cluster y lenguaje de programación

Para descargar las muestras utilizadas por los autores, se recurrió al portal de *BrainScope* ([https://brainscope.gersteinlab.org/data/snrna\\_expr\\_matrices\\_zip/SZBDMulti-Seq.zip](https://brainscope.gersteinlab.org/data/snrna_expr_matrices_zip/SZBDMulti-Seq.zip) 2 Noviembre 2025). La información relativa al sexo, edad y diagnóstico fue proporcionada en como Material suplementario. El dataset original (SZBDMulti-sesq) contiene 72 muestras de *Single-nuclei RNA-seq*, incluyendo pacientes con trastorno bipolar, los cuales se eliminaron al no tener relevancia en el estudio. Las 48 muestras restantes se pueden dividir por sexo (hombres, mujeres) y condición (control-esquizofrenia), de manera que se trabaja con cuatro grupos de 12 (Hombre\_SCZ, Hombre\_CON, Mujeres\_SCZ, Mujeres\_CON) con una matriz de conteo por paciente.

Los datos se almacenaron en el clúster computacional del Centro de Investigación Príncipe Felipe (CIPF) para ejecutar el análisis bioinformático de las mismas. Este cuenta con 20 nodos de 744 procesadores, memoria RAM de 9 Tb y almacenamiento de 1Pb. Se ha escogido RStudio (4.4.2) para replicar este análisis ya que es el que usan los autores y además, es uno de los lenguajes más completos para datos transcriptómicos.

### 2.2 Paquetes

El paquete principal empleado en este trabajo fue Seurat v5, una herramienta muy utilizada para el control de calidad, análisis y exploración de datos de single-cell RNA-seq (sc-RNAseq) y single-nuclei RNA-seq (sn-RNAseq). Adicionalmente, se emplearon paquetes complementarios para visualización, integración de datos, análisis de expresión diferencial y enriquecimiento funcional.

```
library(Seurat)
library(patchwork)
library(ggplot2)
library(R.utils)
library(dplyr)
library(harmony)
library(qs2)
library(EnhancedVolcano)
library(ComplexUpset)
library(clusterProfiler)
library(org.Hs.eg.db)
library(purrr)
library(ComplexHeatmap)
```

```
library(circlize)
library(tidyr)
library(tidytext)# BiocManager::install("GSEAmining")
```

## 2.3 WorkFlow

### 2.3.1 Procesado de datos

#### 1. Creación del Objeto Seurat

Uno de los principales problemas iniciales fue que cada paciente presentaba una matriz de conteos diferente, en la que las columnas correspondían a tipos celulares en lugar de barcodes. Para poder integrar todos los datos en un único objeto Seurat y conservar correctamente la información de los metadatos, las muestras se organizaron en cuatro carpetas según sexo y condición clínica:

- SCZ\_M: hombres con esquizofrenia
- SCZ\_F: mujeres con esquizofrenia
- CON\_H: hombres control
- CON\_F: mujeres control

En primer lugar, los datos fueron cargados individualmente en objetos Seurat.

```
# Load Folder Function

load_folder_to_seurat <- function(dir_path, sex, case) {
  archivos <- list.files(
    path = dir_path,
    pattern = "\\*.txt$",
    full.names = TRUE
  )

  objetos <- lapply(archivos, function(f) {
    m <- read.table(f, header = TRUE, sep = "\t", row.names = 1)

    seu <- CreateSeuratObject(counts = m, assay = "RNA")

    # ID sample
    sample_id <- tools::file_path_sans_ext(basename(f))
    sample_id <- sub("-annotated_matrix$", "", sample_id)
```

```

# Metadatos
seu$sample <- sample_id
seu$sex <- sex
seu$case <- case
seu$group <- paste0(case, "_", substr(sex, 1, 1))

return(seu)
})

return(objetos)
}

# Loading folders ...

SCZ_F_list <- load_folder_to_seurat(
  "/home/cgarcia/files/SCZ_descompr/SCZ_F",
  sex = "Female",
  case = "SCZ"
)

SCZ_M_list <- load_folder_to_seurat(
  "/home/cgarcia/files/SCZ_descompr/SCZ_M",
  sex = "Male",
  case = "SCZ"
)

CON_F_list <- load_folder_to_seurat(
  "/home/cgarcia/files/SCZ_descompr/CON_F",
  sex = "Female",
  case = "CONTROL"
)

CON_M_list <- load_folder_to_seurat(
  "/home/cgarcia/files/SCZ_descompr/CON_M",
  sex = "Male",
  case = "CONTROL"
)

```

Posteriormente, se realizó una intersección de genes, ya que los *dataframes* introducidos debían contener exactamente el mismo conjunto de genes. Este paso es necesario porque Seurat requiere que todos los objetos a fusionar tengan el mismo número de genes. Tras este filtrado, todos los objetos fueron unidos en un único objeto Seurat.

```

# Common Genes
common_genes <- Reduce(intersect, lapply(all_objects,
                                          function(x) rownames(x)))

# Filtering
all_objects_filtered <- lapply(all_objects, function(s) s[common_genes, ])

# Merge
ids <- sapply(all_objects_filtered, function(x) unique(x$sample))

SCZ_48 <- merge(
  x = all_objects_filtered[[1]],
  y = all_objects_filtered[-1],
  add.cell.ids = ids,
  project = "SCZ_project"
)

```

Finalmente, fue imprescindible aplicar la función `JoinLayers`, ya que, de lo contrario, el objeto Seurat habría conservado múltiples capas independientes (una por matriz importada). Las capas están diseñadas para almacenar distintos estudios, pero en este caso el objeto interpretaba cada matriz como un estudio independiente, lo que incrementaba exponencialmente el tamaño del objeto y dificultaba su manejo. Sin este paso, el tamaño final del objeto podría haber alcanzado hasta 128 GB tras el preprocesamiento completo.

La función *JoinLayers* solo puede aplicarse cuando se cumplen las siguientes condiciones:

- Todas las capas pertenecen al mismo assay (RNA).
- Todas las capas contienen el mismo conjunto de genes, previamente filtrados mediante intersección.

```

SCZ_48 <- JoinLayers(SCZ_48, merge.assays = TRUE)
qs_save(SCZ_48, "SCZ_48.qs2") # Qs2: saving/Loading R Objects

```

## 2. Control de Calidad

Para seleccionar núcleos de alta calidad, se aplicaron los siguientes criterios de filtrado:

- `nFeature_RNA > 200`: núcleos que expresan más de 200 genes.
- `percent.mt < 5`: porcentaje de genes mitocondriales inferior al 5 %, ya que valores elevados suelen indicar daño celular.

```
# Searching MT genes
SCZ_48[["percent.mt"]] <- PercentageFeatureSet(SCZ_48, pattern = "^MT-")

# Filtering
SCZ_48 <- subset(SCZ_48, subset = nFeature_RNA > 200 & percent.mt < 5)
```

## 2.3.2 Reducción de Dimensionalidad e Integración de datos

### 4. Normalización

Corrige las diferencias en profundidad de secuenciación entre células, haciendo comparables los niveles de expresión génica entre ellas.

Los conteos de expresión de cada célula se normalizaron dividiendo el total de conteos por célula y multiplicando el resultado por 10 000. Posteriormente, los valores se transformaron logarítmicamente para estabilizar la varianza y reducir el impacto de los genes altamente expresados.

```
SCZ_48 <- NormalizeData(SCZ_48, normalization.method = "LogNormalize",
                        scale.factor = 10000)
```

### 5. Identificación de genes altamente variables (HVGs)

Tras la normalización, se identificaron los 2 000 genes más variables (Highly Variable Genes, HVGs) ya que serán los más informativos a la hora de capturar heterogeneidad biológica. Además, reduce dimensionalidad y mejora la clusterización.

Se seleccionó la función *FindVariableFeatures*, concretamente el método vst, el cual ajusta la relación del logaritmo de la varianza y el logaritmo de la media a partir de la media observada y varianza esperada.

```
SCZ_48 <- FindVariableFeatures(SCZ_48, selection.method = "vst",
                              nfeatures = 2000)

# Identify the 10 most highly variable genes
top10 <- head(VariableFeatures(SCZ_48), 10)

# Variable Features Plot
plot1 <- VariableFeaturePlot(SCZ_48)
plot2 <- LabelPoints(plot = plot1, points = top10, repel = TRUE)
plot2
```

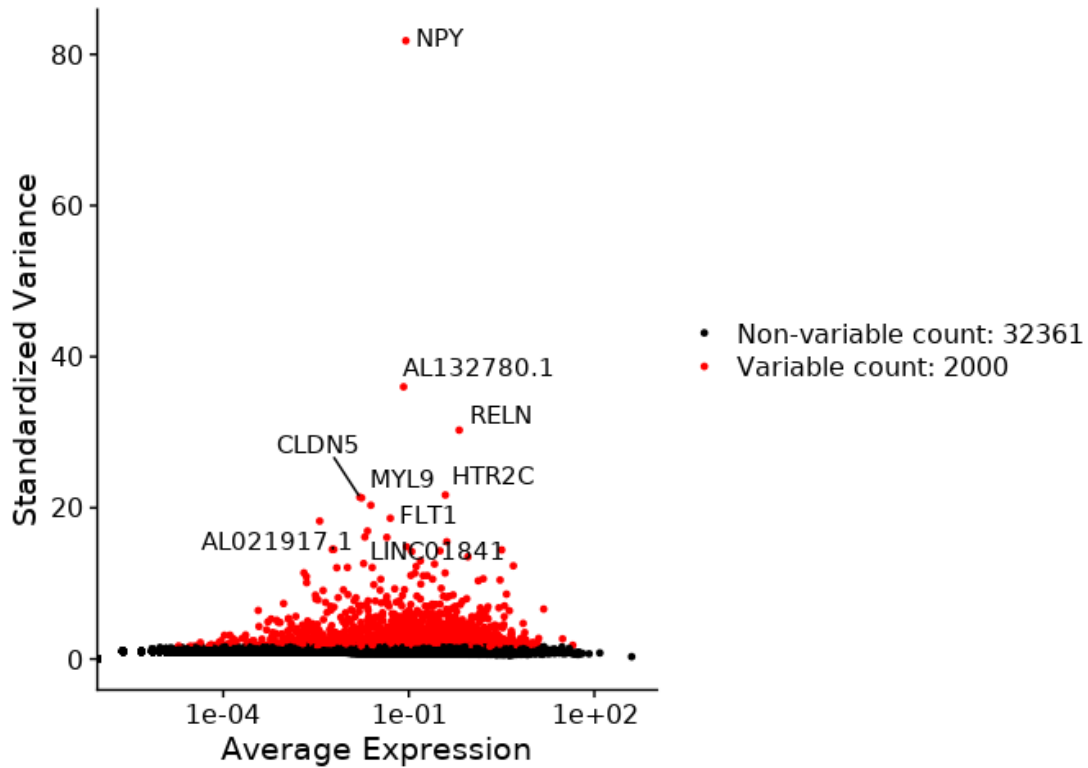


Figure 1: Genes Altamente Variables (HVG)

## 6. Escalado de datos (*Scaling*)

Posteriormente, los datos fueron escalados con el fin de estandarizar la expresión génica para que todos los genes contribuyan de forma comparable en análisis posteriores. Esto previene que los genes HVG dominen el Análisis de Componentes Principales (PCA), el cual reduce la dimensionalidad de los datos manteniendo la máxima variabilidad posible. Se calcularon 50 componentes principales.

```
SCZ_48 <- ScaleData(SCZ_48) # by default, only HVF
SCZ_48 <- RunPCA(SCZ_48,
                  features = VariableFeatures(SCZ_48),
                  npcs = 50)
qs_save(SCZ_48, "SCZ_48_postPCA.qs2")
```

## 7. Integración de datos mediante *Harmony*

Para integrar las 48 muestras y corregir posibles efectos batch, se empleó el paquete Harmony, ampliamente utilizado en metaanálisis de datos de single-cell.

```
SCZ_48 <- RunHarmony(SCZ_48, group.by.vars = ("sample"))
Embeddings(SCZ_48, "harmony")[1:5, 1:5] # Check the results

SCZ_48 <- RunUMAP(SCZ_48, reduction = "harmony", dims = 1:20)
qs_save(SCZ_48, "SCZ_48_UMAP.qs2")
```

En este caso, el algoritmo convergió tras tres iteraciones. A continuación, se calculó una reducción de dimensionalidad UMAP (*Uniform Manifold Approximation and Projection*) para visualizar relaciones entre células en un espacio bidimensional.

### 2.3.3 Clusterización y Anotación celular

La clusterización agrupa células con perfiles de expresión similares. Se llevó a cabo utilizando las funciones *FindNeighbors* y *FindClusters*. Tras la corrección con *Harmony*, se optimizaron los parámetros empleando 20 componentes principales, un valor de k-nearest neighbors de 36 y una resolución de 0.3. Los clústeres que contenían menos de 2 000 células fueron excluidos del análisis.

```
SCZ_48 <- FindNeighbors(SCZ_48, reduction = "harmony", dims = 1:20,
                        k.param = 36)
SCZ_48 <- FindClusters(SCZ_48, resolution = 0.3, verbose = FALSE)

# Filtering clusters < 2000 cells
cluster_sizes <- table(Idents(SCZ_48_cluster))
final_clusters <- names(cluster_sizes[cluster_sizes >= 2000])
SCZ_48_cluster <- subset(SCZ_48_cluster, idents = final_clusters)
```

Como resultado, el número total de clústeres se redujo de 23 iniciales a 19 clústeres finales. Aunque los autores del artículo original utilizaron la función *FindMarkers* para identificar genes marcadores en cada clúster, en este trabajo se optó por *FindAllMarkers*, lo que permitió optimizar y sistematizar el análisis. Los genes marcadores finales se filtraron utilizando un valor de  $p$  ajustado inferior a 0.05 y un  $\log_2\text{FC} > \log_2(1.5)$ .

```
clusterSCZ.markers <- FindAllMarkers(SCZ_48,
                                     only.pos = TRUE,
                                     test.use = "wilcox")

markers_filtrered <- clusterSCZ.markers %>%
  filter(p_val_adj < 0.05 & avg_log2FC > log2(1.5))
```

Para la anotación celular, se utilizó la base de datos *CellKB*, la cual permite identificar subtipos celulares en función de los genes expresados. Se seleccionaron los quince genes principales de cada clúster ordenados por *avg\_log2FC* aunque se tuvieron en cuenta los valores de *pct.1* (proporción de expresión dentro del clúster) y *pct.2* (proporción de expresión en otros clústeres), priorizando genes con alta especificidad.

Finalmente, mediante la función *RenameIdents*, se visualizaron los tipos celulares anotados en el UMAP, observando las distancias relativas entre ellos.

```
SCZ_48 <- RenameIdents(  
  SCZ_48,  
  "0" = "L2/3",  
  "1" = "Oligodendrocyte",  
  "2" = "L2",  
  "3" = "L4/5",  
  "4" = "L6",  
  "5" = "VIP",  
  "6" = "Astrocyte",  
  "7" = "PVALB",  
  "8" = "L4/5",  
  "9" = "L6",  
  "10" = "SST",  
  "11" = "OPC",  
  "12" = "L4",  
  "13" = "LAMP5",  
  "14" = "Microglia",  
  "15" = "L5/6",  
  "16" = "L5",  
  "17" = "L5/6",  
  "18" = "Astrocyte",  
  "19" = "Chandelier Cell"  
)  
  
DimPlot(SCZ_48, reduction = "umap", label = TRUE)
```

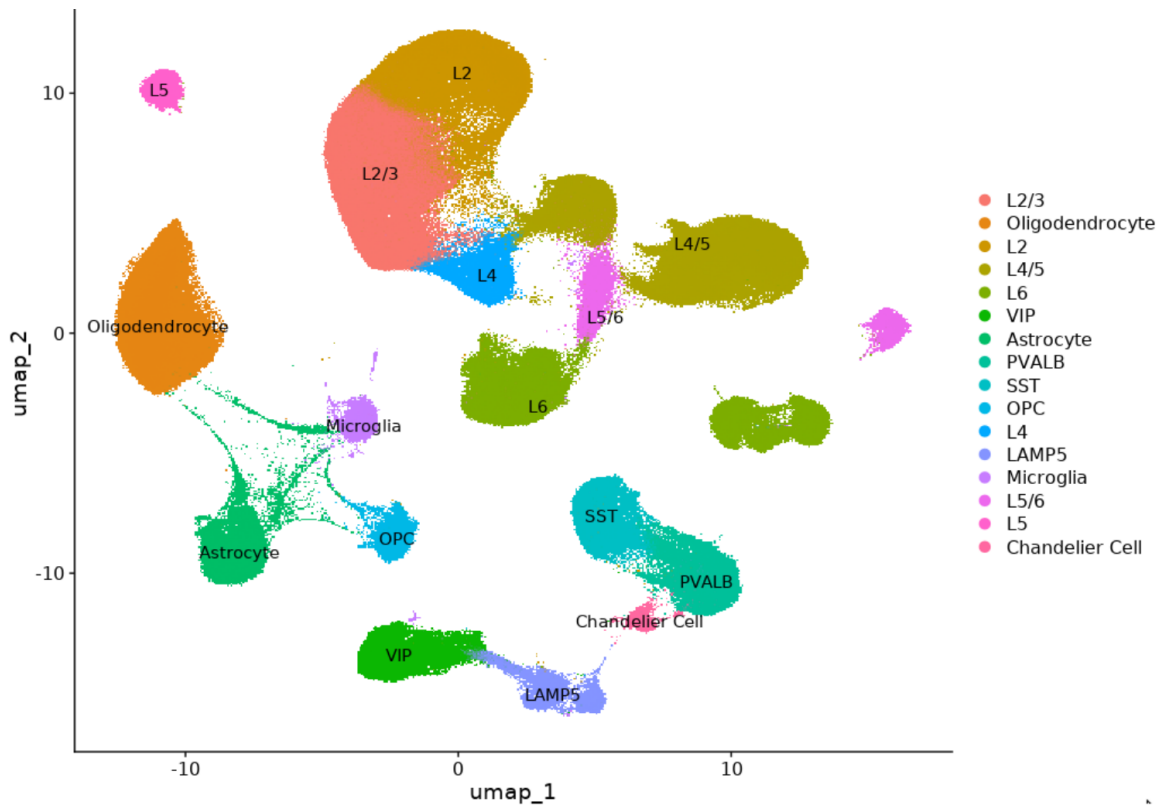


Figure 2: UMAP con Anotación Celular

### Categorías:

- Neuronas Excitatorias: L2, L2/3, L4, L4/5, L5, L5/6 y L6
- Interneuronas Inhibitorias: Células de Chandler, SST, VIP, Lamp5 y Pvalb
- Células no neuronales: células gliales (Astrocitos y Microglía) y linaje oligodendrocito (OPC, Oligodendrocitos)

### 2.3.4 Análisis de Expresión Diferencial

Una vez identificados los tipos celulares, se realizó un análisis de expresión diferencial por tipo celular y por sexo, comparando casos (SCZ) frente a controles (CON). El resultado consistió en una lista de genes diferencialmente expresados para cada tipo celular, estratificada por sexo.

Se utilizó la función *FindMarkers* con un test no paramétrico de *Wilcoxon*, muy empleado en análisis de sc-RNAseq. Además, se estableció un umbral de  $\log FC = 0.25$ , en concordancia con el criterio utilizado por los autores del artículo original.

```

DEG_results <- list()

for (ct in levels(SCZ_48)) {

  obj_ct <- subset(SCZ_48, ids = ct)

  # Male
  male_ct <- subset(obj_ct, subset = sex == "Male")
  Idents(male_ct) <- "case"

  DEG_results[[paste0(ct, "_male")]] <- FindMarkers(
    male_ct,
    ident.1 = "SCZ",
    ident.2 = "CONTROL",
    logfc.threshold = 0.25,
    test.use = "wilcox"
  )

  # Female
  female_ct <- subset(obj_ct, subset = sex == "Female")

  Idents(female_ct) <- "case"

  DEG_results[[paste0(ct, "_female")]] <- FindMarkers(
    female_ct,
    ident.1 = "SCZ",
    ident.2 = "CONTROL",
    logfc.threshold = 0.25,
    test.use = "wilcox"
  )
}

qs_save(DEG_results, "/home/cgarcia/files/scripts/DEG_48.qs2")

```

Los resultados obtenidos fueron filtrados seleccionando aquellos genes con un p ajustado inferior a 0.05 y un log2FC absoluto superior a 1.

```

# DEG Filtered

DEG_results <- qs_read("DEG_48.qs2")
DEG_48 <- lapply(DEG_results, function(df){

```

```
df[df$p_val_adj < 0.05 & abs(df$avg_log2FC) > 1, ]
})
sapply(DEG_48, nrow)
```

### 2.3.5 Enriquecimiento Funcional

Para el análisis de enriquecimiento funcional se empleó GSEA, al considerarse un enfoque más robusto y preciso. El análisis se realizó utilizando el paquete *clusterProfiler*, centrándose en procesos biológicos (BP).

Para cada tipo celular, los genes se dividieron en tres categorías: genes comunes a ambos sexos, genes específicos de hombres y genes específicos de mujeres.

```
# 1. Identifying celltypes
celltypes <- unique(gsub("_ (male|female)$", "", names(DEG_48)))

# 2. List of common and specific genes
DEG_comparative <- lapply(celltypes, function(ct) {

  df_m <- DEG_48[[paste0(ct, "_male")]]
  df_f <- DEG_48[[paste0(ct, "_female")]]

  genes_m <- rownames(df_m)
  genes_f <- rownames(df_f)

  list(
    celltype = ct,
    male = df_m,
    female = df_f,

    common = intersect(genes_m, genes_f),
    male_only = setdiff(genes_m, genes_f),
    female_only = setdiff(genes_f, genes_m)
  )
})

names(DEG_comparative) <- celltypes
```

A partir de estos conjuntos, se identificaron los términos GO asociados tanto a genes regulados positivamente como negativamente.

```

go_results_list <- list()

for (ct in celltypes) {
  message(paste("Procesando tipo celular:", ct))

  m_only <- DEG_comparative[[ct]]$male_only
  f_only <- DEG_comparative[[ct]]$female_only

  m_df <- DEG_comparative[[ct]]$male[m_only, , drop = FALSE]
  f_df <- DEG_comparative[[ct]]$female[f_only, , drop = FALSE]

  subsets <- list(
    Male_Up      = rownames(m_df[m_df$avg_log2FC > 0, , drop = FALSE]),
    Male_Down    = rownames(m_df[m_df$avg_log2FC < 0, , drop = FALSE]),
    Female_Up    = rownames(f_df[f_df$avg_log2FC > 0, , drop = FALSE]),
    Female_Down  = rownames(f_df[f_df$avg_log2FC < 0, , drop = FALSE])
  )
  go_results_list[[ct]] <- subsets
}

# Ranking GSEA
ranking_gsea <- function(df) {
  ranks <- df$avg_log2FC
  names(ranks) <- rownames(df)

  ranks <- ranks[!is.na(ranks)]
  ranks <- sort(ranks, decreasing = TRUE)

  return(ranks)
}

```

Como los datos proceden de humanos, se seleccionó la base *org.Hs.eg.db* y se ajustaron los p-valores con el método de Benjamini-Hochberg.

```

gsea_ct_sex <- function(ct, sex, DEG_comparative) {

  message(paste("GSEA:", ct, "-", sex))

  df <- DEG_comparative[[ct]][[sex]] # <- CLAVE

  ranks <- ranking_gsea(df)

```

```

gsea_res <- gseG0(
  geneList      = ranks,
  OrgDb         = org.Hs.eg.db,
  keyType       = "SYMBOL",
  ont           = "BP",
  minGSSize     = 10,
  maxGSSize     = 500,
  pAdjustMethod = "BH",
  pvalueCutoff  = 1
)

if (is.null(gsea_res) || nrow(gsea_res@result) == 0) {
  return(NULL)
}

gsea_res@result %>%
  mutate(
    celltype = ct,
    sex      = sex,
    direction = ifelse(NES > 0, "Up", "Down")
  )
}

gsea_results <- map_dfr(
  celltypes,
  ~bind_rows(
    gsea_ct_sex(.x, "male", DEG_comparative),
    gsea_ct_sex(.x, "female", DEG_comparative)
  )
)

```

## 3 Resultados

### 3.1 Barplot DEG

Este gráfico permite visualizar de forma clara el número de genes compartidos entre sexos para cada tipo celular, así como aquellos genes que son específicos de hombres o de mujeres.

```

stats_df <- map_df(DEG_comparative, function(x) {
  data.frame(
    celltype = x$celltype,
    n = c(length(x$common), length(x$male_only), length(x$female_only)),
    Categoria = c("Common", "Male_Specific", "Female_Specific"),
    stringsAsFactors = FALSE
  )
})

ggplot(stats_df, aes(x = reorder(celltype, n), y = n, fill = Categoria)) +
  geom_col(position = "dodge") +
  coord_flip() +
  scale_fill_manual(values = c(
    "Common" = "#80b918",
    "Male_Specific" = "#219ebc",
    "Female_Specific" = "#eb5e28"
  )) +
  labs(
    title = "Differential Expression Genes (DEG)",
    x = "Celltype",
    y = "Number of genes",
    fill = "Category"
  ) +
  theme_minimal() +
  theme(panel.grid = element_blank(),
        axis.title.y = element_text(margin = margin(t = 0, r = 20,
                                                    b = 0, l = 0)))

```

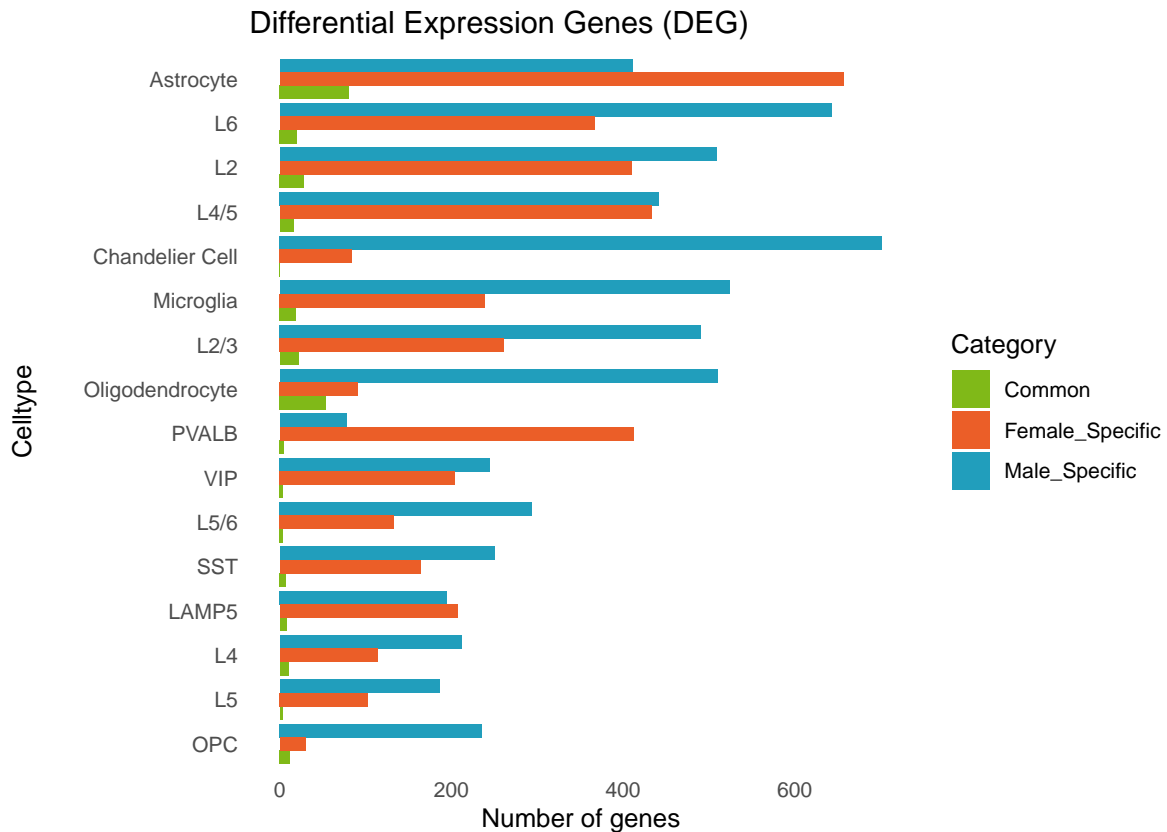


Figure 3: Genes diferencialmente expresados (DEG) por tipo celular y sexo

En general, se observa que los hombres presentan un mayor número de genes diferencialmente expresados específicos en la mayoría de los tipos celulares analizados. En contraste, las mujeres únicamente muestran un mayor número de genes específicos en las neuronas PVALB y en los astrocitos.

Por otro lado, el número de genes comunes entre ambos sexos es reducido en comparación con los genes específicos, lo que sugiere dimorfismo genético asociado a la esquizofrenia.

### 3.2 UpsetPlot

El Upset plot permite observar cuántos genes diferencialmente expresados existen por tipo celular y facilita la identificación de genes únicos y compartidos entre distintos tipos celulares.

```
# Male
genes_m_list <- lapply(DEG_comparative, function(x) rownames(x$male))
```

```

upset_m <- UpSetR::fromList(genes_m_list)

upset_men <- ComplexUpset::upset(
  upset_m,
  intersect = colnames(upset_m),
  name = 'Common Genes',
  width_ratio = 0.2,
  sort_sets = FALSE,
  min_size = 10,
  matrix = intersection_matrix(geom = geom_point(size = 1))
)

# Female
genes_f_list <- lapply(DEG_comparative, function(x) rownames(x$female))
upset_f <- UpSetR::fromList(genes_f_list)

upset_woman <- ComplexUpset::upset(
  upset_f,
  intersect = colnames(upset_f),
  name = 'Common Genes',
  width_ratio = 0.2,
  sort_sets = FALSE,
  min_size = 10,
  matrix = intersection_matrix(geom = geom_point(size = 1))
)

upset_men / upset_woman

```

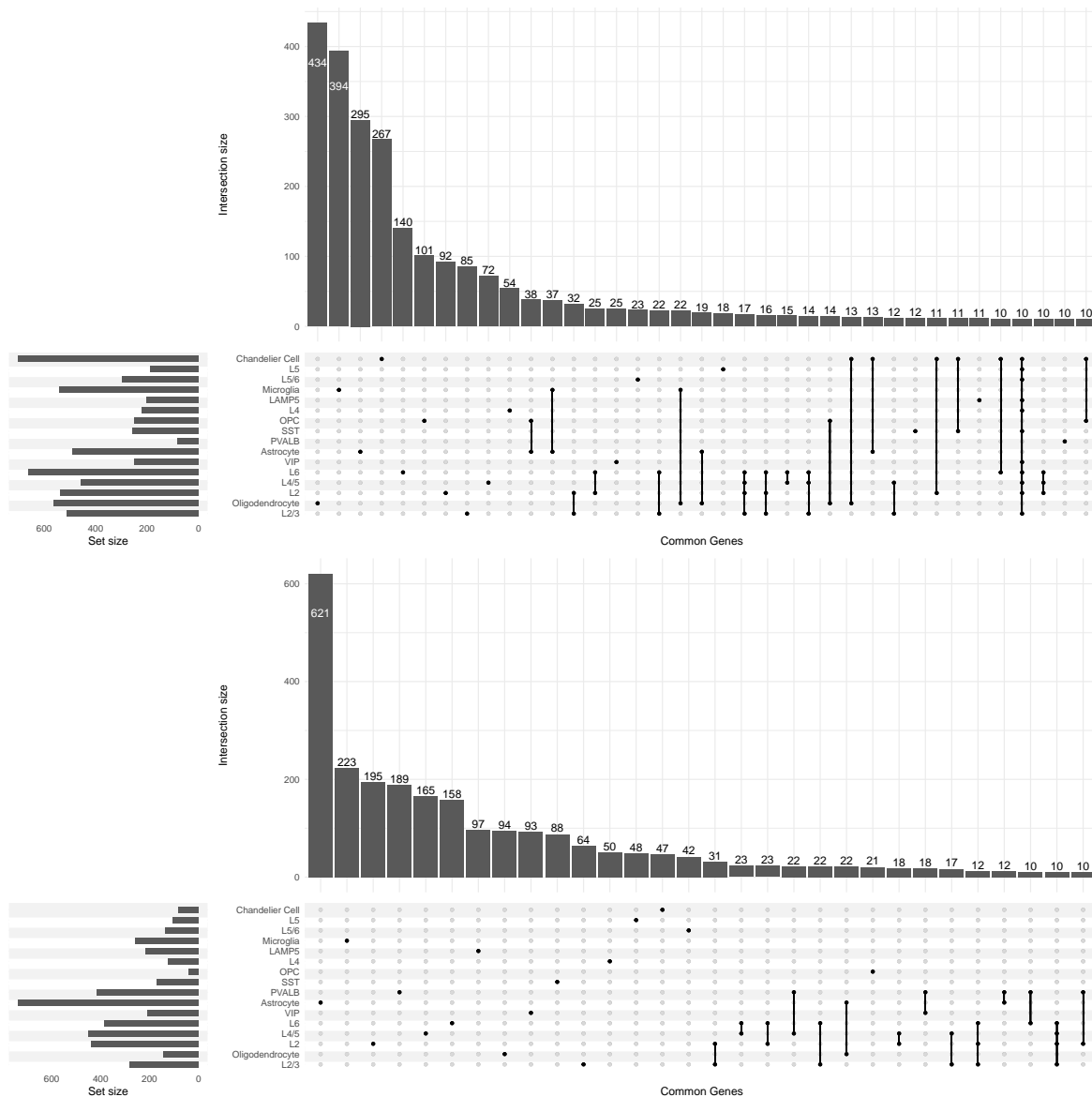


Figure 4: UpsetPlot de cada tipo celular y sexo

Confirma que la cantidad de genes diferencialmente expresados es mayor en hombres que en mujeres, lo que se traduce también en un mayor número de intersecciones entre tipos celulares masculinos.

Las relaciones observadas son coherentes desde el punto de vista biológico ya que se detectan intersecciones relevantes entre capas corticales relacionadas, como L2 y L2/3, que derivan de neuronas excitatorias y comparten características funcionales.

Este gráfico permite estimar el grado de relación de los genes diferencialmente expresados entre distintos tipos celulares. En general, se observa que hay pocas intersecciones con gran número de genes, lo que sugiere la presencia de perfiles transcripcionales altamente específicos y poco compartidos.

### 3.3 Heatmap

El heatmap se construyó exclusivamente a partir de los genes comunes entre ambos sexos, con el objetivo de evaluar si, a pesar de estar presentes en ambos, muestran patrones de expresión diferencial.

```
common_genes <- unique(unlist(lapply(DEG_comparative, function(x) x$common)))

# Matrix
logFC_mat <- matrix(NA, nrow = length(common_genes), ncol = length(DEG_48))
rownames(logFC_mat) <- common_genes
colnames(logFC_mat) <- names(DEG_48)

for (i in seq_along(DEG_48)) {
  df <- DEG_48[[i]]
  genes_presentes <- intersect(common_genes, rownames(df))
  logFC_mat[genes_presentes, i] <- df$avg_log2FC[match(genes_presentes,
                                                    rownames(df))]
}

logFC_mat[is.na(logFC_mat)] <- 0

# Scale
mat_scaled <- t(scale(t(logFC_mat)))
mat_scaled[is.na(mat_scaled)] <- 0
rownames(mat_scaled) <- rownames(logFC_mat)

# Colours
col_fun <- colorRamp2(c(-2, 0, 2), c("#2166ac", "white", "#b2182b"))

# Annotation
cols_interes <- colnames(logFC_mat)
ha_cols <- HeatmapAnnotation(
  Sex = rep(c("M", "F"), length.out = length(cols_interes)),
  col = list(
    Sex = c(M = "#219ebc", F = "#eb5e28")
  )
)
```

```

    )
)

Heatmap(
  mat_scaled,
  name = "Z-score",
  col = col_fun,
  top_annotation = ha_cols,
  cluster_columns = FALSE,
  cluster_rows = TRUE,
  show_row_names = TRUE,
  row_names_gp = gpar(fontsize = 6, fontface = "italic"),
  column_names_gp = gpar(fontsize = 9),
  column_title = "Celltype Common Genes",
  rect_gp = gpar(col = "white", lwd = 0.5)
)

```

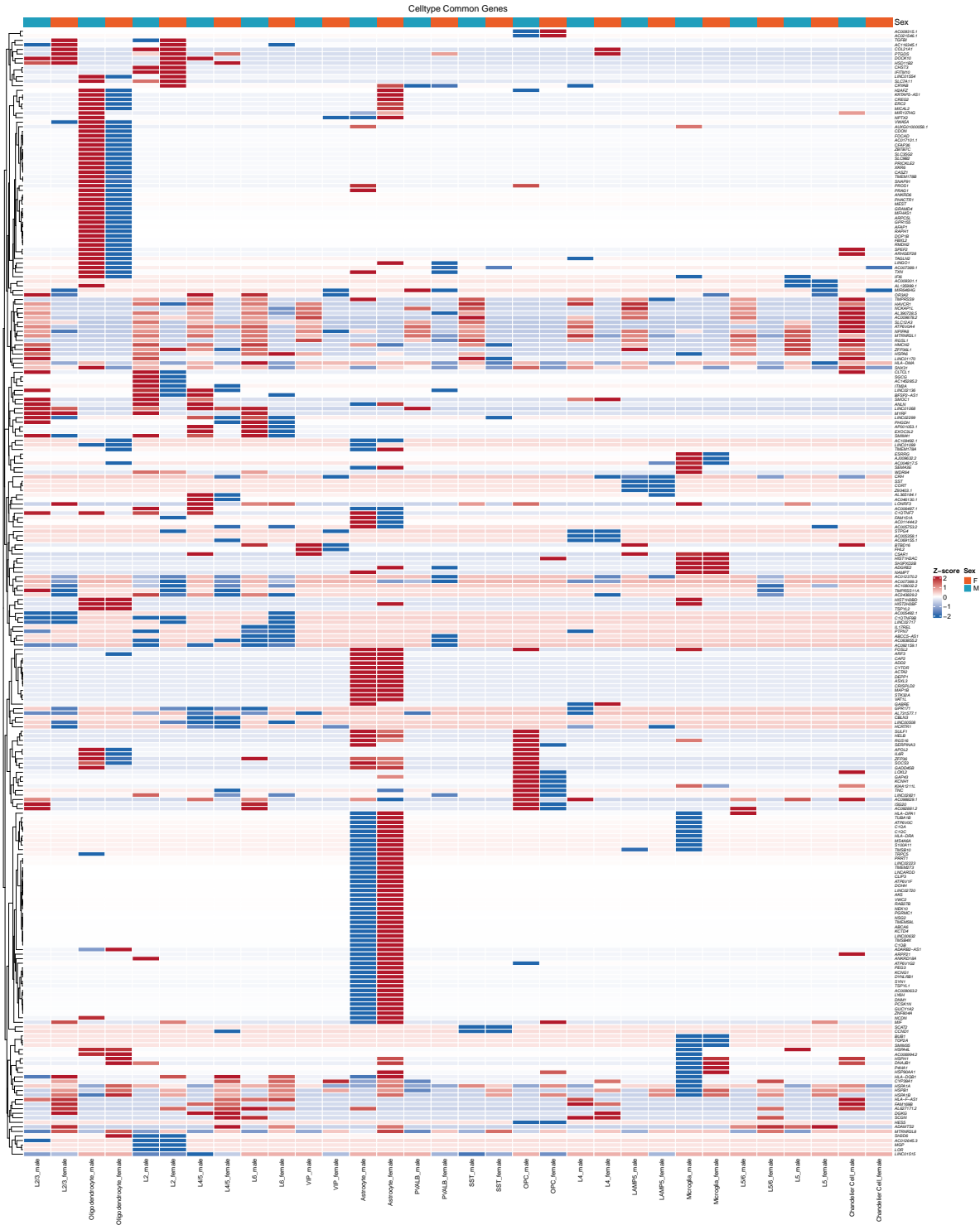


Figure 5: Heatmap de los genes en común entre sexos

Las mayores diferencias de expresión génica se observaron principalmente en astrocitos y oligodendrocitos, donde se aprecia un patrón claramente diferenciado entre sexos.

En concreto, en oligodendrocitos la mayoría de los genes se encuentran *up-regulated* en hombres, mientras que en astrocitos predominan los *down-regulated*. Este patrón se invierte en mujeres, lo que indica una regulación opuesta dependiente del sexo en estos dos tipos celulares.

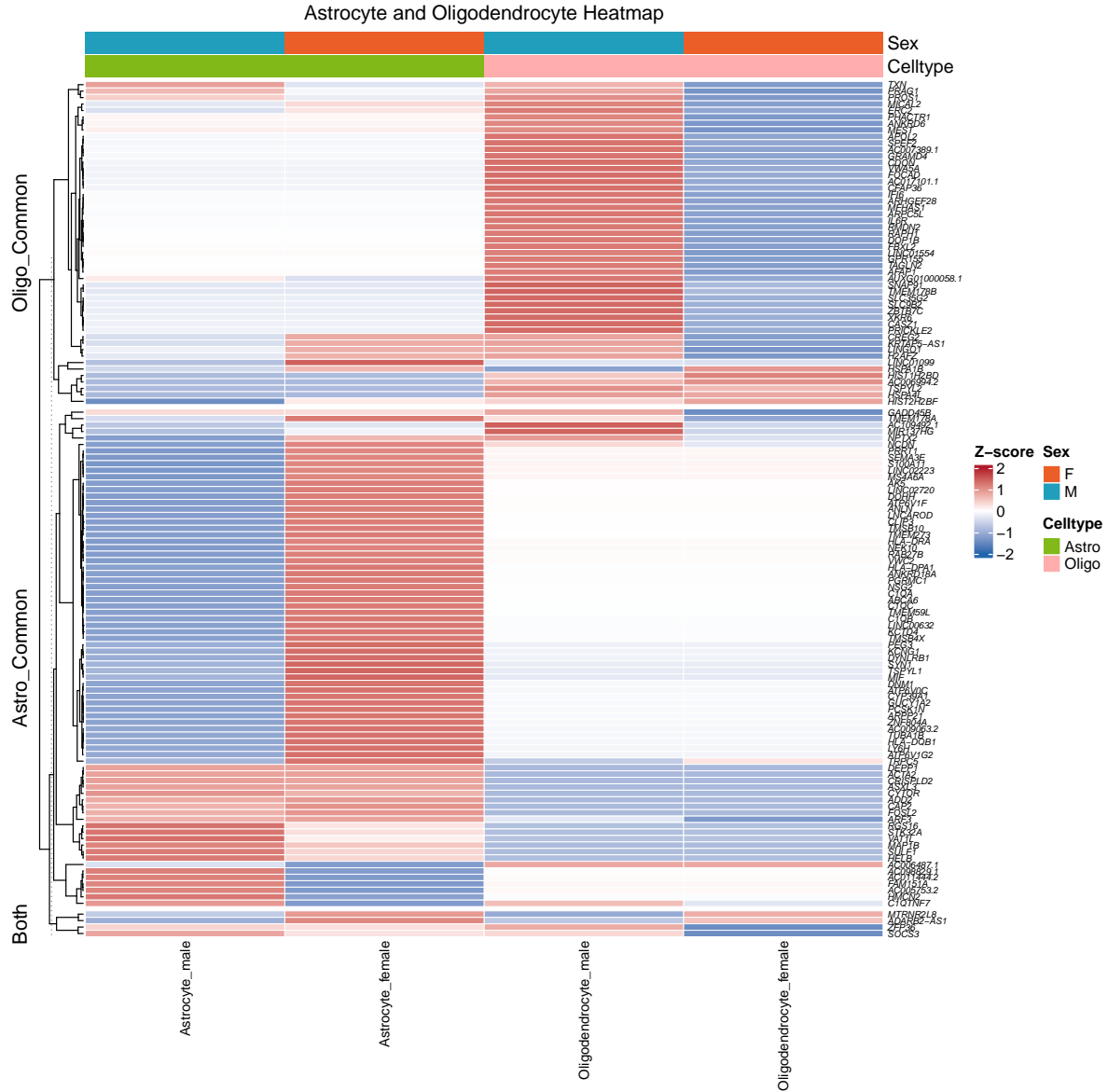


Figure 6: Heatmap de Astrocitos y Oligodendrocitos

### 3.4 Correlation Plot

Este análisis muestra cómo se distribuyen los genes diferencialmente expresados para cada tipo celular al comparar hombres y mujeres. Los genes situados lejos de la diagonal principal indican una mayor diferencia entre sexos, mientras que aquellos próximos a la diagonal representan genes con efectos similares en ambos sexos.

```
correlation_list <- lapply(DEG_comparative, function(x) {

  df_m <- x$male
  df_f <- x$female

  # Data Union
  df_cmp <- inner_join(
    df_m %>% rownames_to_column("gene"),
    df_f %>% rownames_to_column("gene"),
    by = "gene",
    suffix = c("_male", "_female")
  )

  # Correlation
  r_val <- round(cor(df_cmp$avg_log2FC_male, df_cmp$avg_log2FC_female), 2)

  c <- ggplot(df_cmp, aes(x = avg_log2FC_male, y = avg_log2FC_female)) +
    geom_point(
      alpha = 0.5,
      color = "midnightblue",
      size = 2
    ) +
    geom_abline(
      slope = 1,
      intercept = 0,
      linetype = "dashed",
      color = "firebrick"
    ) +
    geom_vline(xintercept = 0, alpha = 0.25) +
    geom_hline(yintercept = 0, alpha = 0.25) +
    scale_x_continuous(limits = c(-4, 4)) +
    scale_y_continuous(limits = c(-4, 4)) +
    coord_fixed() +
    labs(
      title = x$celltype,
```

```

    subtitle = paste("Common Genes:", nrow(df_cmp)),
    caption = paste("r:", r_val),
    x = "Log2FC Male",
    y = "Log2FC Female"
  ) +
  theme_minimal(base_size = 9) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 9, face = "bold"),
    plot.subtitle = element_text(size = 7),
    axis.title = element_text(size = 8),
    axis.text = element_text(size = 7))
  return(c)
})

correlation <- wrap_plots(correlation_list, ncol = 2, nrow = 8)
ggsave("Correlation_plots.pdf", correlation, width = 5, height = 20,
       limitsize = FALSE)

```

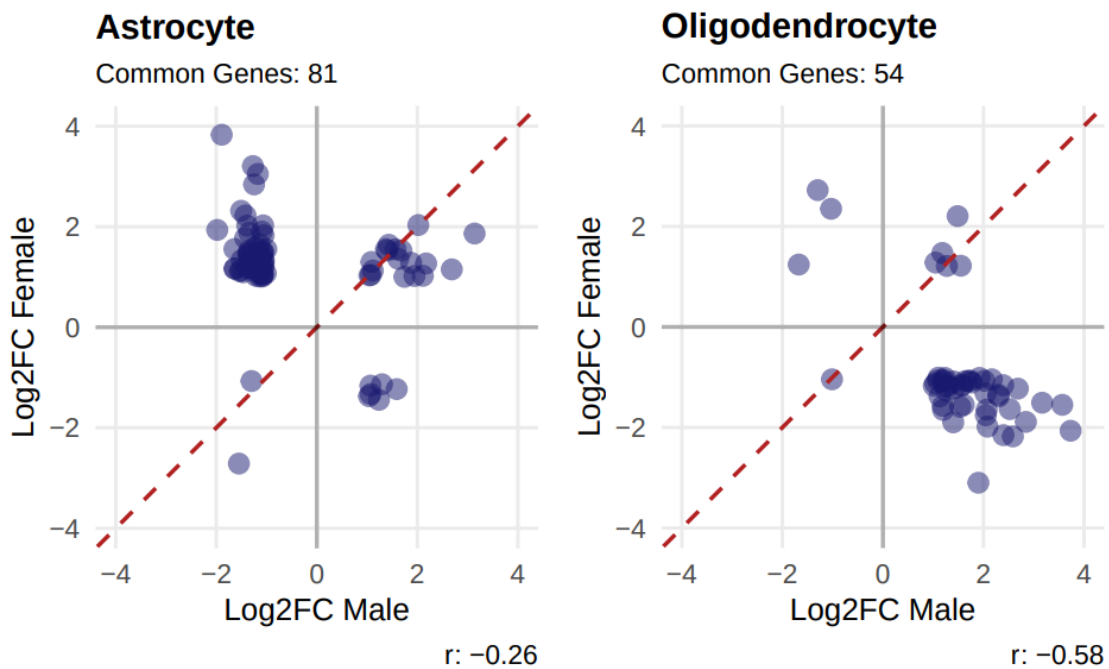


Figure 7: Gráficos de Correlación de Astrocitos y Oligodendrocitos

Estos gráficos refuerzan los observados en el *heatmap*, mostrando que los genes diferencialmente expresados en oligodendrocitos y astrocitos presentan patrones de expresión opuestos entre sexos.

Esta correlación negativa sugiere la existencia de mecanismos reguladores divergentes entre ambos tipos celulares, posiblemente vinculados a diferencias en la función glial asociadas al sexo en este contexto.

### 3.5 Volcano Plot

Los volcano plots permiten visualizar los genes más destacados para cada sexo, resaltando aquellos con mayores cambios de expresión y mayor significación estadística.

```
volcanos_list <- list()
for (celltypes in names(DEG_48)) {
  p <- EnhancedVolcano(DEG_48[[celltypes]],
    lab = rownames(DEG_48[[celltypes]]),
    x = 'avg_log2FC',
    y = 'p_val_adj',
    title = celltypes,
    drawConnectors = TRUE,
    pCutoff = 0.001,
    FCcutoff = 5,
    ylim = c(0, 150),
    xlim = c(-10,10),
    subtitle = NULL,
    caption = NULL,
    legendPosition = 'none'
  )

  volcanos_list[[celltypes]] <- p
}

volcano_union <- wrap_plots(volcanos_list, ncol = 2, nrow = 16)
ggsave("Volcano_plots.pdf", volcano_union, width = 12, height = 60,
  limitsize = FALSE)
```

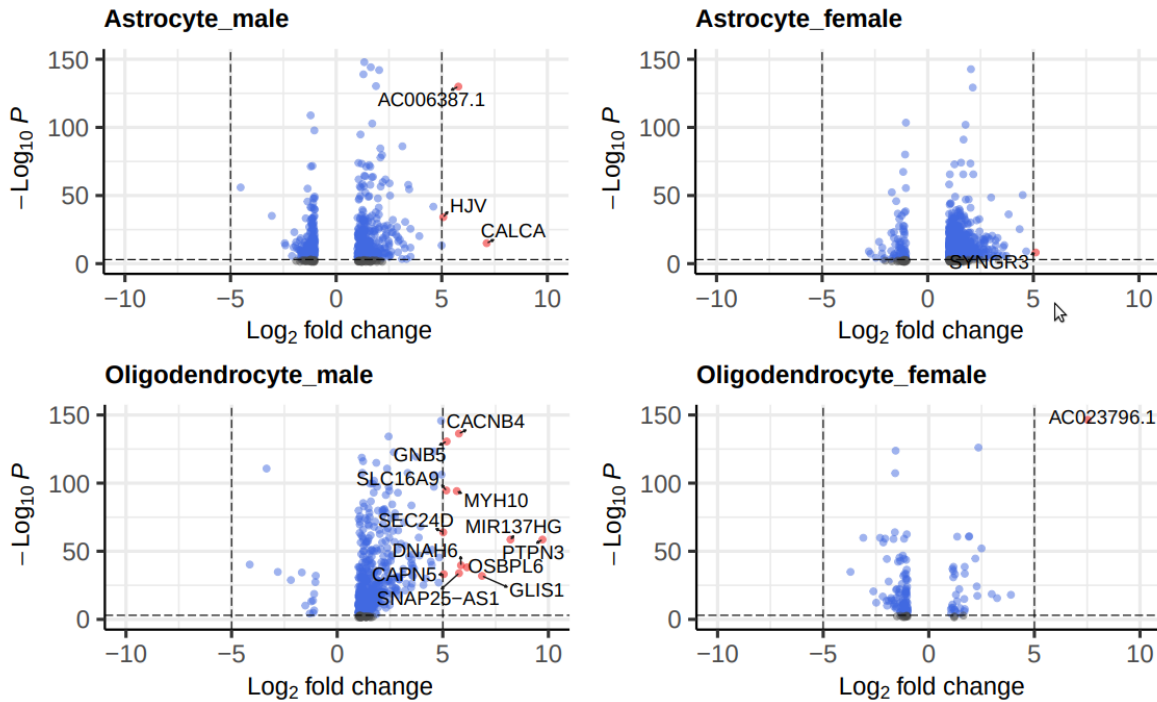


Figure 8: Volcano plot de Astrocitos y Oligodendrocitos

Tal y como se observó en el *barplot*, los hombres presentan un mayor número de genes diferencialmente expresados en comparación con las mujeres. Este patrón se mantiene incluso al aplicar umbrales más estrictos de cambio de expresión ( $FC_{cutoff}$  elevado), lo que indica una respuesta transcripcional más intensa o extensa en hombres.

En contraste, en mujeres el número de genes diferencialmente expresados es considerablemente menor, incluso con criterios menos restrictivos, lo que refuerza la idea de una respuesta más limitada o específica.

### 3.6 Butterfly Plot

Los resultados del enriquecimiento funcional se representaron mediante diagramas tipo butterfly plot, permitiendo comparar de forma directa los procesos biológicos enriquecidos en hombres y mujeres para cada tipo celular.

```
# Defining dataframe with genes up and down-regulated per sex.

butterfly_data <- function(gsea_results, celltype_sel, top_n = 10) {
```

```

gsea_results %>%
  filter(celltype == celltype_sel) %>%
  group_by(sex, direction) %>%
  slice_max(order_by = abs(NES), n = top_n) %>%
  ungroup() %>%
  mutate(
    sex = factor(sex, levels = c("female", "male")),
    direction = factor(direction, levels = c("Down", "Up"))
  )
}

# Defining Butterfly Plot

butterfly_plot <- function(df, selected_sex, show_y_axis = TRUE) {
  plot_data <- df %>% filter(sex == selected_sex)

  p <- ggplot(
    plot_data,
    aes(x = NES, y = reorder_within(Description, NES, sex),
        fill = direction)) +
    geom_col(width = 0.8) +
    geom_vline(xintercept = 0, color = "black") +
    scale_fill_manual(
      values = c("Down" = "#4575b4", "Up" = "#d73027"),
      labels = c("Down-regulated", "Up-regulated")
    ) +
    scale_y_reordered() +
    scale_x_continuous(labels = function(x) abs(x)) +
    labs(
      title = tools::toTitleCase(selected_sex),
      x = "NES",
      y = NULL
    ) +
    theme_minimal(base_size = 13) +
    theme(
      panel.grid = element_blank(),
      panel.border = element_rect(fill = NA, color = "gray80"),
      axis.text.y = if(show_y_axis) element_text(size = 10)
      else element_blank(),
      legend.position = "none")
}

```

```

    return(p)
  }

  pdf("Butterfly_Plots.pdf", width = 30, height = 8)

  for (ct in celltype) {
    df_plot <- butterfly_data(gsea_results, ct)

    if(nrow(df_plot) > 0) {

      tryCatch({

        p_female <- butterfly_plot(df_plot, "female", show_y_axis = TRUE)
        p_male    <- butterfly_plot(df_plot, "male", show_y_axis = TRUE)

        final_plot <- (p_female | p_male) +
          plot_layout(guides = "collect") +
          plot_annotation(
            title = ct,
            theme = theme(plot.title = element_text(size = 14, face = "bold",
                                                    hjust = 0.5))) &
            theme(legend.position = "bottom")

        print(final_plot)

      }, error = function(e) {
        message(paste("Error en", ct, ":", e$message))
      })

    } else {
      message(paste("Saltando", ct, "(sin datos suficientes)"))
    }
  }

  dev.off()

```

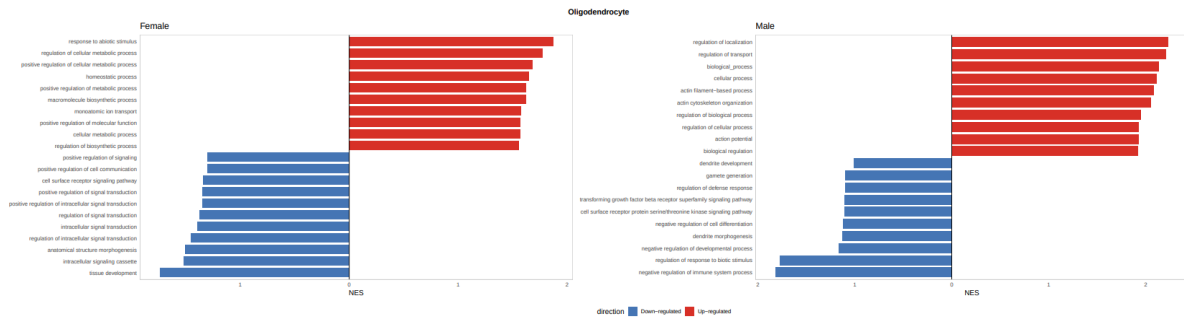


Figure 9: Enrichimiento Funcional de los Oligodendrocitos

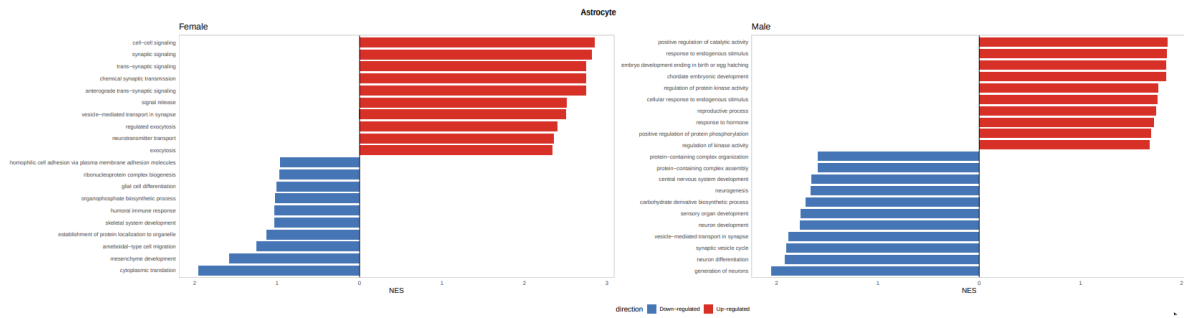


Figure 10: Enrichimiento Funcional Astrocitos

## 4 Discusión

Uno de los objetivos principales de este trabajo fue evaluar la reproducibilidad de los resultados descritos en el estudio original (*Single-Cell Transcriptional Profiling Reveals Cell Type-Specific Sex-Dependent Molecular Patterns of Schizophrenia*). Sin embargo, a pesar de haber aplicado parámetros metodológicos similares, no se obtuvieron exactamente los mismos resultados. No obstante, se propone una reinterpretación más detallada de los mismos, aportando una visión complementaria a la previamente descrita.

A diferencia del artículo original, en el que mediante UMAP se identificaron principalmente los ocho tipos celulares mayoritarios, en este estudio se lograron clasificar hasta 16 tipos celulares, incluyendo distintas capas de neuronas excitatorias que ya estaban presentes en las columnas de las matrices de conteo originales. Esta mayor resolución celular fue posible gracias a un enfoque más detallado de clusterización y a la posterior anotación utilizando la base de datos CellKB.

El análisis de expresión diferencial se realizó comparando casos y controles de forma independiente para cada tipo celular y sexo. Aunque se observaron diferencias claras en la distribución de genes diferencialmente expresados (DEGs), estos patrones no coinciden completamente con

los descritos. En el trabajo original, los autores reportaron un mayor número de DEGs en células gliales en hombres y un mayor número de DEGs en neuronas en mujeres. En cambio, en este estudio las mujeres solo mostraron un mayor número de DEGs en astrocitos e interneuronas PVALB, mientras que en el resto de tipos celulares los hombres presentaron, de forma general, una mayor carga.

La distribución de los DEGs representada mediante Upsetplots muestra que los hombres presentan un mayor número de genes compartidos entre distintos tipos celulares, lo cual es consistente con el hecho de que en este grupo se detecta un mayor número de DEGs. No obstante, no se observaron diferencias significativas en los patrones de relación entre tipos celulares entre hombres y mujeres. En ambos sexos, los tipos celulares tienden a agruparse por similitud biológica, de modo que las neuronas excitatorias de distintas capas comparten más genes entre sí, mientras que las células gliales muestran un mayor solapamiento dentro de su propio linaje.

Este patrón también se refleja en los volcano plots, donde, a pesar de aplicar un umbral de  $\log_2FC$  muy restrictivo, se detecta un número mayor de genes significativos de lo esperado en hombres, especialmente en las células de Chandelier, astrocitos y oligodendrocitos. En mujeres, en cambio, la cantidad de DEGs es considerablemente menor, incluso con criterios menos estrictos.

Es importante destacar que, mientras que el estudio original utilizó el método de corrección de Bonferroni, en este trabajo se optó por aplicar el ajuste de Benjamini–Hochberg, considerado más adecuado para este tipo de contrastes. El umbral de  $\log_2FC$  se mantuvo constante.

Además, mientras que en el estudio de referencia se definieron los genes dimórficos como aquellos comunes a ambos sexos pero con direcciones opuestas de cambio, en este estudio se adoptó una aproximación diferente: primero se identificaron los genes comunes y posteriormente se clasificaron como específicos de sexo cuando su expresión diferencial solo se detectaba en uno de los grupos. En la mayoría de los tipos celulares, menos del 20 % de los DEGs eran compartidos entre sexos, lo que concuerda con la fuerte especificidad observada.

A pesar de que el número de genes compartidos es reducido, el *heatmap* revela que estos genes presentan niveles de expresión claramente distintos entre sexos, especialmente en astrocitos y oligodendrocitos. En particular, se observa que en oligodendrocitos los hombres presentan una mayor proporción de genes *up-regulated* mientras que en astrocitos se invierte. Las mujeres muestran un patrón totalmente contrario al del otro sexo, mostrando una regulación opuesta entre ambos tipos celulares. Este comportamiento inverso también queda reflejado en los correlation plots, reforzando la hipótesis de una regulación transcripcional diferencial dependiente del sexo en células gliales.

El análisis de enriquecimiento funcional se llevó a cabo utilizando la base de datos *Gene Ontology* (GO), separando los resultados por tipo celular y sexo. Para ello, se aplicó un enfoque basado en *Gene Set Enrichment Analysis* (GSEA), que permite integrar los niveles de expresión génica con la actividad funcional de los genes asociados a esquizofrenia. Se

seleccionaron los diez términos GO más relevantes, tanto *up-regulated* como *down-regulated*, ordenados según el *Normalized Enrichment Score* (NES).

Dado el papel destacado observado en los análisis previos, se puso especial énfasis en oligodendrocitos y astrocitos. En oligodendrocitos femeninos, los procesos *up-regulated* se asocian principalmente con el metabolismo celular y la homeostasis, mientras que los procesos *down-regulated* están relacionados con la señalización celular e intercelular. En hombres, los procesos *up-regulated* se vinculan con la localización y transporte intracelular, especialmente relacionados con el citoesqueleto y los filamentos de actina, mientras que los procesos *down-regulated* incluyen la morfogénesis, diferenciación y desarrollo celular, así como procesos relacionados con dendritas.

En el caso de los astrocitos, las mujeres muestran una regulación positiva de procesos relacionados con la señalización célula-célula, la sinapsis y los neurotransmisores, mientras que los procesos *down-regulated* incluyen adhesión celular, diferenciación, biogénesis y migración celular. En hombres, los procesos *up-regulated* se asocian principalmente con actividad catalítica, respuesta a estímulos endógenos, procesos reproductivos y respuesta hormonal, mientras que los procesos *down-regulated* incluyen neurogénesis, neurodesarrollo, diferenciación neuronal y sinapsis.

En conjunto, estos resultados sugieren que el análisis a mayor resolución celular revela una heterogeneidad transcripcional dependiente del sexo más compleja de lo previamente descrito, especialmente en células gliales. Las diferencias observadas entre este estudio y el trabajo original no contradicen necesariamente sus conclusiones, sino que amplían y refinan los hallazgos previos, poniendo de manifiesto la importancia de considerar subtipos celulares y enfoques analíticos más detallados en el estudio de la esquizofrenia.

## 5 Conclusión

En este trabajo se ha llevado a cabo la replicación parcial de un estudio transcriptómico de esquizofrenia evaluando tanto la reproducibilidad como la interpretación biológica de los resultados. Aunque no se obtuvieron las mismas soluciones, el enfoque adoptado permitió una caracterización celular más detallada y puso de manifiesto patrones consistentes de dimorfismo sexual en la regulación génica. En particular, se observó que los hombres presentan una mayor carga global de genes diferencialmente expresados, mientras que en las mujeres las alteraciones se concentran en tipos celulares específicos como astrocitos e interneuronas PVALB.

El análisis de expresión diferencial, correlación y enriquecimiento funcional señala a las células gliales, especialmente astrocitos y oligodendrocitos, como elementos clave en las diferencias moleculares dependientes del sexo en esquizofrenia. La regulación opuesta observada en estos tipos celulares sugiere la implicación de mecanismos biológicos divergentes que podrían estar relacionados con procesos hormonales, metabólicos y de señalización celular. En conjunto, estos resultados subrayan la importancia de incorporar la variable sexo en estudios transcriptómicos

y apoyan el uso de estrategias de single-cell RNA-seq como herramientas fundamentales para avanzar hacia una comprensión más precisa y personalizada de la enfermedades complejas como la esquizofrenia.

## 6 Bibliografía

- [1]Amiri, E., Baghaei, R., Habibzadeh, H., & Ebrahimi, H. (2025). The role of personal traits on the dignity of individuals living with schizophrenia: a qualitative study. *BMC Psychiatry*, 25, 1000. <https://doi.org/10.1186/s12888-025-07477-w>
- [2]Anwar, A., Mustafa, A. M., Abdou, K., Rabie, M. A., El-Shiekh, R. A., & El-Dessouki, A. M. (2025). A comprehensive review on schizophrenia: epidemiology, pathogenesis, diagnosis, conventional treatments, and proposed natural compounds used for management. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 398(11), 15231-15255. <https://doi.org/10.1007/s00210-025-04351-0>
- [3]Batiuk, M. Y., Tyler, T., Dragicevic, K., Mei, S., Rydbirk, R., Petukhov, V., Deviatiiarov, R., Sedmak, D., Frank, E., Feher, V., Habek, N., Hu, Q., Igolkina, A., Roszik, L., Pfisterer, U., Garcia-Gonzalez, D., Petanjek, Z., Adorjan, I., Kharchenko, P. V., & Khodosevich, K. (2022). Upper cortical layer-driven network impairment in schizophrenia. *Science Advances*, 8(41), eabn8367. <https://doi.org/10.1126/sciadv.abn8367>
- [4]Faden, J., & Citrome, L. (2018). Resistance is not futile: treatment-refractory schizophrenia – overview, evaluation and treatment. *Expert Opinion on Pharmacotherapy*, 20(1), 11–24. <https://doi.org/10.1080/14656566.2018.1543409>
- [5]Institute for Health Metrics and Evaluation (IHME). (2025). GBD Results. *VizHub*. <https://vizhub.healthdata.org/gbd-results/> (Accedido el 19/11/2025).
- [6]Luvsannyam, E., Jain, M. S., Pormento, M. K. L., Siddiqui, H., Balagtas, A. R. A., Emuze, B. O., & Poprawski, T. (2022). Neurobiology of Schizophrenia: A Comprehensive Review. *Cureus*, 14(4), e23959. <https://doi.org/10.7759/cureus.23959>
- [7]Orrico-Sanchez, A., López-Lacort, M., Muñoz-Quiles, C., Sanfélix-Gimeno, G., & Díez-Domingo, J. (2020). Epidemiology of schizophrenia and its management over an 8-years period using real-world data in Spain. *BMC Psychiatry*, 20, 149. <https://doi.org/10.1186/s12888-020-02538-8>
- [8]Ruzicka, W. B., Mohammadi, S., Fullard, J. F., Davila-Velderrain, J., Subburaju, S., Tso, D. R., Hourihan, M., Jiang, S., Lee, H., Bendl, J., Voloudakis, G., Haroutunian, V., Hoffman, G. E., Roussos, P., Kellis, M., Akbarian, S., Abyzov, A., Ahituv, N., Arasappan, D., ... Zintel, T. M. (2024). Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science*, 384(6698), eadg5136. <https://doi.org/10.1126/science.adg5136>

[9]Zhou, R., Zhang, T., & Sun, B. (2025). Single-Cell Transcriptional Profiling Reveals Cell Type-Specific Sex-Dependent Molecular Patterns of Schizophrenia. *International Journal of Molecular Sciences*, 26(5), 2227. <https://doi.org/10.3390/ijms26052227>