Genética de las disparidades de género en la enfermedad de Alzheimer

Olga Reyes Malia

Máster de Bioinformática y Bioestadística

Biología del desarrollo, cáncer, biología molecular y farmacología

Nombre Consultor/a

Ivette Olivares Castiñeira

Nombre Profesor/a responsable de la asignatura

Marc Maceira Duch

05 de enero de 2021

FICHA DEL TRABAJO FINAL

Título del trabajo	*Genética de las disparidades de género en la enfermedad de Alzheimer*
Nombre del autor	*Olga Reyes Malia*
Nombre del consultor/a	*Ivette Olivares Castiñeira*
Nombre del PRA	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa)	05/2021
Titulación	*Máster de Bioinformática y Bioestadística*
Área del Trabajo Final	*Biología del desarrollo, cáncer, biología molecular y farmacología*
Idioma del trabajo	*Español*
Palabras clave	*Enfermedad de Alzheimer, disparidades de género, expresión génica.*

Resumen del Trabajo

La enfermedad de Alzheimer es la forma más común de demencia que existe. Algunos estudios sugieren que el género de las personas puede ser un factor de riesgo para su prevalencia y desarrollo. Con este proyecto se pretende valorar si realmente el sexo puede influir en la clínica o la genética de los pacientes. Para ello se realiza un análisis estadístico donde se valoran las diferencias clínicas significativas entre hombres y mujeres afectados por la enfermedad. A continuación, se realiza un análisis de expresión diferencial de dos zonas cerebrales también incluidas en el análisis estadístico: hipocampo y corteza entorrinal. Se valora en ambas zonas cerebrales, qué genes se expresan de forma diferencial en hombres y mujeres. Por último, se realiza un análisis de *pathways* sobre los genes destacados y, finalmente, se lleva a cabo un análisis comparativo con el objetivo de relacionar los resultados de los dos análisis anteriormente realizados.

Los resultados obtenidos indican que las mujeres con Alzheimer tienen un volumen cerebral general inferior al de los hombres, especialmente la zona del hipocampo, que es la única zona que se confirma en ambos análisis. En esta zona cerebral se detectan varios genes diferencialmente expresados, siendo los más significativos *Sp1, MBLN1* y *ALCAM.* Adicionalmente también se observa una disminución del volumen del hipocampo a medida que aumenta la edad de los pacientes.

Como conclusión, se identifican diferencias clínicas entre sexos en pacientes con Alzheimer y también genes diferencialmente expresados en el hipocampo. Indicando todo esto, que existen disparidades de género en la enfermedad de Alzheimer.

Abstract:

Alzheimer's disease is the most common existing form of dementia. Some research studies suggest that people's gender may be a risk factor for its prevalence and development. This project aims to assess whether sex can influence the clinic or the genetics of patients. For this, a statistical analysis is carried out where the significant clinical differences between men and women affected by the disease are assessed. This is followed by differential expression analysis of two brain areas also included in the statistical analysis: hippocampus and entorhinal cortex. It is assessed in both brain areas, which genes are expressed differentially in men and women. Pathway analysis is executed on the identified genes. Finally, a comparative analysis is carried out to relate the results of the two previously performed analyzes.

The results obtained indicate that women with Alzheimer's have a lower general brain volume than men, especially the hippocampus area. This is the only area with confirmed gender differences in both analyzes and where several differentially expressed genes are detected, the most significant being *Sp1*, *MBLN1*, and *ALCAM*. Additionally, a decrease in the hippocampus volume is observed as the patients' age increases.

In conclusion, clinical differences between sexes in Alzheimer's patients, and genes differentially expressed in the hippocampus are identified. All this indicating, that there are gender disparities in Alzheimer's disease.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1. Contexto y justificación del trabajo

1.1.1. Estado del arte

La demencia es una de las principales causas de discapacidad entre las personas mayores en el mundo entero, tal y como indica la Organización Mundial de la Salud (OMS), por lo que representa un síndrome de interés social y científico desde hace décadas. El deterioro de la función cognitiva es la característica principal de la demencia, la cual da lugar a la afectación de funciones cerebrales básicas [1].

La forma más común de demencia es la enfermedad de Alzheimer, que representa entre un 60% y un 70% de los casos [1]. Existen dos tipos de Alzheimer según la edad de aparición (temprana o tardía). El Alzheimer de aparición tardía es el más común y los primeros síntomas aparecen a partir de los 60 años, aunque el riesgo a padecerla aumenta con la edad [2]. En consecuencia, la edad es uno de los factores de riesgo más importantes para esta enfermedad, además de la genética. No obstante, se conocen otros factores que pueden también influir en el riesgo a padecerla, como por ejemplo las enfermedades cardiovasculares, el tiempo de educación o lesiones traumáticas cerebrales, entre otros[3].

El Alzheimer de aparición tardía surge de forma esporádica, es decir, no tiene una agregación familiar. Sin embargo, esta forma de la enfermedad es una de las patologías humanas multifactoriales con el nivel más alto de heredabilidad [4]. Desde hace varias décadas, se conoce que el gen *apolipoproteína E* (*APOE*) está asociado al riesgo a desarrollar Alzheimer, en concreto el alelo ε4. Por consiguiente, este es el gen con un mayor impacto en el riesgo de padecer la forma tardía de Alzheimer [3,4]. La proteína APOE ε4 actúa en el transporte, la regulación de la producción y la distribución de partículas lipídicas a través del plasma y líquidos intersticiales entre órganos y tejidos [5]. A partir de los años 2000, gracias a los avances en investigación, se comenzaron a descubrir nuevos posibles factores de riesgo genéticos, así como, *CLU, CR1 y PICALM* [4].

Según diversos estudios, el género de las personas puede ser otro factor de riesgo para la enfermedad de Alzheimer, debido a que parece tener un papel importante en la heterogeneidad de muchos aspectos de la enfermedad, como la prevalencia, las manifestaciones clínicas, la patología o el pronóstico [6,7].

Hasta el momento, uno de los motivos que se consideran para explicar estas disparidades de género, es el hecho de que las mujeres tienen una mayor esperanza de vida que los hombres. Se cree que, aunque esta puede ser una de las causas de las diferencias observadas, es probable que no sea la única [8].

1.1.2. Descripción general

El trabajo final se centra en el estudio de la enfermedad de Alzheimer mediante el uso de herramientas y técnicas bioinformáticas. Concretamente, se enfoca en valorar las disparidades que existen entre pacientes de diferente sexo.

Por una parte, se realiza un estudio estadístico para evaluar las diferencias clínicas entre pacientes de diferente género. Con esto se pretende comparar aspectos como la sintomatología o la edad, para comprobar si se observa alguna tendencia o correlación entre el sexo y alguno de los datos clínicos estudiados.

Por otra parte, se lleva a cabo un estudio de expresión génica con el objetivo de encontrar genes cuya expresión sea específica de sexo para la enfermedad. Se analiza en qué región genómica se encuentran esos genes y en qué vías celulares se ven involucrados.

Por último, se valora si existe una relación entre los resultados de ambos estudios que permita explicar y entender mejor las causas de las disparidades que se dan entre géneros.

Con este trabajo se pretende generar nueva información que sea de utilidad para comprender mejor la enfermedad de Alzheimer y poder avanzar en su investigación con un enfoque distinto al convencional. De esta forma, se aspira a poder dar un paso en el descubrimiento de nuevos genes candidatos con el fin de abrir puertas hacia nuevas investigaciones. La finalidad de esto es avanzar en la búsqueda de tratamientos eficientes y diagnósticos tempranos más personalizados a los pacientes que la padecen.

1.1.3. Justificación del trabajo

De la misma manera que el desarrollo biológico de ambos sexos es diferente, es lógico pensar que las enfermedades puedan adoptar distintos mecanismos en función del género de la persona que lo padece. A raíz de esto, surge la necesidad de adoptar nuevas estrategias de estudio que permitan tener en cuenta estas diferencias. Existen evidencias de numerosas enfermedades y patologías que actúan y afectan de forma diferente en hombres y mujeres [9]. El Alzheimer podría ser un ejemplo más de este tipo de patologías y, tenerlo en cuenta a la hora de investigar esta forma de demencia, permitiría descubrir nuevos diagnósticos y terapias que se adecúen a todos los pacientes, así como facilitar y llevar la comprensión de la enfermedad a otro nivel.

La mayor parte de los avances hechos en la investigación genética del Alzheimer se debe a la realización de estudios de asociación de genoma completo (GWASs) o de expresión, con los que se analizan gran cantidad de genes. A pesar de las evidencias sobre las asociaciones sexo-específicas en ciertos rasgos humanos, estos tipos de estudios se suelen realizar combinando datos de ambos géneros, sin diferenciarlos.

Actualmente comienzan a encontrarse variantes genéticas y asociaciones específicas de sexo en esta enfermedad llevadas a cabo con GWAS sexo-específicos[10], así como

diferencias en la expresión génica de ambos géneros[11]. Aun así, es necesaria mucha más investigación de este tipo para conseguir resultados relevantes.

Este enfoque diferente se hace especialmente necesario en el Alzheimer, debido a que es una enfermedad que lleva muchísimo tiempo investigándose sin obtener resultados satisfactorios que permitan tratarla de forma eficiente, además de ser tan común y devastadora, tanto para los pacientes como para sus familiares.

En conclusión, se ha escogido esta área debido a la importancia que tiene el Alzheimer y lo presente que está en nuestra sociedad y en la necesidad que existe de nuevos hallazgos que lleven a la comprensión total de la enfermedad y al desarrollo de terapias y diagnósticos eficientes. Se ha decidido enfocar de esta manera el trabajo a causa de la falta de resultados relevantes producto de estudios convencionales. Nuevas formas de investigación son necesarias, así como estudios orientados desde distintos puntos de vista, con tal de replantear nuevas opciones y abrir fronteras a la investigación del Alzheimer.

1.2. Objetivos

Los objetivos del trabajo son los siguientes. El primer nivel de clasificación corresponde a los objetivos generales mientras que el segundo nivel corresponde a los objetivos específicos.

I. Comprobar estadísticamente las **diferencias clínicas** entre pacientes de Alzheimer de ambos géneros.
    i. Identificar diferencias significativas entre sexos, mediante un análisis estadístico.
    ii. Interpretar los resultados obtenidos y definir posibles correlaciones entre los datos.
    iii. Contrastar los resultados obtenidos con datos conocidos de otros estudios.

II. Detectar **genes de expresión diferencial** específicos de sexo en el Alzheimer.
    i. Detectar genes que se expresen diferencialmente en ambos sexos mediante un estudio de expresión génica.
    ii. Determinar en qué vías están involucradas esos genes mediante análisis de *pathways*.
    iii. Identificar en qué región genómica se encuentran los genes identificados.

III. Evaluar posibles **relaciones entre los resultados** obtenidos en ambos análisis previos. Es decir, comprobar si los resultados del estudio estadístico pueden asociarse de alguna forma con los del estudio de expresión o viceversa.

1.3. Enfoque y método seguido

La metodología principal seguida, ha consistido primero, en valorar si existen diferencias entre géneros realizando comparaciones de los datos clínicos en ambos grupos. Para ello, se ha realizado un análisis estadístico descriptivo que ha permitido valorar cuáles son los datos observados que más varían. Este primer análisis ha proporcionado información

clave para visualizar cómo se comportan los datos dentro de cada grupo y si estos se diferencian en algún aspecto en concreto.

Una vez identificadas las variables que se comportan de forma distinta en cada grupo, se ha llevado a cabo un proceso estadístico con el objetivo de estimar las relaciones entre las variables y observar cómo se correlacionan. Para hacerlo, se han realizado una serie de contrastes de hipótesis y análisis de regresión lineal.

Todo el análisis estadístico se ha realizado con R, ya que dispone de herramientas estadísticas potentes que facilitan la investigación, devolviendo los resultados de forma gráfica.

Para la segunda parte del proyecto se pretende encontrar el origen genético de las diferencias clínicas entre sexos para así avanzar en el conocimiento sobre el Alzheimer. Para realizar esto, existen varias estrategias posibles que se han tenido en cuenta a la hora de plantear el estudio: análisis de asociación de GWAs o análisis de expresión. Finalmente se ha optado por realizar la segunda opción, de forma que se han evaluado los genes expresados diferencialmente en pacientes con Alzheimer teniendo en cuenta su sexo.

Los datos a utilizar para el análisis de expresión se han extraído de la base de datos Gene Expression Omnibus (GEO). Todo el análisis se ha realizado también con paquetes específicos de R y Bioconductor, además de con el uso de otras bases de datos y navegadores genómicos como *National Center for Biotecnology Information* (NCBI), que ha permitido visualizar de forma clara aspectos importantes de los resultados, como la región genómica donde se encuentra cada gen.

Una vez obtenidos todos los resultados de ambos estudios se ha comprobado si estos pueden correlacionarse de alguna manera, para obtener una posible explicación de las disparidades de género en el Alzheimer o, si de lo contrario, los resultados no permiten extraer ninguna conclusión al respecto.

Como parte final del proyecto, además de realizar el análisis comparativo, se han contrastado los resultados con datos ya conocidos de otras investigaciones. De esta manera se ha valorado si las conclusiones extraídas de este trabajo son coincidentes con las obtenidas en otros artículos ya publicados.

1.4. Planificación

Durante la realización de cada tarea especificada, se han ido anotando los scripts utilizados en cada análisis, referenciando la bibliografía pertinente y redactando en forma de esbozo los resultados y las conclusiones obtenidas, con el objetivo de hacer la redacción final más sencilla y rápida.

1.4.1. Tareas

Objetivo I – i

1. Seleccionar datos clínicos de pacientes enfermos de Alzheimer, donde se especifique el género de cada uno de ellos junto con sus características clínicas.
2. Explorar el origen de los datos.
3. Explorar los datos para obtener una idea general de la información contenida.
4. Seleccionar las variables que interesen ser incluidas en el análisis.
5. Lectura del archivo en R.
6. Depurar y corregir los datos para evitar que haya valores NA, por ejemplo.
7. Obtener una idea general de la distribución de los datos.
8. Realizar el análisis descriptivo y valorar las diferencias encontradas a simple vista.

Objetivo I – ii

9. Realizar contrastes de hipótesis para identificar diferencias entre sexos.
10. Realizar análisis de regresión para identificar asociaciones concretas.
11. Interpretar los resultados obtenidos, valorando si son significativos o no.

Objetivo I – iii

12. Selección de varios estudios similares que también analicen las diferencias clínicas entre sexos.
13. Comparar los resultados de los estudios seleccionados con los resultados obtenidos durante el análisis estadístico.

Objetivo II – i

14. Seleccionar datos resultantes de microarray para realizar el estudio de expresión, a poder ser de *Homo sapiens*. Intentar que los datos a escoger puedan tener cierta relación con las variables estudiadas en el estudio clínico para que todo el proyecto tenga coherencia.
15. Explorar el diseño experimental que se ha utilizado para obtener esos datos.
16. Comprobar que el formato de los ficheros sea de tipo .CEL, para poder realizar el análisis de expresión.
17. Crear un archivo *targets.csv* en función de cómo se distribuyan las muestras.
18. Lectura de archivos en R.
19. Realizar los controles de calidad necesarios.
20. Suprimir datos de mala calidad.
21. Normalizar los datos.
22. Segundo control de calidad.
23. Filtrar de forma no específica los datos.
24. Identificar los genes expresados diferencialmente.
25. Anotar los resultados obtenidos.
26. Realizar una comparación múltiple entre los grupos.
27. Visualización de perfiles de expresión.
28. Clasificar los genes identificados.

29. Visualización post-analítica e interpretación de resultados. Identificar genes diferencialmente expresados definitivos.

Objetivo II – ii

30. Analizar la significación biológica de los genes identificados.
31. Realizar un análisis de *pathways* para determinar en qué vías se ven involucrados los genes identificados.
32. Representar de forma visual las vías identificadas.
33. Interpretar los resultados.

Objetivo II – iii

34. Localizar en el genoma de referencia los genes seleccionados, indicando en qué cromosomas están, cuáles son sus coordenadas, etc.

Objetivo III

35. Comprobar si pueden combinarse todos los resultados obtenidos en el proyecto para extraer una conclusión que explique la disparidad entre sexos en personas con Alzheimer, o si, por el contrario, se han obtenido resultados completamente diferentes y no puede extraerse una conclusión clara.
36. Redactar las conclusiones obtenidas.

Detalles finales

37. Redactar memoria definitiva.
    a. Introducción, materiales y métodos.
    b. Resultados, discusión y conclusiones.
    c. Resto de apartados como *abstract*, bibliografía, etc.
38. Revisar redacción y estructura.
39. Hacer presentación Power Point.
40. Preparar defensa pública.

1.4.2. Hitos

A continuación, se explican los hitos del proyecto:

1. **Selección de datos.** La parte más importante del trabajo es la correcta elección de los datos que van a analizarse. Estos datos deben contener un tamaño muestral óptimo y deben tener una coherencia que permita relacionar el análisis estadístico y el análisis de expresión. Es decir, ambos datos escogidos deben tener en común alguna variable o característica que pueda asociarse para poder extraer conclusiones coherentes. Además, es importante que esté identificado el género de cada individuo.

   Por este motivo, es oportuno invertir el tiempo necesario en escoger los datos que se adecúen a nuestro estudio y, esto debe haberse hecho como máximo el 15 de octubre.

2. Análisis estadístico. Finalizar el análisis estadístico a tiempo es esencial para poder interpretar los resultados obtenidos de forma adecuada y compararlos con los resultados de otros estudios para poder extraer las primeras conclusiones del objetivo I. Esto debería estar finalizado el 26 de octubre.

3. Control de calidad. Este paso es fundamental para poder iniciar el análisis de expresión de forma correcta. Además, es importante comprobar que los scripts funcionan sin problemas. El control de calidad debería estar hecho para el día 5 de noviembre.

4. PEC 2 – Desarrollo del trabajo, Fase 1. La primera actividad de evaluación consiste en realizar un informe de seguimiento del proyecto. Esto debe ser entregado como máximo el día 16 de noviembre.

5. Identificación de genes expresados diferencialmente. El análisis debe estar finalizado como muy tarde el día 18 de noviembre. Esto es importante para poder interpretar los resultados y extraer las primeras conclusiones del análisis.

6. Anotación de genes. Esta parte es importante e imprescindible para que los resultados obtenidos hasta este momento tengan sentido y significado. La anotación permitirá saber qué genes están involucrados en las diferencias entre sexos y el lugar del genoma donde se encuentran. Esto debería estar hecho el 30 de noviembre.

7. PEC 3 – Desarrollo del trabajo, Fase 2. De nuevo, se realiza un segundo seguimiento del proyecto que debe entregarse como máximo el día 14 de diciembre.

8. Comparación de resultados. Para poder comenzar con la redacción de la memoria debería estar terminado el objetivo III, es decir, se debería haber comparado los resultados obtenidos en ambos análisis para intentar redactar un borrador de lo que serían las conclusiones. Es importante que esté terminado el día 15 de diciembre.

9. Redacción memoria. Lo idóneo sería tener la redacción terminada varios días antes de la entrega para poder hacer un repaso y perfeccionar los últimos detalles. Por este motivo, es importante que esté terminada para el día 31 de diciembre.

10. PEC 4 – Cierre de la memoria. La memoria debe estar finalizada y entregada el día 5 de enero.

11. PEC 5a – presentación. El día 10 de enero debería estar lista la presentación PowerPoint para poder preparar la defensa de forma adecuada.

12. PEC 5b – Defensa pública. A partir del 13 de enero se preparará la defensa pública.

1.4.3. Calendario

El siguiente diagrama de Gantt (Figura 1) muestra la calendarización de las diferentes tareas e hitos definidos en los apartados anteriores. Como se observa, están representados diferentes colores y formas:

- Rectángulos: representan las tareas.
  - Color azul: correspondientes al objetivo I.
  - Color salmón: correspondientes al objetivo II.
  - Color gris: correspondientes al objetivo III.
  - Color lila: correspondientes a detalles finales.
  - Color amarillo: correspondientes a la presentación y defensa.
- Rombos: representan los hitos.
  - Color fucsia: hitos del proyecto.
  - Color verde: actividades de evaluación (PECs).

1.5. Breve sumario de productos obtenidos

Como resultado del proyecto se obtienen evidencias significativas sobre diferencias clínicas y genéticas específicas de género en personas con la enfermedad de Alzheimer.

A partir del análisis estadístico se demuestra que las mujeres con Alzheimer tienen, en general, un volumen cerebral menor al de los hombres con Alzheimer. Además, se observa una disminución del volumen del hipocampo en los pacientes, a medida que aumenta la edad. Los resultados obtenidos en el análisis estadístico se muestran en las figuras 8-30.

Por otra parte, a raíz del análisis de expresión diferencial, se confirma únicamente la diferencia de volumen del hipocampo entre hombres y mujeres con la enfermedad. En este caso, se identifican 193 genes diferencialmente expresados que podrían ser clave para futuras investigaciones. Los 5 más diferencialmente expresados y, por tanto, considerados los más importantes son *Sp1, MBNL1, ALCAM, RAD21* y *CLIC4.*

Los resultados obtenidos en ambos análisis se adjuntan en el enlace Github https://github.com/olrema/TFM, en forma de tablas, archivos *html* y gráficos.

La Tabla 1 resume claramente los principales resultados obtenidos en los dos análisis:

Tabla 1. Sumario de resultados obtenidos. Los "✓" indican que se han encontrado diferencias significativas; las "X" que no se han encontrado diferencias; Los "-" indican "no evaluado".

	Diferencias significativas específicas de sexo			
	Hipocampo	Ventrículos	Corteza entorrinal	Cerebro completo
Análisis estadístico	✓	✓	✓	✓
Análisis de expresión	✓	-	X	-

Los códigos y datos utilizados para llevar a cabo el proyecto también se encuentran en el enlace Github anterior. Todo su contenido se detalla en Materiales y métodos.

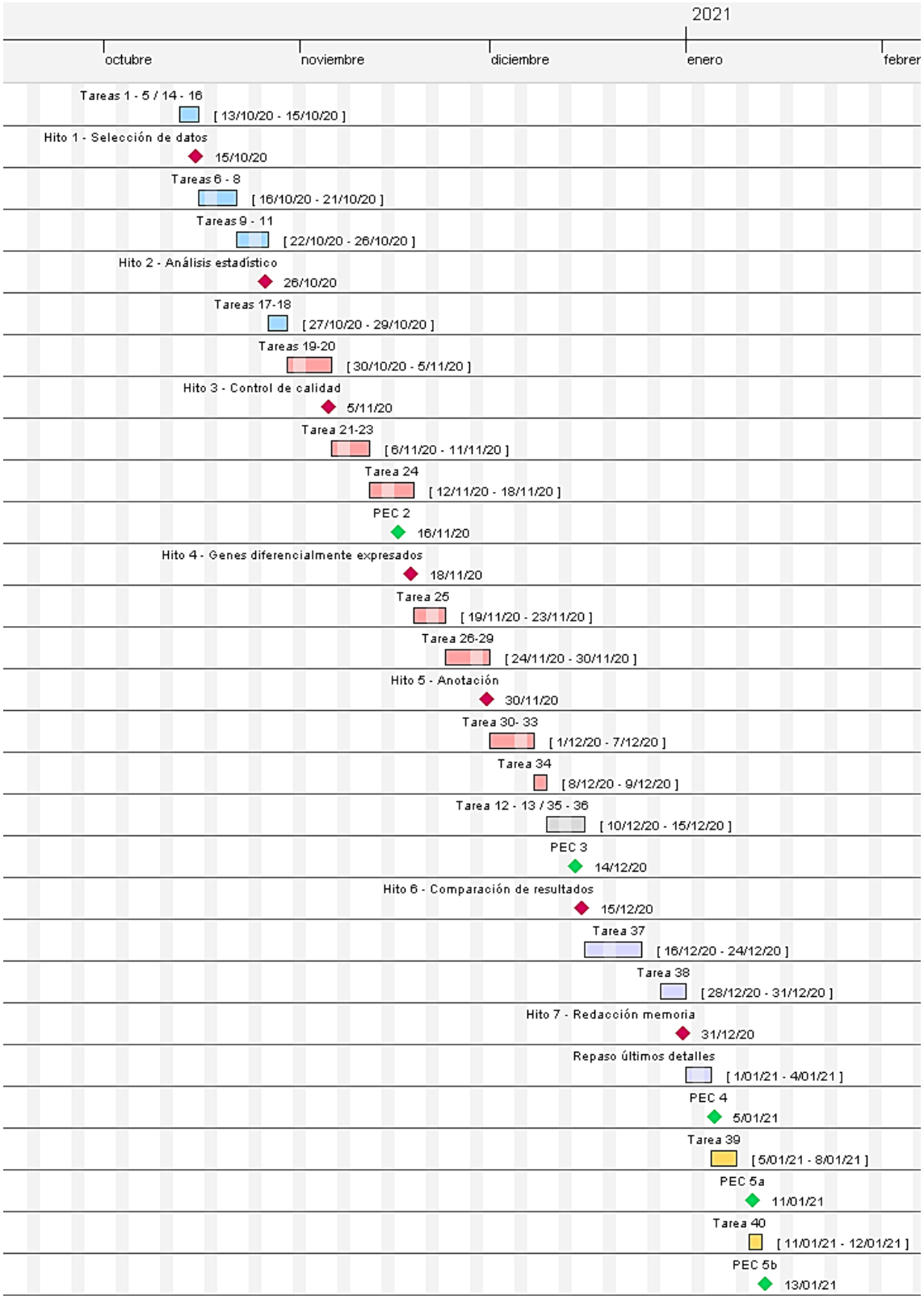

Figura 1. Diagrama de Gantt con la planificación del proyecto.

1.6. Breve descripción del resto de capítulos

A continuación, se explica brevemente el contenido del resto de capítulos:

- Materiales y métodos:
  - Datos: se explica de forma detallada toda la información disponible sobre los datos utilizados en los dos análisis del proyecto. Se especifica el estudio de origen, la fuente de donde se han extraído, el tipo de dato, el formato, la información contenida y cualquier otra información que sea de interés y/o necesaria.
  - Análisis estadístico: aquí se explica el programa utilizado para llevar a cabo el estudio y toda la metodología seguida en los diferentes procesos realizados:
    - Análisis descriptivo.
    - Contraste de hipótesis.
    - Análisis de regresión.
  - Análisis de expresión: se definirá todo el proceso del análisis, indicando los paquetes de R utilizados, los valores definidos para obtener los resultados, así como todos los pasos que se han llevado a cabo (pipeline) y la metodología utilizada.
- Resultados:
  - Análisis descriptivo: se muestra cómo se distribuyen los datos clínicos con los que se trabaja, centrando siempre el interés en observar las diferencias entre sexos para cada variable.
  - Análisis estadístico:
    - Contraste de hipótesis. Se indican los resultados de cada contraste realizado, mostrando en cada caso si existen diferencias significativas entre géneros y acabando de especificar las pruebas llevadas a cabo junto con los resultados numéricos obtenidos.
    - Análisis de regresión. De la misma manera que en el apartado anterior, se define el modelo lineal de regresión obtenido, así como la recta de mínimos cuadrados. Además, se especifican los resultados obtenidos para el análisis de regresión múltiple, mostrando las variables que presentan mayor relación con la variable respuesta (en este caso el volumen del hipocampo).
  - Análisis de expresión: se presentan los resultados para el análisis de expresión. Aquí se muestran tanto los resultados obtenidos durante todo el proceso, así como los genes finales identificados como diferencialmente expresados. Se muestran también los resultados del análisis de pathways de forma visual para poder identificar claramente las vías involucradas. Por último, se identifica la región del genoma en la que se encuentran los genes seleccionados.

- **Discusión**: se analizan los resultados obtenidos en cada apartado, intentando encontrar el sentido y la relación biológica que pueden tener. También se hace una valoración crítica del trabajo realizado y se valoran posibles mejoras.
  - **Análisis estadístico**.
  - **Análisis de expresión diferencial**.
  - **Estudio comparativo**: se explica la relación entre los resultados obtenidos de ambos análisis. Se especifica cómo se cree que pueden combinarse esos datos y se presentan artículos de estudios que pueden confirmar esa relación.

2. Materiales y métodos

El proyecto se ha estructurado en dos tipos de análisis diferentes, uno estadístico donde se valoran las diferencias clínicas de género entre los pacientes de Alzheimer y uno de expresión, con el que se pretende identificar genes expresados diferencialmente para evaluar las disparidades genéticas. Cada análisis requiere un tipo de dato concreto, motivo por el que los datos utilizados no provienen del mismo estudio de origen. No obstante, los datos de ambos análisis tienen un nexo a partir del cual pueden relacionarse para extraer conclusiones conjuntas.

Toda la información externa referenciada a lo largo del proyecto se ha gestionado con la aplicación Mendeley. Además, todos los artículos científicos han sido extraídos de PubMed.

Con el fin de servir como anexo para todo aquel que quiera reproducir de nuevo todos los análisis realizados o visualizar de forma completa los resultados obtenidos, en el siguiente enlace de GitHub (https://github.com/olrema/TFM) se adjunta toda la información que se refleja en este trabajo, desde los códigos R ejecutados hasta los datos utilizados para cada análisis y los resultados obtenidos (Tabla 2).

Tabla 2. Correspondencia de archivos adjuntados como anexo en GitHub. Se marcan en negrita las carpetas y en cursiva el resto de los archivos.

Memoria final del proyecto	*Memoria_OlgaReyes.pdf*
AE	
Datos clínicos originales	*DC_originales.xlsx*
Datos clínicos procesados	*DC_procesados.xlsx*
Datos clínicos finales (.xlsx)	*DC_finales.xlsx*
Datos clínicos finales (.csv)	*DC_finales.csv*
Código R y resultados obtenidos del análisis estadístico (.html)	Statistic.html
Código R y resultados obtenidos del análisis estadístico (.Rmd)	Statistic.Rmd
AED/Hip	
Muestras iniciales junto con el archivo *targets.csv*	***data1***
Muestras finalmente utilizadas junto con el archivo *targets.csv* definitivo	***data***
Código R y resultados obtenidos del análisis de expresión (.html)	*Microarray.html*
Código R utilizado para el análisis de expresión (.Rmd)	*Microarray.Rmd*

results	
Datos representados en el *heatmap*	*data4Heatmap.csv*
Datos normalizados de cada muestra (.csv)	*normalized.Data.zip*
Datos normalizados de cada muestra (.Rda)	*normalized.Data.Rda*
Datos normalizados de cada muestra después del filtraje	*normalized.Filtered.Data.csv*
Genes ordenados de mayor a menor expresión diferencial (según el p-valor) en individuos con Alzheimer	*topAnnotated_FEMvsMALE.AD.csv*
Genes ordenados de mayor a menor expresión diferencial (según el p-valor) en individuos control	*topAnnotated_FEMvsMALE.CT.csv*
Resultados del control de calidad de los datos normalizados	*QCDir.Norm*
Resultados del control de calidad de los datos crudos	*QCDir.Raw*
AED/CE	
Muestras finalmente utilizadas junto con el archivo *targets.csv* definitivo	*data*
Código R y resultados obtenidos del análisis de expresión (.html)	*Microarray-EC.html*
Código R utilizado para el análisis de expresión (.Rmd)	*Microarray-EC.Rmd*
results	
Datos normalizados de cada muestra (.csv)	*normalized.Data.zip*
Datos normalizados de cada muestra (.Rda)	*normalized.Data.Rda*
Datos normalizados de cada muestra después del filtraje	*normalized.Filtered.Data.csv*
Genes ordenados de mayor a menor expresión diferencial (según el p-valor) en individuos con Alzheimer	*topAnnotated_FEMvsMALE.AD.csv*
Genes ordenados de mayor a menor expresión diferencial (según el p-valor) en individuos control	*topAnnotated_FEMvsMALE.CT.csv*
Resultados del control de calidad de los datos normalizados	*QCDir.Norm*
Resultados del control de calidad de los datos crudos	*QCDir.Raw*

2.1. Datos

Para llevar a cabo el **análisis estadístico** sobre las características clínicas de los pacientes se han recogido datos procedentes de la plataforma *Alzheimer's Disease Neuroimagin Initiative* (ADNI). Los datos utilizados corresponden a un subconjunto de los presentados en la tabla "*adnimerge"* de la plataforma, la cual combina varias de las variables clave de diversos casos y resúmenes de protocolos ADNI. Esta tabla se actualiza diariamente directamente desde *Alzheimer's Disease Coorporative Study* (ADCS). La información recogida en la tabla surge a partir de un estudio longitudinal que se realiza en pacientes con diferentes grados de demencia y distintos diagnósticos médicos, en el que se registran los datos de cada uno de los pacientes cada cierto tiempo para observar su evolución clínica [12].

Para poder acceder a los datos ofrecidos por la plataforma ADNI y descargarlos, es necesario solicitar un permiso especial que es complicado de obtener, por lo que no ha sido posible utilizar los datos oficiales actualizados del estudio. En su lugar, se ha accedido a datos procedentes del mismo proyecto, un poco más antiguos, gracias a la publicación extraoficial de éstos en la plataforma Github [13].

La tabla "*adnimerge*" original contiene 85 variables, incluyendo los datos clínicos iniciales de cada paciente junto con los obtenidos en la última valoración médica. Para realizar este análisis únicamente se han seleccionado los registros iniciales de aquellos pacientes con la enfermedad de Alzheimer y se han escogido sólo 12 variables de interés, las cuales se explican a continuación y se reflejan en la Tabla 3. Este proceso de selección de datos se ha llevado a cabo mediante la aplicación Microsoft Excel.

- RID: número de fila de los datos.
- PTID: código identificativo de los pacientes.
- AGE: edad (años).
- PTGENDER: género (1: mujer; 0: hombre)
- PTEDUCAT: tiempo de educación (años).
- PTMARRY: estado civil.
    - 0: casada/o.
    - 1: divorciada/o.
    - 2: nunca ha estado casada/o.
    - 3: viuda/o.
- APOE4: cantidad de alelos de *ApoE4* (0; 1; 2).
- EXAMDATE_bl: fecha en que se recogen los datos.
- Ventricles_bl: volumen de los ventrículos (mm^3).
- Hippocampus_bl: volumen del hipocampo (mm^3).
- WholeBrain_bl: volumen de todo el cerebro (mm^3).
- Entorhinal_bl: volumen de la corteza entorrinal (mm^3).

Tabla 3. Parte de los datos y las variables escogidas para el estudio.

RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
1	pac_0001	81.3	0	18	0	1	12/09/05	84599	5319	1129830	1791
2	pac_0002	75.4	0	10	0	1	06/10/05	25704	6729	875798	2050
3	pac_0003	73.9	1	12	0	1	10/11/05	26820	5485	1033540	2676
4	pac_0004	64.1	0	18	0	1	31/10/05	45401	7359	1222380	2895
5	pac_0005	80.1	0	12	0	0	29/11/05	77780	6232	1033070	2187
6	pac_0006	78.3	0	12	0	1	12/01/06	30910	4719	940406	1972

Con el objetivo de no disponer de datos vacíos en el análisis, se ha realizado un segundo filtraje con el programa R, en el que estos registros son eliminados del estudio. De esta forma, el tamaño muestral final queda en 239 pacientes.

De forma paralela, para realizar el **análisis de expresión diferencial**, se han utilizado datos publicados en la plataforma *Gene Expression Omnibus* (GEO), concretamente los correspondientes al número de acceso GSE48350. Estos datos han sido generados por Berchtold NC y Cotman CW, del Instituto UC Irvine para Deficiencias de Memoria y Trastornos Neurológicos (UCI MIND), en Estados Unidos (EE. UU.), utilizando la plataforma *Affymetrix Human Genome U133 Plus 2.0 Array*. En estos se recogen varios perfiles de expresión de microarray de Homo sapiens de personas sanas y de personas con la enfermedad de Alzheimer, de cuatro regiones del cerebro distintas: el hipocampo, la corteza entorrinal, la corteza frontal superior y la circunvolución post-central. Este conjunto de datos fue utilizado para analizar los cambios en la expresión génica relacionados con la enfermedad de Alzheimer, la región cerebral y la edad de los pacientes. Todos los datos procedentes de los individuos sanos (controles) fueron recogidos de GSE11882. Para obtener los datos, los investigadores recogieron tejido cerebral post-mórtem de bancos de cerebros y escogieron aquellos casos en los que había disponibles tres o más regiones cerebrales. Aunque estos datos fueron publicados en abril de 2014, la última actualización se llevó a cabo en junio de 2019 [14].

Para este proyecto se han utilizado los datos de expresión de sólo las dos zonas cerebrales que se incluyen en el análisis estadístico: hipocampo y corteza entorrinal. No obstante, en este caso, no se ha tenido en cuenta la edad de los pacientes y se han definido cuatro grupos de muestras para cada zona cerebral para llevar a cabo el análisis, diferenciando entre el género y la presencia o no de enfermedad (Figura 2).

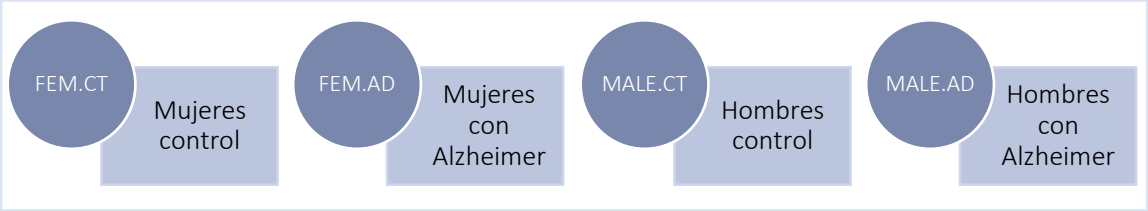

Figura 2. Grupos en los que se clasifican las muestras.

En GSE48350 aparecen publicados un total de 253 archivos en formato .CEL correspondientes a datos de expresión para diferentes participantes y según la zona cerebral evaluada. Este proyecto se ha basado en un subconjunto de muestras

correspondientes a las dos zonas cerebrales de interés. Este subconjunto se ha seleccionado de forma aleatoria hasta obtener un número similar de muestras en cada grupo definido dentro de cada zona cerebral:

- Hipocampo: 36 individuos.
- Corteza entorrinal: 28 individuos.

Aunque en el análisis estadístico y en el análisis de expresión llevados a cabo en este proyecto no se han utilizado datos de los mismos individuos, mismo año, lugar u origen, ambos conjuntos de datos tienen un nexo que permitirá relacionar los resultados obtenidos. Este nexo corresponde a las dos zonas cerebrales que se han estudiado independientemente en los dos análisis: hipocampo y corteza entorrinal. Estas dos variables son las de real interés en el proyecto, ya que son las que permitirán extraer conclusiones a partir de la comparación de los resultados de los análisis. De esta manera, se ha realizado un total de dos análisis de expresión, uno para cada zona cerebral.

2.2. Análisis estadístico

Antes de realizar el análisis estadístico de los datos explicados, se ha llevado a cabo una exploración general y un análisis descriptivo de éstos para observar cómo se distribuyen. Ambos procesos se han efectuado con el programa R.

2.2.1. Análisis descriptivo

Para visualizar la distribución de los datos se ha llevado a cabo un análisis descriptivo, estudiando los rasgos más importantes de cada variable y observando cómo se distribuyen en función de la variable de interés para el estudio (género). Para visualizar estos aspectos se han generado varias tablas y gráficos para cada variable independientemente y según el género de los pacientes.

2.2.2. Contraste de hipótesis

Se ha iniciado el análisis estadístico con un contraste de hipótesis para evaluar las diferencias presentes en ciertas variables de interés según el género de los pacientes. Se han valorado las diferencias de volumen cerebral y sus partes, en concreto, el volumen del hipocampo, del cerebro completo, de los ventrículos y de la corteza entorrinal.

Para cada contraste, se han definido unas hipótesis. Para la variable volumen del hipocampo:

- Hipótesis nula: no existen diferencias significativas entre las personas de diferente sexo y su volumen de hipocampo.
- Hipótesis alternativa: sí existen diferencias significativas entre las personas de diferente sexo y su volumen de hipocampo.

Para el volumen del cerebro completo:

- Hipótesis nula: no existen diferencias significativas entre las personas de diferente sexo y su volumen de cerebro completo.

- Hipótesis alternativa: sí existen diferencias significativas entre las personas de diferente sexo y su volumen de cerebro completo.

Para el volumen de los ventrículos:

- Hipótesis nula: no existen diferencias significativas entre las personas de diferente sexo y su volumen de ventrículos.
- Hipótesis alternativa: sí existen diferencias significativas entre las personas de diferente sexo y su volumen de ventrículos.

Para el volumen de la corteza entorrinal:

- Hipótesis nula: no existen diferencias significativas entre las personas de diferente sexo y su volumen de corteza entorrinal.
- Hipótesis alternativa: sí existen diferencias significativas entre las personas de diferente sexo y su volumen de corteza entorrinal.

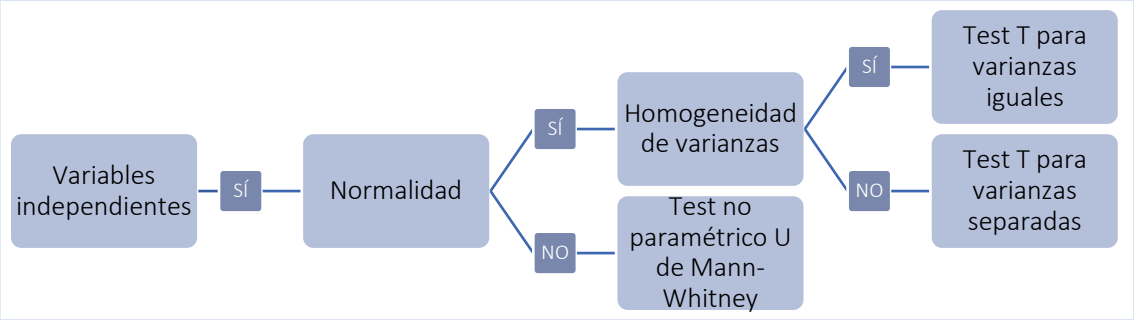

Figura 3. Estructura seguida para el contraste de hipótesis.

Para realizar cada uno de los contrastes de hipótesis, es necesario evaluar distintas propiedades de los datos. En función de las propiedades que muestra cada variable, se han llevado a cabo unas pruebas estadísticas u otras, siguiendo la estructura de la Figura 3. Se parte del hecho de que las variables estudiadas son independientes. Por lo tanto, en primer lugar, se ha procedido en todos los contrastes de la misma manera, comprobando la normalidad de los datos gráficamente y mediante el test estadístico Kolmogorov-Smirnov (`ks.test` en R), debido a que la muestra supera en los cuatro casos las 50 observaciones. En el caso de los datos que resultan no ser normales, se ha realizado la prueba no paramétrica U de Mann-Whitney. De lo contrario, se ha evaluado la homogeneidad de las varianzas y en función del resultado obtenido, se ha efectuado un tipo de test T concreto.

2.2.3. Análisis de regresión

Para finalizar el análisis estadístico se ha realizado un análisis de regresión lineal simple y un análisis de regresión lineal múltiple. Con el análisis de regresión lineal simple se pretende valorar si la edad de los individuos con Alzheimer influye en su volumen del hipocampo. Además, para valorar si el volumen del hipocampo puede verse influido por alguna otra variable, se realiza también el análisis de regresión múltiple.

Por lo tanto, se definen las siguientes hipótesis para el análisis de regresión lineal simple:

- Hipótesis nula: no existe relación estadísticamente significativa entre la edad y el volumen del hipocampo.
- Hipótesis alternativa: existe relación estadísticamente significativa entre la edad y el volumen del hipocampo.

Hipótesis para el análisis de regresión lineal múltiple:

- Hipótesis nula: no existe relación estadísticamente significativa entre el volumen del hipocampo y alguna otra variable.
- Hipótesis alternativa: existe relación estadísticamente significativa entre el volumen del hipocampo y alguna otra variable.

2.3. Análisis de expresión

Se han llevado a cabo un total de dos análisis de expresión, uno para el hipocampo y otro para la corteza entorrinal. Las etapas que se han seguido para los dos análisis de microarrays se resumen en la Figura 4 y se han llevado a cabo mediante líneas de código ejecutadas en el programa R y haciendo uso de las librerías desarrolladas para el análisis de microarray en el proyecto Bioconductor.

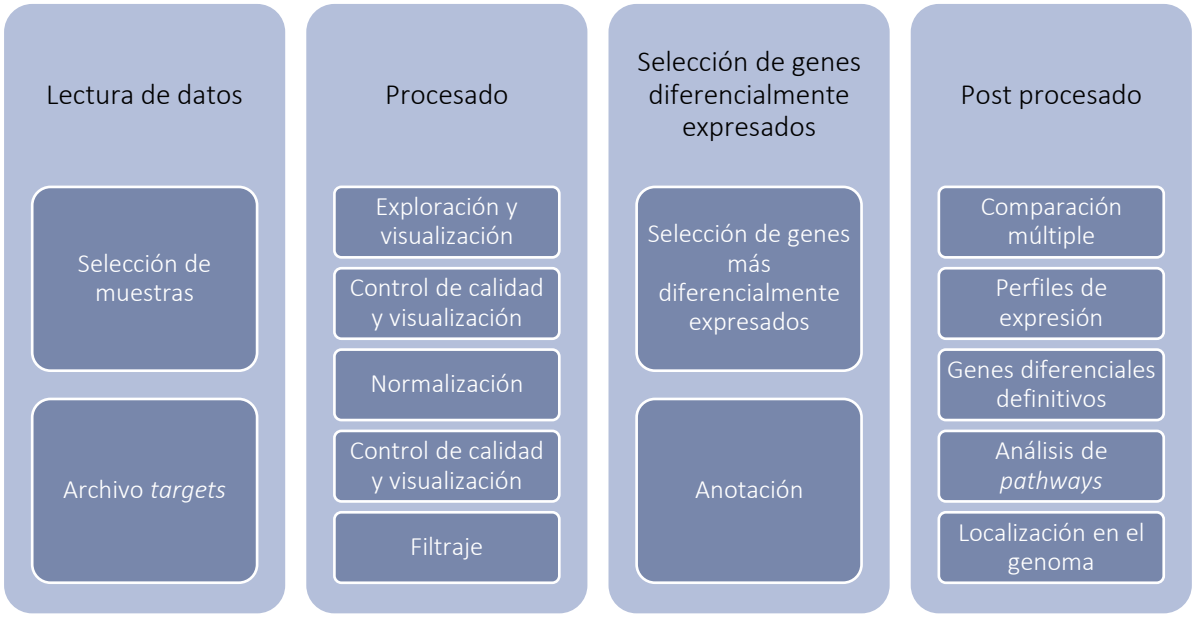

Figura 4. Pipeline o etapas seguidas en el análisis de expresión diferencial del hipocampo y la corteza entorrinal.

Los siguientes listados corresponden a los principales paquetes instalados y utilizados para los análisis de expresión.

Paquetes de Bioconductor:

```
> oligo
> Biobase
> arrayQualityMetrics
> genefilter
> hgu133plus2.db
> limma
> ReactomePA
```

Paquetes de R:

```
> BiocManager
> ggplot2
> ggrepel
> gplots
```

> `clusterProfiler`
> `enrichplot`

Después de localizar unos datos aptos para el proyecto y de valorar la información contenida en ellos, se ha realizado una selección aleatoria de las muestras de microarrays para cada análisis de forma manual, descargando los archivos en formato .CEL desde GEO. Posteriormente, se han definido los grupos en los que se dividen estas muestras seleccionadas comentados anteriormente (Figura 2).

A continuación, se ha creado un archivo *targets.csv,* que contiene la identificación de cada archivo con la asignación de las muestras a cada condición experimental o grupo. Este archivo *targets.csv* es prácticamente igual en los dos análisis exceptuando las muestras que aparecen. Consiste en una tabla donde cada fila corresponde a una muestra y las columnas corresponden a: el nombre identificativo de la muestra, el grupo al que pertenece la muestra, el género, la condición del paciente y el nombre identificativo que recibe el archivo (Tabla 4).

Con el uso de los paquetes `oligo` y `Biobase` de Bioconductor se ha realizado la lectura de los datos, utilizando tanto el contenido del archivo *targets* como el de los archivos .CEL. Gracias a este método, es posible leer los datos al mismo tiempo que se asigna a cada muestra el grupo para el análisis.

Tabla 4. Ejemplo de contenido del archivo "targets.csv".

FileName	Group	Gender	Genome	ShortName
GSM300168	FEM.CT	FEM	CT	FEM.CT.1
GSM300169	MALE.CT	MALE	CT	MALE.CT.1
GSM1176212	FEM.AD	FEM	AD	FEM.AD.1
GSM1176221	MALE.AD	MALE	AD	MALE.AD.1

Una vez se han leído los datos con el programa, se ha procedido con la exploración y la visualización de éstos con el fin de obtener una idea de cómo se distribuyen. Para ello, se han generado gráficos de densidad y diagramas de cajas. Además, se han representado las muestras en gráficos de análisis de componentes principales (PCA), con el objetivo de observar posibles agrupamientos de las muestras y su distribución. Todos estos gráficos se han generado a partir de los paquetes `ggplot2` y `ggrepel`.

Antes de normalizar los datos, se ha valorado la calidad de los mismos con el paquete `arrayQualityMetrics`[15]. Únicamente las muestras que han demostrado una calidad suficiente se han sometido al resto del análisis. En este caso, el paquete utiliza tres métodos distintos para detectar *outliers:* distancia entre arrays, diagrama de cajas y gráfico MA. Cuando alguna muestra es detectada como *outlier* por alguno de estos tres métodos, es marcada en una tabla resumen. Para este proyecto, las muestras marcadas por los tres criterios en el primer control de calidad han sido excluidas del análisis.

A continuación, se han normalizado los datos utilizando el método *Robust Multiarray Average* (RMA)[16], un proceso de tres etapas que incluye la corrección de fondo, la normalización y la sumarización de las sondas asociadas a cada grupo de sondas para dar un único valor. Seguidamente se ha repetido el control de calidad, esta vez con los datos normalizados con el fin de valorar si la calidad de estos ha mejorado y se han representado de nuevo gráficamente para observar su distribución.

Antes de realizar el filtraje no específico de las muestras, se ha realizado una detección de los genes más variables y se ha generado un gráfico para observar esa distribución de variabilidad génica. Este gráfico (mostrado el correspondiente al hipocampo en la Figura 5) es muy similar en los dos análisis y ambos resultan en una desviación estándar de la variabilidad del 0,5. Este valor ha ayudado en la definición del umbral de filtraje para eliminar los genes con menor variabilidad. Una vez definido este valor numérico, se ha procedido a realizar el filtraje no específico de aquellos genes que no han superado el umbral, con el paquete **`genefilter`**. Durante este filtraje se ha seleccionado e introducido el paquete de anotación que se utilizará en las siguientes etapas del análisis (**`hgu133plus2.db`**). Como resultado, se han filtrado en los dos análisis un total de 44.595 genes, pasando de tener 54.675 genes iniciales a 10.080.

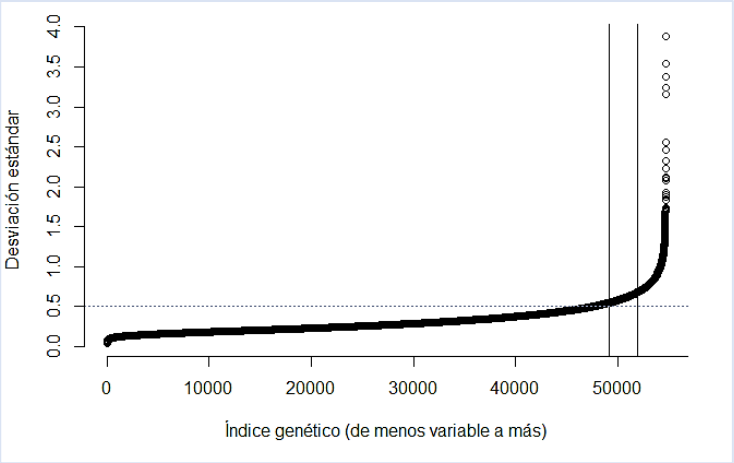

Figura 5. Distribución de la variabilidad génica para el hipocampo.

La selección de genes diferencialmente expresados entre condiciones experimentales se ha basado en el método que utiliza el paquete **`limma`**, el cual consiste en usar modelos lineales para analizar experimentos diseñados y evaluar la expresión diferencial. **`limma`** utiliza métodos empíricos bayesianos para combinar la información de cada gen con la de todos los genes restantes, lo cual resulta idóneo para este proyecto[17]. Para llevar a cabo la selección de genes diferenciales, se han definido unos contrastes entre grupos con tal de comparar la expresión entre hombres y mujeres según si son control o tienen Alzheimer (Figura 6). A partir de estos contrastes se han creado unas matrices de diseño y de contrastes que, en este caso, son iguales para las dos zonas cerebrales analizadas.

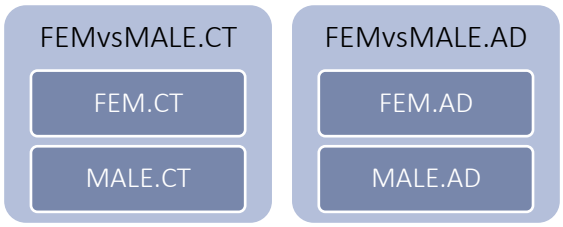

Figura 6. Contrastes definidos para la selección de genes diferencialmente expresados.

A partir de estas matrices y los datos normalizados, se ha ajustado el modelo lineal para cada gen con el paquete **`limma`** y se ha generado la selección de genes diferencialmente expresados.

A la hora de seleccionar los genes diferenciales se ha realizado un ajuste de p-valores utilizando el método de Benjamini y Hochberg con el fin de controlar el porcentaje de falsos positivos que puedan surgir del alto número de contrastes realizados de forma simultánea[18]. Se ha tenido en cuenta, entonces, este p-valor ajustado a la hora de filtrar

y seleccionar los genes más diferencialmente expresados en las etapas siguientes del análisis.

Para visualizar los genes más relevantes seleccionados en las dos comparaciones en función de su *p-valor* y *Log Fold Change* (LFC) se han resaltado utilizando *volcano-plots*, que organizan los genes en dos dimensiones. Esto permite observar el cambio medio de expresión entre los grupos comparados en el eje horizontal y los genes con un *p-valor* inferior en el eje vertical.

Con el objetivo de poder identificar fácilmente qué genes son los que se han seleccionado como diferencialmente expresados se han anotado utilizando la función `annotatedTopTable`, de forma que se visualiza el símbolo de cada gen y su nombre completo.

A continuación, se ha realizado una comparación múltiple de los genes seleccionados como diferencialmente expresados en cada contraste, con el objetivo de afinar aún más la selección y ver qué genes cambian simultáneamente. En los dos análisis de expresión se ha realizado esta comparación múltiple con un *p-valor* ajustado (adj.P.Val) definido de 0,05 y un *LFC* de 1. Esto se realiza con la función `decideTests`, con la que se obtiene una matriz que incluye cada comparación junto con los genes expresados diferencialmente. A partir de la interpretación de esta matriz, se ha generado una tabla que contiene el número total de genes sobreexpresados (*up*), subexpresados (*down*) o no significativos (*NotSig*) para cada contraste. Los resultados obtenidos en dicha tabla representan los genes diferenciales definitivos y se han representado mediante un diagrama de Venn, que permite visualizar claramente la cantidad de genes diferenciales compartidos por cada contraste.

Seguidamente se han utilizado *heatmaps* o mapas de colores con el objetivo de agrupar la expresión de cada gen seleccionado como diferencialmente expresado con tal de identificar perfiles de expresión en los que haya genes que se encuentren *up* o *down* regulados simultáneamente. Gracias a esta representación se consigue visualizar cómo es la expresión de cada gen según la muestra.

Con el objetivo de obtener información sobre el proceso biológico en el que están involucrados los genes seleccionados como diferenciales, se ha realizado una clasificación utilizando la base de datos *Gene Ontology* (GO) con la función `groupGO` a un nivel de especificidad 3. De esta forma se han agrupado los genes en los diferentes procesos biológicos en los que participan.

Por último, se ha realizado un análisis de *pathways* a de estos genes. Dentro de todas las posibilidades que existen a la hora de realizar este procedimiento[19], se ha decidido llevarlo a cabo mediante dos métodos distintos. Por una parte, se ha realizado un análisis de *pathways* con el uso del paquete `clusterProfiler` utilizando la función `enrichKEGG`, que ejecuta el análisis extrayendo la información sobre las vías biológicas publicadas en el repositorio público *Kyoto Encyclopedia of Genes and Genomes* (KEGG)[20]. Paralelamente se ha realizado un segundo análisis de pathways utilizando la función `enrichPathway` del paquete `ReactomePA`, cuya información se recoge de la base de

datos *Reactome*[21]. Ambos métodos ofrecen información sobre las vías en las que los genes identificados pueden estar involucrados, según dos bases de datos distintas. Los dos análisis de *pathways* han sido ejecutados con un p-valor de 0.1.

Los resultados de estos dos análisis se han representado gráficamente mediante diversas funciones del paquete `enrichplot.`

Para finalizar el análisis de expresión, una vez obtenidos los cinco genes más diferencialmente expresados, se ha realizado una búsqueda en NCBI para localizarlos en el genoma humano, Además, para conocer las funciones realizadas por las proteínas codificadas en estos genes, se ha utilizado *UniProt*.

3. Resultados

3.1. Análisis descriptivo

La Tabla 5 muestra una primera idea de los datos con los que se trabaja. Es un resumen estadístico de las variables numéricas presentadas en los datos.

Tabla 5. Resumen estadístico de las variables numéricas de los datos. Las filas muestran la siguiente información respectivamente: valor mínimo, primer cuartil, mediana, media, tercer cuartil y valor máximo.

	AGE	PTEDUCAT	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
Min.	55.10	6.00	9166	3091	727478	1438
1st Qu.	69.50	13.00	30804	4996	887963	2285
Median	75.10	16.00	44550	5629	966403	2773
Mean	74.30	15.23	48361	5774	974504	2817
3rd Qu.	80.05	18.00	59906	6429	1046700	3233
Max.	90.90	20.00	59906	9572	1364690	5430

Tal y como se ve representado en la Figura 7, forman parte del estudio más cantidad de hombres que de mujeres. En concreto se incluye un total de 131 hombres y 108 mujeres.

La edad media de los pacientes hombres es superior a la de las mujeres, no obstante, la de ambos sexos ronda los 73-75 años (Figura 8). Tanto la paciente de mayor edad, cuyos datos clínicos se encuentran en La Tabla 6, como la de menor edad, representada en la Tabla 7, son mujeres que coinciden en no disponer ningún alelo *ApoE4.*

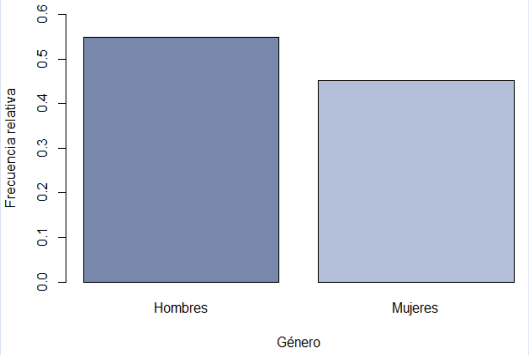

Figura 7. Frecuencia relativa de hombres y mujeres.

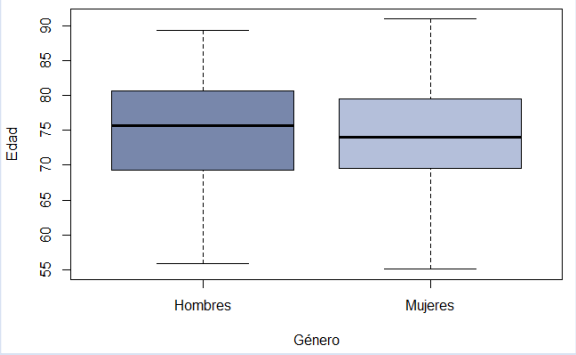

Figura 8. Distribución de edades según el género.

Tabla 6. Datos del paciente de mayor edad.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
97	97	pac_0097	90.9	1	16	3	0	16/11/06	39494	5694	824632	2982

Tabla 7. Datos del paciente de menor edad.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
150	150	pac_0150	55.1	1	18	0	0	15/05/07	41709	6648	961588	2621

En la Figura 9 se puede observar cómo los hombres del estudio disponen de más años de educación que las mujeres. La media de años de educación de las mujeres es 14,4, mientras que la de los hombres es 15,87.

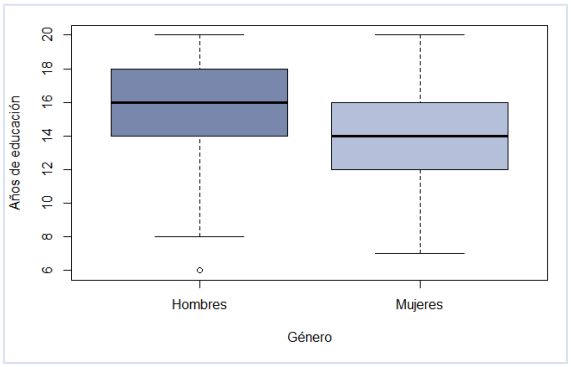

Figura 9. Distribución del tiempo de educación según el género.

En cuanto al estado civil de los pacientes, la mayoría de ellos están casados/as y un mínimo número de ellos no se ha casado nunca, tal y como se muestra en la Figura 10.

Al tener en cuenta el género de los pacientes a la hora de observar su estado civil, se aprecia como gran parte de los individuos de ambos géneros del estudio están casados, siendo superior la cantidad de hombres que de mujeres. En el resto de los estados civiles representados, las mujeres prevalecen sobre los hombres, aunque la diferencia entre géneros no es tan marcada como sucede con el estado civil "casado/a" (Figura 11).

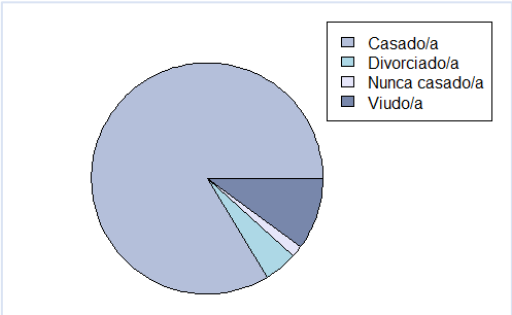

Figura 10. Frecuencia relativa del estado civil de los pacientes.

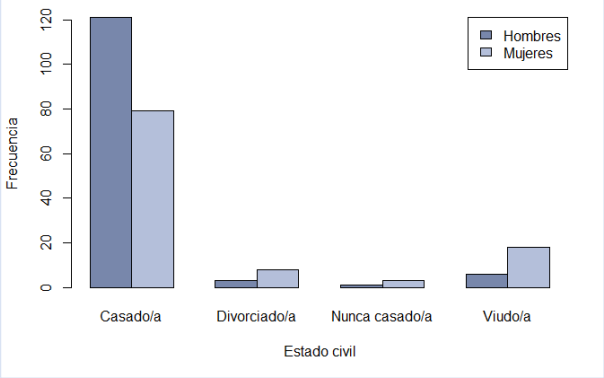

Figura 11. Frecuencia del estado civil de hombres y mujeres.

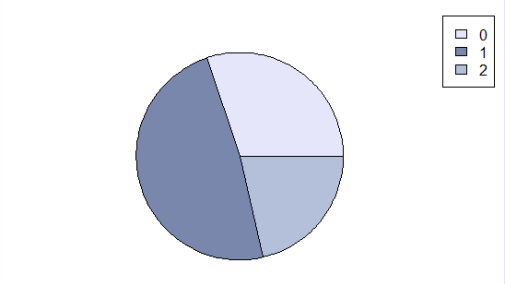

Figura 12. Frecuencia relativa de la cantidad de alelos ApoE4 de los pacientes.

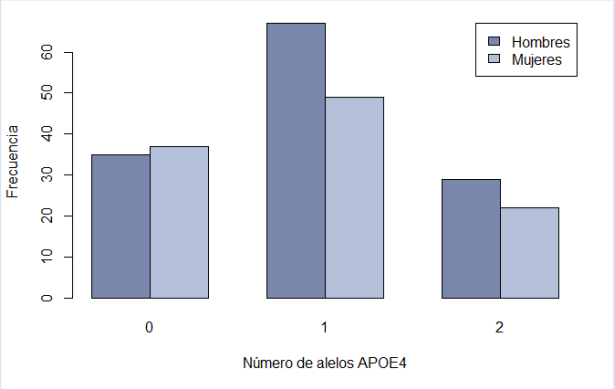

Figura 13. Frecuencia de alelos ApoE4 en hombres y mujeres.

La mayoría de los pacientes involucrados en el estudio, independientemente de su género, disponen únicamente de un alelo *ApoE4*. La distribución de la disposición de 0 o 2 alelos es bastante similar (Figura 12). Cuando se tiene en cuenta el género de los pacientes, se observa que hay más cantidad de hombres que disponen de uno o dos alelos *ApoE4.* Finalmente, aunque la cantidad de hombres y mujeres es muy similar en la disposición de 0 alelos *ApoE4*, las mujeres son mínimamente superiores en número (Figura 13).

Tanto el volumen del cerebro completo como el de sus diferentes partes evaluadas en el estudio (hipocampo, corteza entorrinal y ventrículos) resulta ser superior en hombres que en mujeres. Aunque esto coincida en las cuatro variables, el volumen de éstas no es proporcional. Es decir, el paciente con el mayor volumen cerebral completo no es el mismo paciente con el mayor volumen de hipocampo. Esto podemos verlo en las Tablas 8-13.

Tabla 8. Datos del paciente con mayor volumen cerebral completo.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
28	28	pac_0028	56.4	0	16	0	2	22/05/06	25658	7354	1364690	4101

Tabla 9. Datos del paciente con menor volumen cerebral completo.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
112	112	pac_0112	72.8	1	16	0	2	19/12/06	78281	4732	727478	2553

Tabla 10. Datos del paciente con mayor volumen de hipocampo.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
11	11	pac_0011	62.4	0	20	0	0	03/01/06	9166	9572	993236	4196

Tabla 11. Datos del paciente con menor volumen de hipocampo.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
96	96	pac_0096	86.3	1	10	3	0	21/11/06	52909	3091	768300	2067

Tabla 12. Datos del paciente con el mayor volumen de corteza entorrinal.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
34	34	pac_0034	74.5	1	16	0	0	11/05/06	24880	7945	967608	5430

Tabla 13. Datos del paciente con el menor volumen de corteza entorrinal.

	RID	PTID	AGE	PTGENDER	PTEDUCAT	PTMARRY	APOE4	EXAMDATE_bl	Ventricles_bl	Hippocampus_bl	WholeBrain_bl	Entorhinal_bl
262	262	pac_0262	79	0	18	0	1	02/05/12	38585	4065	974302	1438

Las siguientes Figuras 14-17 muestran de forma gráfica la distribución del volumen de las partes del cerebro comentadas anteriormente, donde puede apreciarse de forma clara como los hombres acostumbran a tener un volumen superior al de las mujeres.

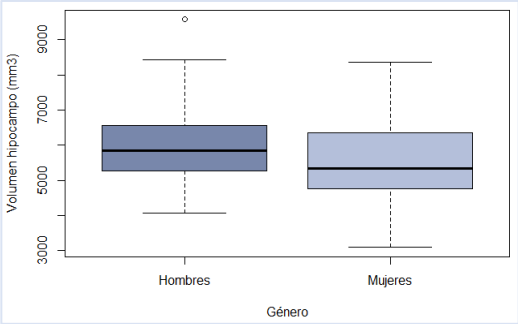

Figura 14. Distribución del volumen del hipocampo según el género.

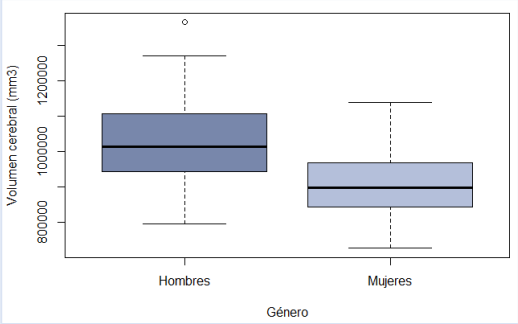

Figura 15. Distribución del volumen del cerebro completo según el género.

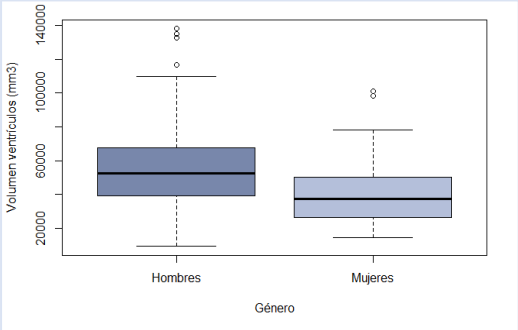

Figura 16. Distribución del volumen de los ventrículos según el género.

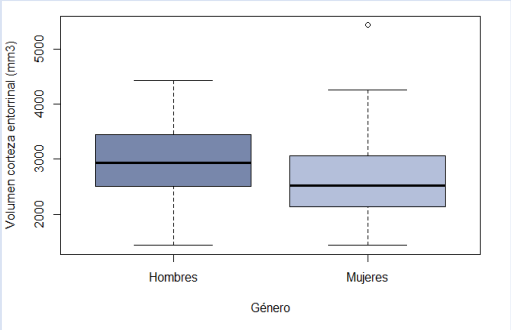

Figura 17. Distribución del volumen de la corteza entorrinal según el género.

3.2. Análisis estadístico

3.2.1. Contraste de hipótesis

En los siguientes apartados se muestran los resultados obtenidos a partir de los contrastes de hipótesis realizados para valorar las diferencias entre géneros para distintas variables.

3.2.1.1. Volumen del hipocampo

La primera variable que se evalúa es el volumen del hipocampo. La Figura 14 sugiere una diferencia entre géneros para esta variable, que se valora a continuación.

Tanto la Figura 18 como el test de normalidad (p-valor 0.03362) indican que los datos no siguen una distribución normal, por lo que para valorar si existen diferencias significativas se realiza la prueba no paramétrica de Mann-Whitney (`wilcox.test` en R).

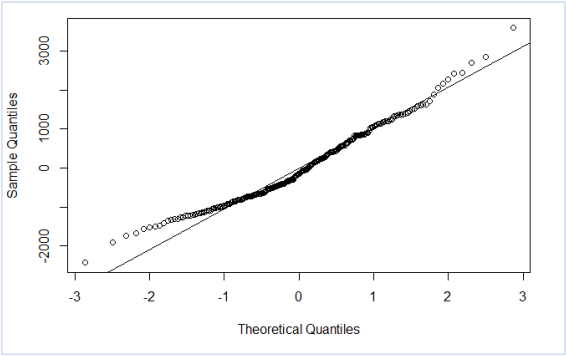

Figura 18. Gráfico Q-Q para las variables hipocampo y género.

Como resultado del test final se obtiene un p-valor de $2{,}2e^{-16}$ y un intervalo de confianza entre -5766 y -5531 con una probabilidad del 95% (Figura 19). Con todo esto, se rechaza la hipótesis nula planteada, indicando que existen diferencias significativas entre sexos para el volumen del hipocampo.

```
 Wilcoxon rank sum test with continuity correction

data:  clinic_data$PTGENDER and clinic_data$Hippocampus_bl
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -5766 -5531
sample estimates:
difference in location
                 -5629
```

Figura 19. Resumen del test Mann-Whitney para las variables hipocampo y género.

3.2.1.2. Volumen del cerebro completo

Se evalúa ahora el volumen del cerebro completo. Como aparece en la Figura 15 también se aprecia una diferencia entre ambos sexos.

En este caso, el test de normalidad (p-valor 0.5584) y la Figura 20 indican que el supuesto de normalidad se cumple para estas variables. Como consecuencia, antes de decidir la prueba estadística para testar si las diferencias son significativas, se evalúa la

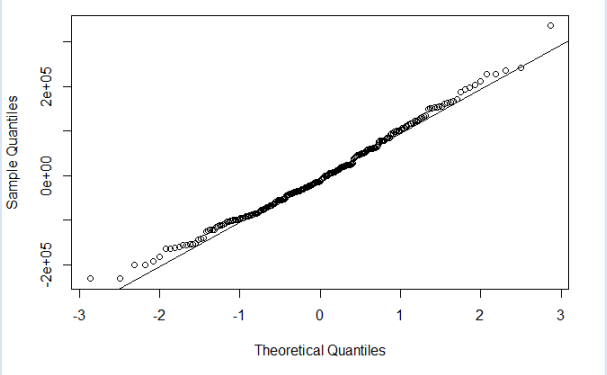

Figura 20. Gráfico Q-Q para las variables género y volumen cerebral completo.

homocedasticidad con el test Levene, con el que se obtiene una homogeneidad de varianzas negativa (p-valor 0.03034).

A partir de aquí se procede con un test T de varianzas separadas, cuyo resumen aparece en la Figura 21. Con este test se obtiene un p-valor de $2{,}2e^{-16}$ y un intervalo de confianza entre 959787.1 y 989219.0 con una probabilidad del 95%. Como conclusión se rechaza la hipótesis nula, indicando que sí existen diferencias significativas entre sexos para el volumen del cerebro completo.

```
 Welch Two Sample t-test

data:  clinic_data$WholeBrain_bl and clinic_data$PTGENDER
t = 130.45, df = 238, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 959787.1 989219.0
sample estimates:
   mean of x    mean of y
9.745035e+05 4.518828e-01
```

Figura 21. Resumen del test T de varianzas separadas para las variables volumen cerebral completo y género.

3.2.1.3. Volumen de los ventrículos

A continuación, se valora el volumen de los ventrículos de los pacientes. En la Figura 16 vuelve a observarse una posible diferencia de volumen entre géneros.

Con un test de normalidad con p-valor 0.06137 (Figura 22), se concluye que las variables siguen una distribución normal. Para valorar la homocedasticidad se realiza un test Levene, con el que se obtiene un p-valor de 0.004346, indicando falta de homogeneidad de varianzas.

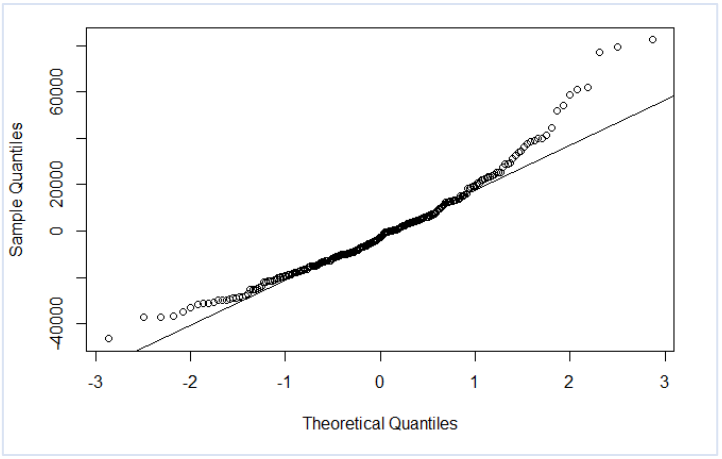

Figura 22. Gráfico Q-Q para las variables género y volumen de los ventrículos.

Debido a estos resultados, se procede con un test T de varianzas separadas, con el que se obtiene un p-valor de $2{,}2e^{-16}$ y un intervalo de confianza entre 45420.80 y

```
 Welch Two Sample t-test

data:  clinic_data$Ventricles_bl and clinic_data$PTGENDER
t = 32.41, df = 238, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 45420.80 51299.71
sample estimates:
   mean of x    mean of y
4.836071e+04 4.518828e-01
```

Figura 23. Resumen del test T de varianzas separadas para las variables ventrículos y género.

51299.71 con un 95% de probabilidad (Figura 23). Con estos resultados, se concluye que sí existen diferencias significativas entre sexos para el volumen de los ventrículos.

3.2.1.4. Volumen de la corteza entorrinal

Por último, se evalúa el volumen de la corteza entorrinal de los pacientes. En la Figura 17 se puede apreciar una diferencia entre géneros.

En este caso vuelve a cumplirse el supuesto de normalidad con un p-valor de 0.5942 (Figura 24). Con el test Leven realizado se obtiene un p-valor de 0.4634, indicando que las muestras sí disponen de homocedasticidad y, por tanto, se procede con un test T para varianzas iguales.

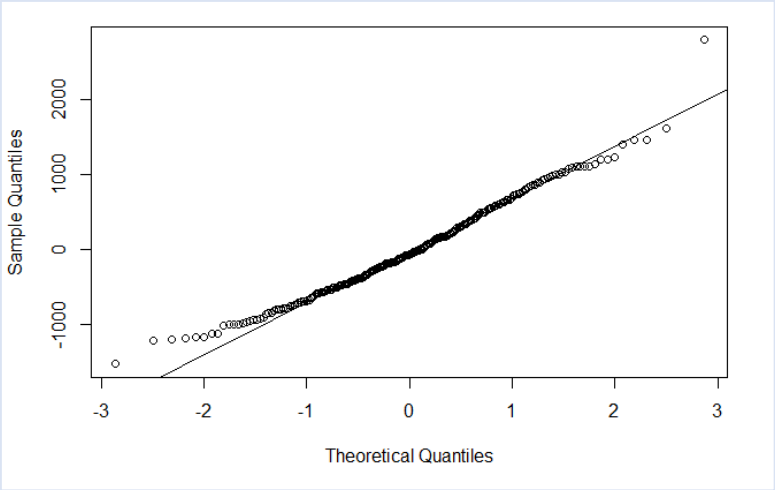

Figura 24. Gráfico Q-Q para las variables género y volumen de la corteza entorrinal.

Con este último test se obtiene un p-valor de $2.2e^{-16}$ y un intervalo de confianza entre 45428.30 y 51292.22 con una probabilidad del 95%, tal y como muestra la Figura 25. Este resultado indica que sí existen diferencias estadísticamente significativas entre sexos para el volumen de corteza entorrinal.

```
 Two Sample t-test

data:  clinic_data$Ventricles_bl and clinic_data$PTGENDER
t = 32.41, df = 476, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 45428.30 51292.22
sample estimates:
   mean of x    mean of y
4.836071e+04 4.518828e-01
```

Figura 25. Resumen del test T de varianzas iguales para las variables corteza entorrinal y género.

3.2.2. Análisis de regresión

Primeramente, se ha realizado un análisis de regresión lineal evaluando las variables edad y volumen de hipocampo. El coeficiente de correlación para ambas variables es -0,3858, lo que indica que sí existe cierta relación negativa entre la edad y el volumen de hipocampo de los pacientes.

Para comprobar si esta posible relación es significativa, se genera el modelo lineal de regresión, resumido en la Figura 26. Según este modelo se obtiene un p-valor de $6.66e^{-10}$ y un coeficiente de determinación de 0.15.

La recta de mínimos cuadrados resultante de este modelo se representa gráficamente en la Figura 27 y su ecuación corresponde a:

$$Y = -0{,}003051X + 91{,}916779$$

```
Call:
lm(formula = Hippocampus_bl ~ AGE, data = clinic_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2216.7  -739.8   -61.3   727.0  3217.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9399.212    566.247  16.599  < 2e-16 ***
AGE          -48.787      7.578  -6.438 6.66e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 932.3 on 237 degrees of freedom
Multiple R-squared:  0.1489, Adjusted R-squared:  0.1453
F-statistic: 41.45 on 1 and 237 DF,  p-value: 6.658e-10
```

Figura 26. Resumen del modelo generado definiendo el volumen del hipocampo como variable respuesta y la edad como variable predictora.

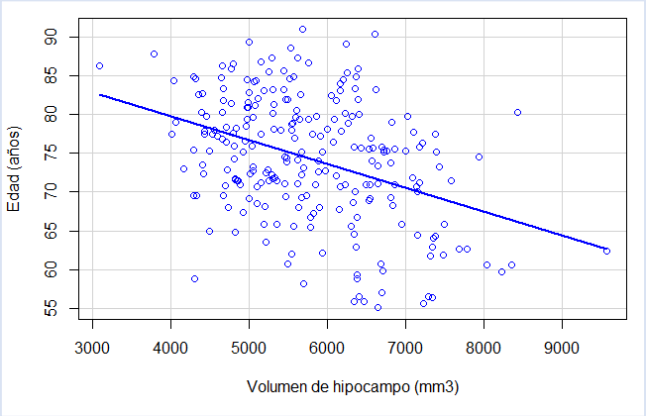

Figura 27. Gráfico de dispersión y recta de mínimos cuadrados para las variables edad y volumen de hipocampo.

Como resultado de este análisis de regresión lineal simple, se obtiene la evidencia de que la edad y el volumen del hipocampo tienen una relación estadísticamente significativa. Observando la Figura 27 se puede terminar de concluir que esta relación se define de manera que a medida que aumenta la edad de los pacientes, el volumen del hipocampo disminuye.

Para finalizar el análisis estadístico se comprueba si alguna otra variable de las presentes en el estudio puede influir en el volumen del hipocampo mediante un análisis de regresión lineal múltiple. Antes de llevarlo a cabo, se observa cuál es la variable que más puede influir en el volumen del hipocampo calculando los coeficientes de correlación para todas las variables que pueden cuantificarse del estudio (Figura 28).

Tal y como se observa, la variable que presenta una mayor relación con el volumen del hipocampo es la que representa el volumen de la corteza entorrinal (Entorhinal_bl).

	Hippocampus_bl
RID	0.06327624
AGE	-0.38581635
PTGENDER	-0.22727006
PTEDUCAT	0.20276556
PTMARRY	-0.18024427
APOE4	-0.10497909
Ventricles_bl	0.03103572
Hippocampus_bl	1.00000000
WholeBrain_bl	0.52389370
Entorhinal_bl	0.60301911

Figura 28. Coeficientes de correlación para la variable volumen del hipocampo y el resto de las variables numéricas del estudio.

Suponiendo todas las variables excepto la del hipocampo como predictores a la hora de generar el modelo de regresión múltiple se obtiene un coeficiente de determinación de 0.52 y un p-valor significativo de $2.2e^{16}$ (Figura 29).

```
Call:
lm(formula = clinic_data$Hippocampus_bl ~ clinic_data$RID + clinic_data$AGE +
    clinic_data$PTGENDER + clinic_data$PTEDUCAT + clinic_data$APOE4 +
    clinic_data$Ventricles_bl + clinic_data$WholeBrain_bl + clinic_data$Entorhinal_bl +
    clinic_data$PTMARRY)

Residuals:
     Min       1Q   Median       3Q      Max
-1876.47  -461.43    23.28   457.11  2195.01

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.272e+03  8.523e+02   3.839  0.00016 ***
clinic_data$RID           -6.255e-01  5.382e-01  -1.162  0.24638
clinic_data$AGE           -2.857e+01  6.712e+00  -4.257 3.02e-05 ***
clinic_data$PTGENDER       6.663e+01  1.178e+02   0.565  0.57235
clinic_data$PTEDUCAT       2.704e+01  1.691e+01   1.599  0.11114
clinic_data$APOE4         -2.190e+02  6.726e+01  -3.256  0.00130 **
clinic_data$Ventricles_bl -6.745e-04  2.274e-03  -0.297  0.76699
clinic_data$WholeBrain_bl  2.951e-03  5.421e-04   5.444 1.34e-07 ***
clinic_data$Entorhinal_bl  5.789e-01  7.909e-02   7.319 4.19e-12 ***
clinic_data$PTMARRY       -1.180e+01  5.439e+01  -0.217  0.82851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 711 on 229 degrees of freedom
Multiple R-squared:  0.5216, Adjusted R-squared:  0.5028
F-statistic: 27.75 on 9 and 229 DF,  p-value: < 2.2e-16
```

Figura 29. Resultado del modelo lineal múltiple definiendo la variable hipocampo como respuesta y todas las variables numéricas restantes como predictoras.

Como resultado final del análisis se obtiene que los mejores predictores para la variable estudiada son, a parte de la edad (relación obtenida también en el análisis de regresión lineal simple), los años de educación, el número de alelos *ApoE4* y el volumen cerebral y de la corteza entorrinal (CE). Por lo tanto, estas son las variables que más influyen en el volumen del hipocampo, siendo el modelo de regresión más significativo el que indica la Figura 30.

```
Call:
lm(formula = clinic_data$Hippocampus_bl ~ clinic_data$AGE + clinic_data$PTEDUCAT +
    clinic_data$APOE4 + clinic_data$WholeBrain_bl + clinic_data$Entorhinal_bl)

Residuals:
     Min       1Q   Median       3Q      Max
-1902.63  -513.35     9.49   484.92  2278.47

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.643e+03  6.999e+02   5.205 4.25e-07 ***
clinic_data$AGE          -3.047e+01  6.083e+00  -5.010 1.08e-06 ***
clinic_data$PTEDUCAT      2.299e+01  1.615e+01   1.424 0.155860
clinic_data$APOE4        -2.207e+02  6.613e+01  -3.337 0.000986 ***
clinic_data$WholeBrain_bl  2.691e-03  4.517e-04   5.957 9.41e-09 ***
clinic_data$Entorhinal_bl  5.766e-01  7.760e-02   7.430 2.04e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 707.5 on 233 degrees of freedom
Multiple R-squared:  0.5181, Adjusted R-squared:  0.5078
F-statistic: 50.11 on 5 and 233 DF,  p-value: < 2.2e-16
```

Figura 30. Modelo de regresión más significativo para la variable respuesta volumen de hipocampo.

3.3. Análisis de expresión

A continuación, se muestran los resultados obtenidos en los análisis de expresión diferencial realizados en el hipocampo y en la corteza entorrinal.

3.3.1. Hipocampo

En el hipocampo se trabaja con un total de 36 muestras, con 9 de ellas en cada grupo.

En primer lugar, se muestran los resultados de la exploración visual de los datos antes del control de calidad y la normalización. Se puede observar tanto en la Figura 31 como en la Figura 33 que las muestras presentan heterogeneidad y además sugieren algún problema en los datos ya que alguna de las muestras se desvía más que el resto. Esto se confirma con los resultados surgidos en el control de calidad, que pueden observarse de forma resumida en la Figura 32 o de forma completa en la carpeta anexada "AED/Hip/results/QCDir.Raw". Como se puede ver, el control de calidad detecta varias muestras (sobre todo de mujeres con la enfermedad de Alzheimer) como *outliers.* No

obstante, únicamente una de ellas es detectada por los tres métodos aplicados: FEM.AD.5, que resulta ser la misma muestra que se observa desviada en las figuras anteriormente comentadas.

El gráfico PCA representado en la Figura 34 ofrece información sobre la variabilidad de las muestras y, como se observa, el primer componente engloba el 67% de la variabilidad que, principalmente es generada por las mujeres con la enfermedad. En este gráfico también puede apreciarse como la muestra FEM.AD.5 se encuentra alejada del resto.

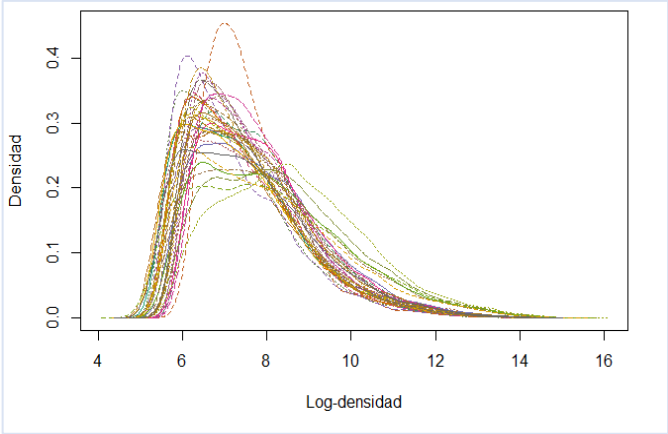

Figura 31. Gráfico de densidad de los datos crudos para el hipocampo.

	array	sampleNames	*1	*2	*3	Group	Gender	Genome	ShortName
☐	1	FEM.CT.1				FEM.CT	FEM	CT	FEM.CT.1
☐	2	FEM.CT.2				FEM.CT	FEM	CT	FEM.CT.2
☐	3	FEM.CT.3			x	FEM.CT	FEM	CT	FEM.CT.3
☐	4	FEM.CT.4				FEM.CT	FEM	CT	FEM.CT.4
☐	5	FEM.CT.5				FEM.CT	FEM	CT	FEM.CT.5
☐	6	FEM.CT.6			x	FEM.CT	FEM	CT	FEM.CT.6
☐	7	FEM.CT.7				FEM.CT	FEM	CT	FEM.CT.7
☐	8	FEM.CT.8				FEM.CT	FEM	CT	FEM.CT.8
☐	9	FEM.CT.9				FEM.CT	FEM	CT	FEM.CT.9
☐	10	MALE.CT.1				MALE.CT	MALE	CT	MALE.CT.1
☐	11	MALE.CT.2			x	MALE.CT	MALE	CT	MALE.CT.2
☐	12	MALE.CT.3			x	MALE.CT	MALE	CT	MALE.CT.3
☐	13	MALE.CT.4				MALE.CT	MALE	CT	MALE.CT.4
☐	14	MALE.CT.5				MALE.CT	MALE	CT	MALE.CT.5
☐	15	MALE.CT.6				MALE.CT	MALE	CT	MALE.CT.6
☐	16	MALE.CT.7			x	MALE.CT	MALE	CT	MALE.CT.7
☐	17	MALE.CT.8				MALE.CT	MALE	CT	MALE.CT.8
☐	18	MALE.CT.9				MALE.CT	MALE	CT	MALE.CT.9
☐	19	FEM.AD.1	x		x	FEM.AD	FEM	AD	FEM.AD.1
☐	20	FEM.AD.2	x		x	FEM.AD	FEM	AD	FEM.AD.2
☐	21	FEM.AD.3	x		x	FEM.AD	FEM	AD	FEM.AD.3
☐	22	FEM.AD.4	x		x	FEM.AD	FEM	AD	FEM.AD.4
☐	23	FEM.AD.5	x	x	x	FEM.AD	FEM	AD	FEM.AD.5
☐	24	FEM.AD.6				FEM.AD	FEM	AD	FEM.AD.6
☐	25	FEM.AD.7				FEM.AD	FEM	AD	FEM.AD.7
☐	26	FEM.AD.8				FEM.AD	FEM	AD	FEM.AD.8
☐	27	FEM.AD.9	x		x	FEM.AD	FEM	AD	FEM.AD.9
☐	28	MALE.AD.1				MALE.AD	MALE	AD	MALE.AD.1
☐	29	MALE.AD.2				MALE.AD	MALE	AD	MALE.AD.2
☐	30	MALE.AD.3				MALE.AD	MALE	AD	MALE.AD.3
☐	31	MALE.AD.4				MALE.AD	MALE	AD	MALE.AD.4
☐	32	MALE.AD.5				MALE.AD	MALE	AD	MALE.AD.5
☐	33	MALE.AD.6				MALE.AD	MALE	AD	MALE.AD.6
☐	34	MALE.AD.7			x	MALE.AD	MALE	AD	MALE.AD.7
☐	35	MALE.AD.8				MALE.AD	MALE	AD	MALE.AD.8
☐	36	MALE.AD.9				MALE.AD	MALE	AD	MALE.AD.9

*Figura 32. Resumen del control de calidad de las muestras del hipocampo. Las columnas: *1 corresponde al método de detección por distancias entre arrays; *2 corresponde al método de detección por diagrama de cajas; *3 corresponde al método de detección por gráficos MA. Las muestras detectadas como "outliers" son marcadas con una "X" en el método correspondiente.*

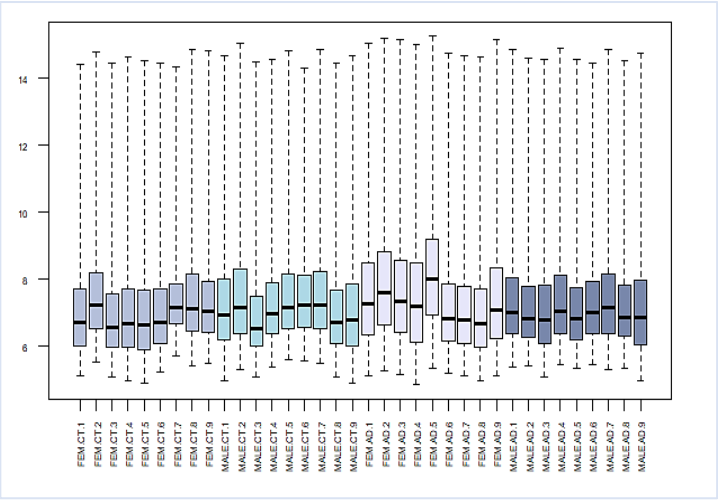

Figura 33. Diagrama de cajas de los datos crudos para el hipocampo.

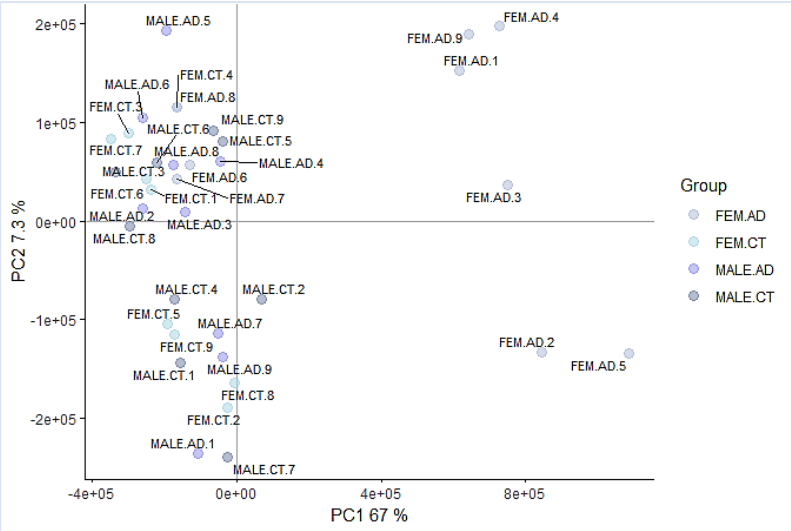

Figura 34. Gráfico PCA de los datos crudos para el hipocampo.

Debido a estos resultados obtenidos, la muestra número 23 es excluida del análisis.

Se muestran ahora los mismos gráficos generados anteriormente, a partir de los datos ya normalizados y la muestra problemática eliminada. Tanto la Figura 35 como la Figura 36, confirman que la normalización de los datos se ha llevado a cabo adecuadamente y que ahora las muestras son más homogéneas y cumplen las condiciones necesarias para continuar con el análisis de forma correcta. En el gráfico PCA en este caso, la variabilidad del primer componente se ha visto disminuida, viéndose las muestras más separadas, pero sin ningún patrón observable a simple vista (Figura 37).

Además, el control de calidad de los datos normalizados es satisfactorio para todas las muestras ya que únicamente se detectan como *outliers* dos muestras por sólo uno de los tres métodos (resultados en la carpeta anexada "AED/Hip/results/QCDir.Norm").

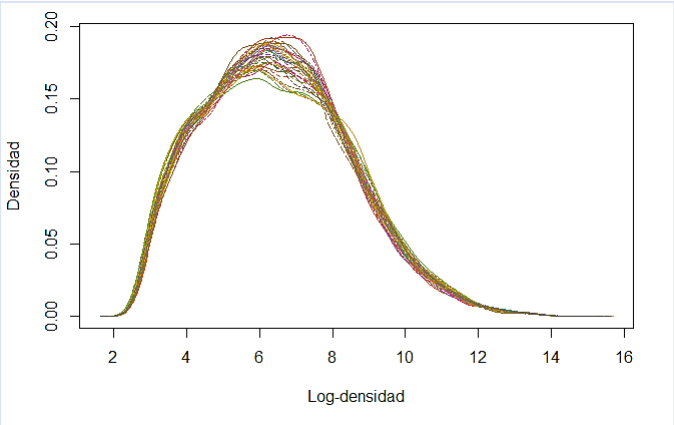

Figura 35. Gráfico de densidad de los datos normalizados para el hipocampo.

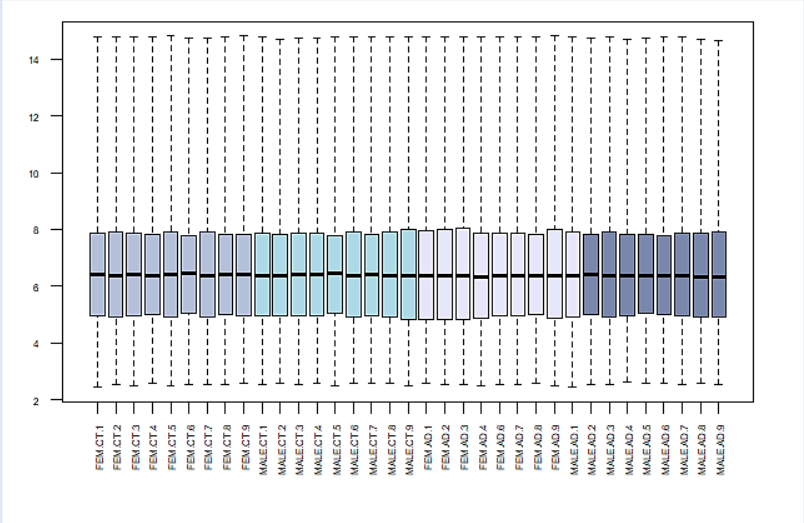

Figura 36. Diagrama de cajas de los datos normalizados para el hipocampo.

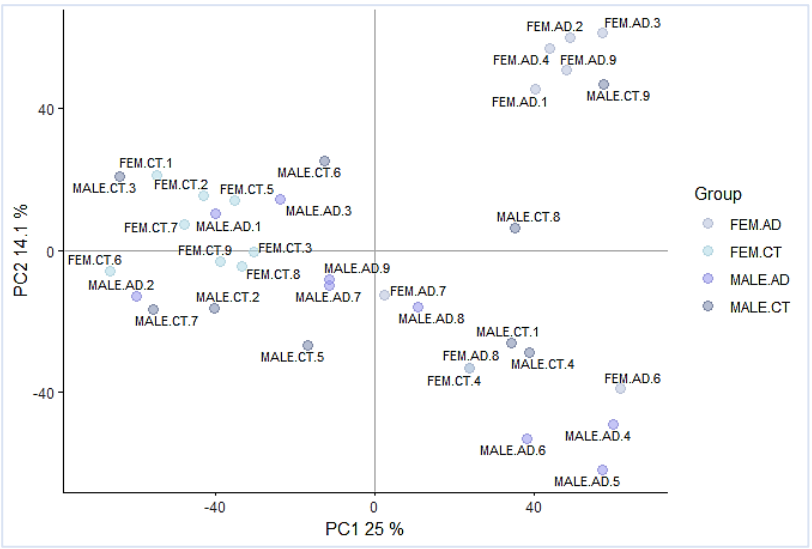

Figura 37. Gráfico PCA de los datos normalizados para el hipocampo.

Como resultado de la selección de genes y su anotación se obtienen dos tablas, una para cada comparación, con los genes ordenados de menor a mayor p-valor. La tabla anexada "AED/Hip/results/topAnnotated_FEMvsMALE.CT" contiene los genes seleccionados para el contraste de hombres y mujeres control. La tabla anexada "AED/Hip/results/topAnnotated_FEMvsMALE.AD", para el contraste de hombres y mujeres con Alzheimer.

Parte de los resultados de la selección de genes puede observarse en los *volcano plots*, los cuales muestran los cinco genes más diferencialmente expresados de cada contraste, tomando como referencia el *p-valor* y el *LFC.* En la Figura 38, correspondiente a los individuos control, se observa cómo no hay ningún gen que supere los valores umbrales definidos en el filtraje. En la Figura 39, en cambio, sí que se detectan genes que superan estos umbrales.

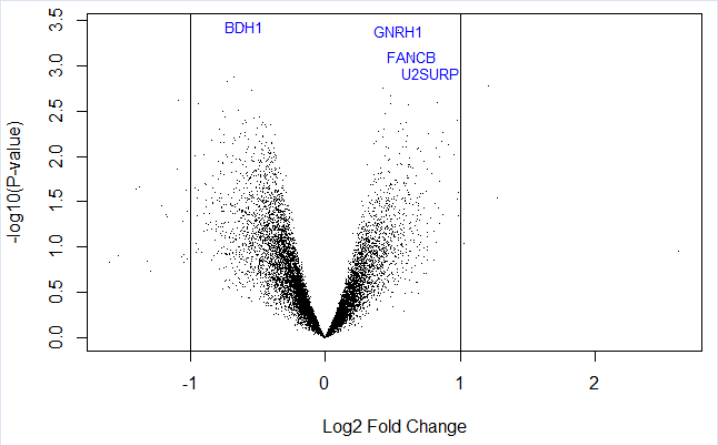

Figura 38. Genes más diferencialmente expresados en el hipocampo de pacientes control.

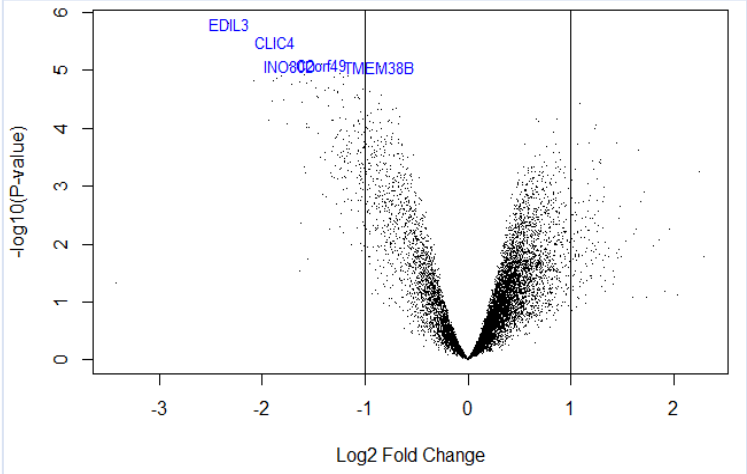

Figura 39. Genes más diferencialmente expresados en el hipocampo de pacientes con Alzheimer.

La selección génica se reduce aún más después de la comparación múltiple, que genera como resultado la Tabla 14. Tal y como indicaban los *volcano plots,* en el contraste de personas control no se obtiene ningún gen diferencial, mientras que en el contraste de individuos enfermos se obtienen 150 genes subexpresados y 43 genes sobreexpresados. Por tanto, y tal y como muestra la Figura 40, ninguno de los 193 genes seleccionados es compartido por los dos contrastes.

Tabla 14. Resumen de la comparación múltiple. Número de genes diferencialmente expresados en el hipocampo definitivos.

	FEMvsMALE.CT	FEMvsMALE.AD
Down	0	150
NotSig	10080	9887
Up	0	43

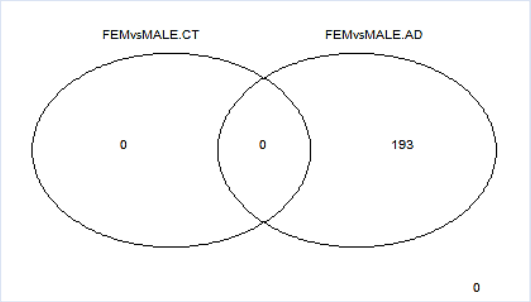

Figura 40. Genes diferenciales en común en cada contraste para el hipocampo.

La Figura 41 muestra el *heatmap* generado a partir de la expresión de los genes seleccionados. Como se observa, parece haber un patrón en el que gran parte de las muestras de mujeres con la enfermedad, presentan una expresión génica muy similar, la cual es prácticamente contraria a la del resto de muestras. Además, parecen haber otros patrones menos claros a simple vista.

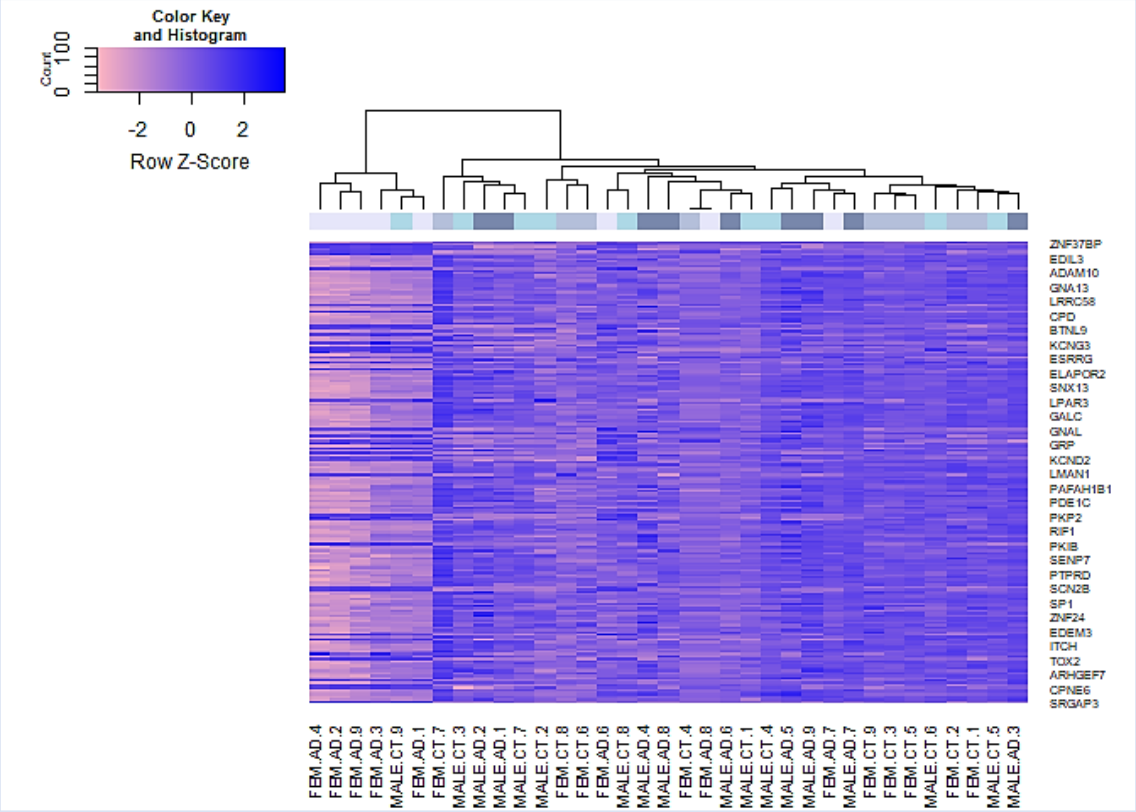

Figura 41. Heatmap de los genes seleccionados como diferencialmente expresados en el hipocampo, con las muestras agrupadas según su similitud.

Los procesos biológicos más representados de esta lista de genes se muestran en la Figura 42. La regulación y el desarrollo del sistema inmune junto con procesos biológicos reproductivos son los más destacados en este conjunto de genes.

Como resultado del análisis de *pathways* realizado a partir de la base de datos KEGG, se obtiene como más representada la vía de señalización AMPK. Los genes más representativos involucrados en esta vía se muestran en la Figura 43.

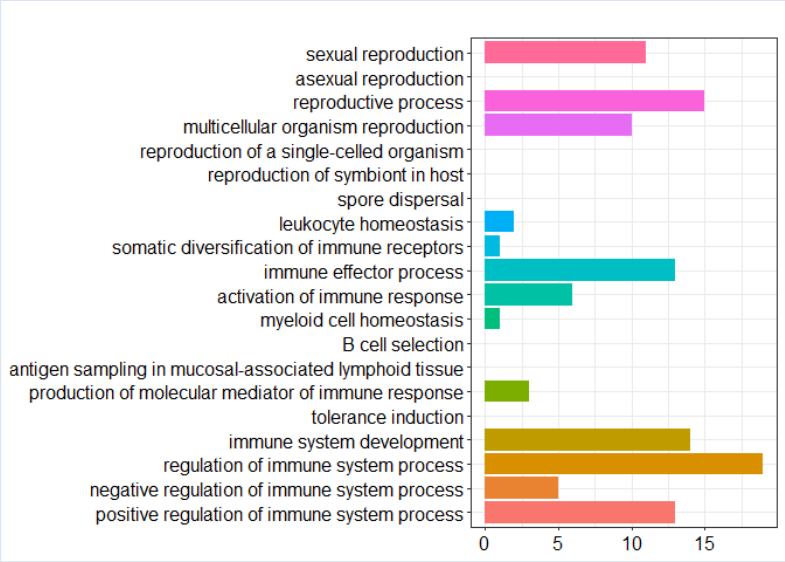

Figura 42. Agrupación de los genes seleccionados como diferenciales en el hipocampo según los procesos biológicos en los que participan.

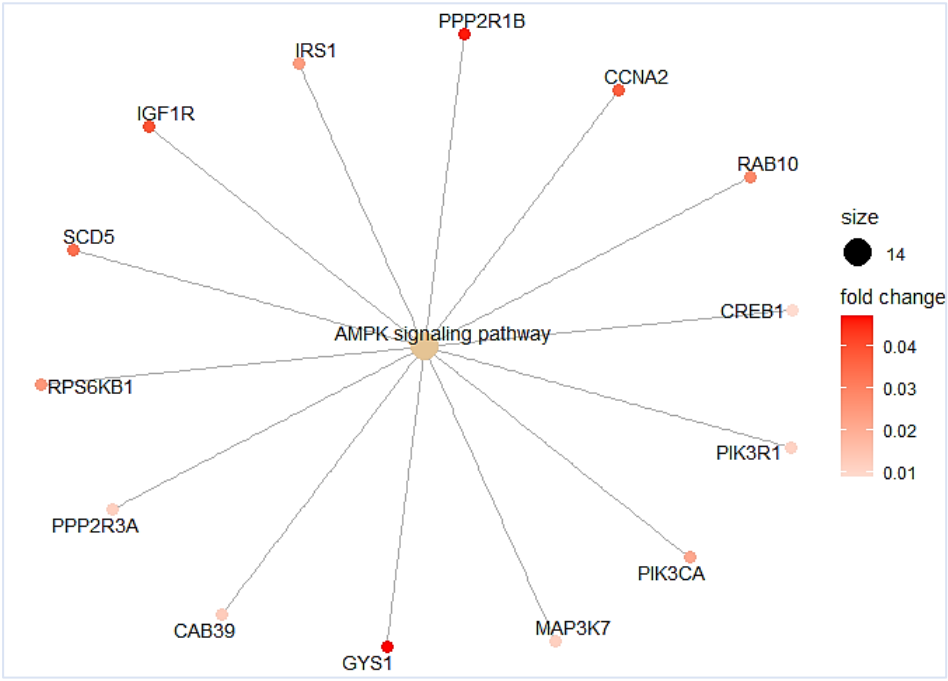

Figura 43. Resultado del análisis de pathways realizado a partir de KEGG en el hipocampo. Vía más representativa junto con los genes involucrados en ella.

Los resultados para el análisis de *pathways* mediante ReactomePA se encuentran en la Figura 44. En este caso, las vías más representativas son el metabolismo y la síntesis de

lípidos, junto con la señalización de EGFR, entre otras. Los genes involucrados en algunas de estas vías y la relación entre ellas se pueden apreciar en la Figura 45.

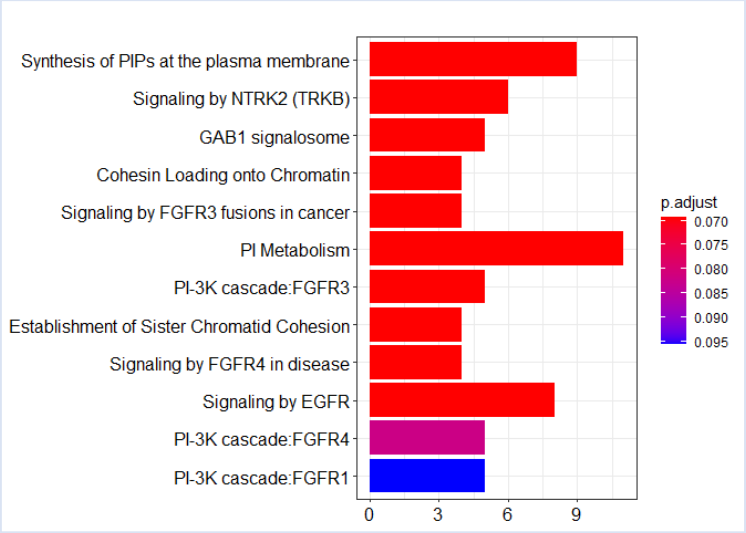

Figura 44. Resultados del análisis de pathways realizado a partir de ReactomePA en el hipocampo. Vías más representativas.

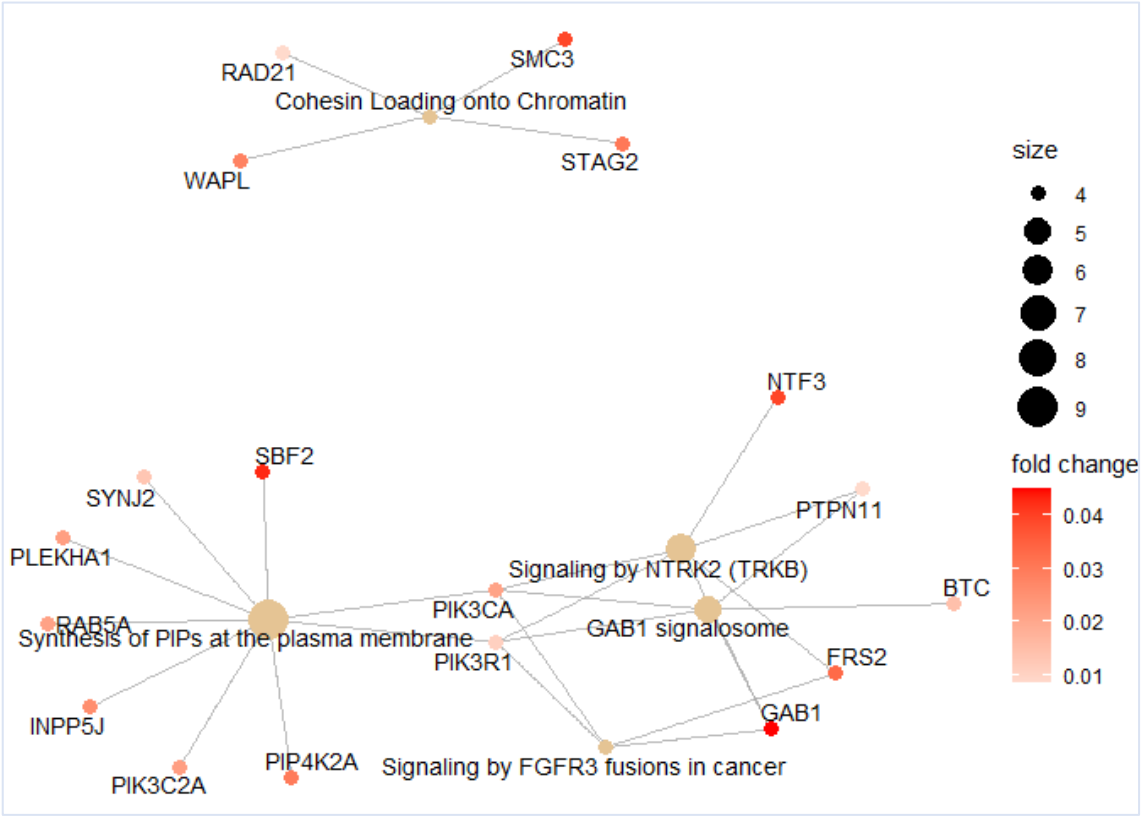

Figura 45. Resultados del análisis de pathways realizado a partir de ReactomePA en el hipocampo. Genes involucrados en algunas de las vías.

De estos 193 genes finalmente seleccionados, los 5 más diferencialmente expresados teniendo en cuenta el menor *p-valor* ajustado son los mostrados en la Tabla 15. Además, en esta tabla se contempla la localización y las coordenadas de cada gen.

Tabla 15. Los cinco genes más diferencialmente expresados como resultado final del análisis de expresión en el hipocampo junto con su localización en el genoma.

SÍMBOLO	NOMBRE	LOCALIZACIÓN
SP1	Sp1 transcription factor	12q13.13
MBNL1	Muscleblind like splicing regulator 1	3q25.1-q25.2
ALCAM	Activated leukocyte cell adhesion molecule	3q13.11
RAD21	RAD21 cohesin complex component	8q24.11
CLIC4	Chloride intracellular channel 4	1p36.11

3.3.2. Corteza entorrinal

En la corteza entorrinal se trabaja con un total de 28 muestras, con 7 en cada grupo.

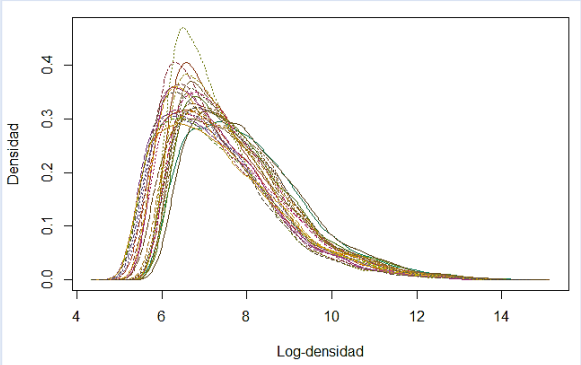

Figura 46. Gráfico de densidad de los datos crudos para la corteza entorrinal.

Los resultados de la exploración visual de los datos antes del control de calidad y la normalización se muestran a continuación. En primer lugar, la Figura 46 y la Figura 47 reflejan la heterogeneidad inicial de las muestras. Aunque la primera figura puede insinuar la presencia de una muestra problemática, la segunda figura y el control de calidad (Figura 48 o "AED/CE/results/QCDir.Raw") realizado posteriormente no detectan ninguna muestra desviada. En segundo lugar, el gráfico PCA representado en la Figura 49 presenta una distribución de las muestras separadas y con una variabilidad del primer componente baja (37,3%).

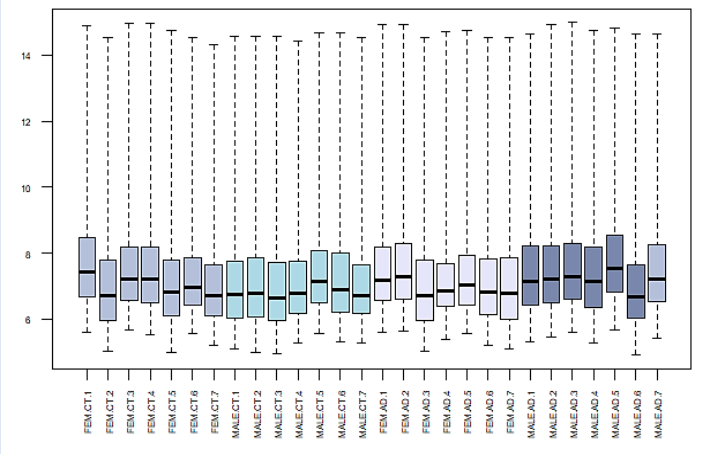

Figura 47. Diagrama de cajas de los datos crudos para la corteza entorrinal.

	array	sampleNames	*1	*2	*3	Group	Gender	Genome	ShortName
☐	1	FEM.CT.1	x		x	FEM.CT	FEM	CT	FEM.CT.1
☐	2	FEM.CT.2				FEM.CT	FEM	CT	FEM.CT.2
☐	3	FEM.CT.3				FEM.CT	FEM	CT	FEM.CT.3
☐	4	FEM.CT.4			x	FEM.CT	FEM	CT	FEM.CT.4
☐	5	FEM.CT.5				FEM.CT	FEM	CT	FEM.CT.5
☐	6	FEM.CT.6				FEM.CT	FEM	CT	FEM.CT.6
☐	7	FEM.CT.7			x	FEM.CT	FEM	CT	FEM.CT.7
☐	8	MALE.CT.1				MALE.CT	MALE	CT	MALE.CT.1
☐	9	MALE.CT.2				MALE.CT	MALE	CT	MALE.CT.2
☐	10	MALE.CT.3				MALE.CT	MALE	CT	MALE.CT.3
☐	11	MALE.CT.4			x	MALE.CT	MALE	CT	MALE.CT.4
☐	12	MALE.CT.5				MALE.CT	MALE	CT	MALE.CT.5
☐	13	MALE.CT.6				MALE.CT	MALE	CT	MALE.CT.6
☐	14	MALE.CT.7			x	MALE.CT	MALE	CT	MALE.CT.7
☐	15	FEM.AD.1				FEM.AD	FEM	AD	FEM.AD.1
☐	16	FEM.AD.2			x	FEM.AD	FEM	AD	FEM.AD.2
☐	17	FEM.AD.3				FEM.AD	FEM	AD	FEM.AD.3
☐	18	FEM.AD.4			x	FEM.AD	FEM	AD	FEM.AD.4
☐	19	FEM.AD.5				FEM.AD	FEM	AD	FEM.AD.5
☐	20	FEM.AD.6				FEM.AD	FEM	AD	FEM.AD.6
☐	21	FEM.AD.7				FEM.AD	FEM	AD	FEM.AD.7
☐	22	MALE.AD.1			x	MALE.AD	MALE	AD	MALE.AD.1
☐	23	MALE.AD.2			x	MALE.AD	MALE	AD	MALE.AD.2
☐	24	MALE.AD.3			x	MALE.AD	MALE	AD	MALE.AD.3
☐	25	MALE.AD.4			x	MALE.AD	MALE	AD	MALE.AD.4
☐	26	MALE.AD.5	x		x	MALE.AD	MALE	AD	MALE.AD.5
☐	27	MALE.AD.6				MALE.AD	MALE	AD	MALE.AD.6
☐	28	MALE.AD.7			x	MALE.AD	MALE	AD	MALE.AD.7

*Figura 48. Resumen del control de calidad de las muestras de la corteza entorrinal. Las columnas: *1 corresponde al método de detección por distancias entre arrays; *2 corresponde al método de detección por diagrama de cajas; *3 corresponde al método de detección por gráficos MA.*

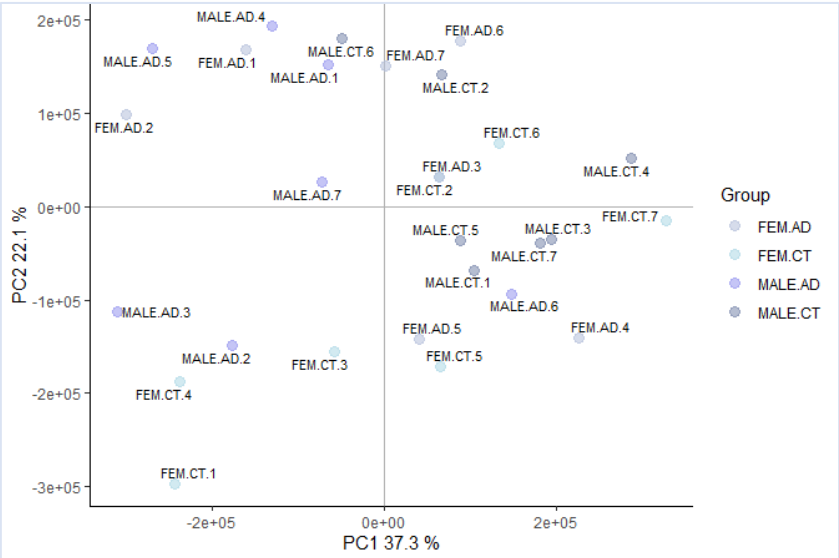

Figura 49. Gráfico PCA de los datos crudos para la corteza entorrinal.

Con los resultados obtenidos, se decide no eliminar ninguna muestra debido a que ninguna de ellas se ha detectado como *outlier* en los 3 criterios del control de calidad.

Seguidamente, se muestran los mismos gráficos generados anteriormente, ahora con los datos después de la normalización (Figuras 50-52). Como puede apreciarse, la normalización se ha realizado correctamente ya que las muestras presentan una

distribución más homogénea. El gráfico PCA (Figura 53), en cambio, no muestra modificaciones importantes.

De nuevo, el control de calidad de los datos normalizados confirma que la normalización se ha llevado a cabo correctamente con solo tres muestras detectadas como *outliers* por el criterio número dos ("AED/CE/results/QCDir.Norm").

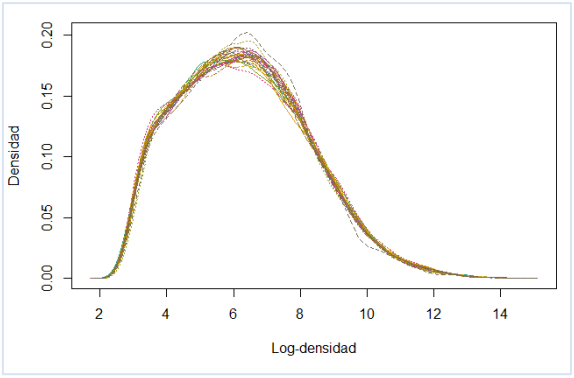

Figura 50. Gráfico de densidad de los datos normalizados para la corteza entorrinal.

	array	sampleNames	*1	*2	*3	Group	Gender	Genome	ShortName
☐	1	FEM.CT.1				FEM.CT	FEM	CT	FEM.CT.1
☐	2	FEM.CT.2				FEM.CT	FEM	CT	FEM.CT.2
☐	3	FEM.CT.3				FEM.CT	FEM	CT	FEM.CT.3
☐	4	FEM.CT.4				FEM.CT	FEM	CT	FEM.CT.4
☐	5	FEM.CT.5				FEM.CT	FEM	CT	FEM.CT.5
☐	6	FEM.CT.6				FEM.CT	FEM	CT	FEM.CT.6
☐	7	FEM.CT.7				FEM.CT	FEM	CT	FEM.CT.7
☐	8	MALE.CT.1				MALE.CT	MALE	CT	MALE.CT.1
☐	9	MALE.CT.2				MALE.CT	MALE	CT	MALE.CT.2
☐	10	MALE.CT.3				MALE.CT	MALE	CT	MALE.CT.3
☐	11	MALE.CT.4				MALE.CT	MALE	CT	MALE.CT.4
☐	12	MALE.CT.5				MALE.CT	MALE	CT	MALE.CT.5
☐	13	MALE.CT.6		x		MALE.CT	MALE	CT	MALE.CT.6
☐	14	MALE.CT.7				MALE.CT	MALE	CT	MALE.CT.7
☐	15	FEM.AD.1				FEM.AD	FEM	AD	FEM.AD.1
☐	16	FEM.AD.2		x		FEM.AD	FEM	AD	FEM.AD.2
☐	17	FEM.AD.3				FEM.AD	FEM	AD	FEM.AD.3
☐	18	FEM.AD.4		x		FEM.AD	FEM	AD	FEM.AD.4
☐	19	FEM.AD.5				FEM.AD	FEM	AD	FEM.AD.5
☐	20	FEM.AD.6				FEM.AD	FEM	AD	FEM.AD.6
☐	21	FEM.AD.7		x		FEM.AD	FEM	AD	FEM.AD.7
☐	22	MALE.AD.1				MALE.AD	MALE	AD	MALE.AD.1
☐	23	MALE.AD.2				MALE.AD	MALE	AD	MALE.AD.2
☐	24	MALE.AD.3				MALE.AD	MALE	AD	MALE.AD.3
☐	25	MALE.AD.4				MALE.AD	MALE	AD	MALE.AD.4
☐	26	MALE.AD.5				MALE.AD	MALE	AD	MALE.AD.5
☐	27	MALE.AD.6		x		MALE.AD	MALE	AD	MALE.AD.6
☐	28	MALE.AD.7				MALE.AD	MALE	AD	MALE.AD.7

*Figura 51. Resumen del control de calidad de los datos normalizados para la corteza entorrinal. Las columnas: *1 corresponde al método de detección por distancias entre arrays; *2 corresponde al método de detección por diagrama de cajas; *3 corresponde al método de detección por gráficos MA.*

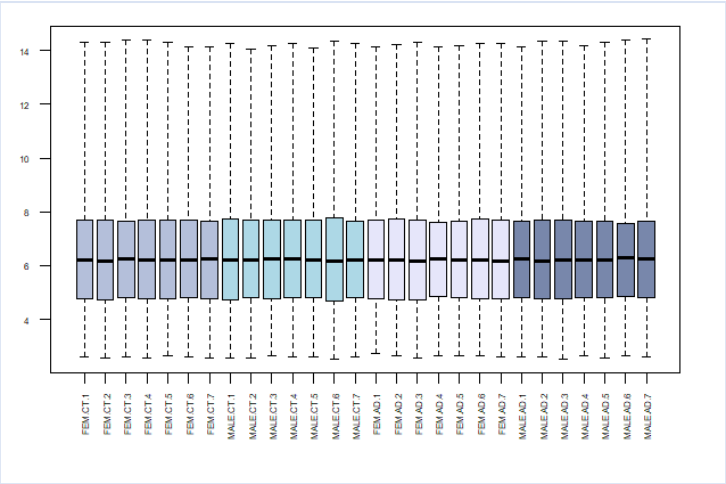

Figura 52. Diagrama de cajas de los datos normalizados.

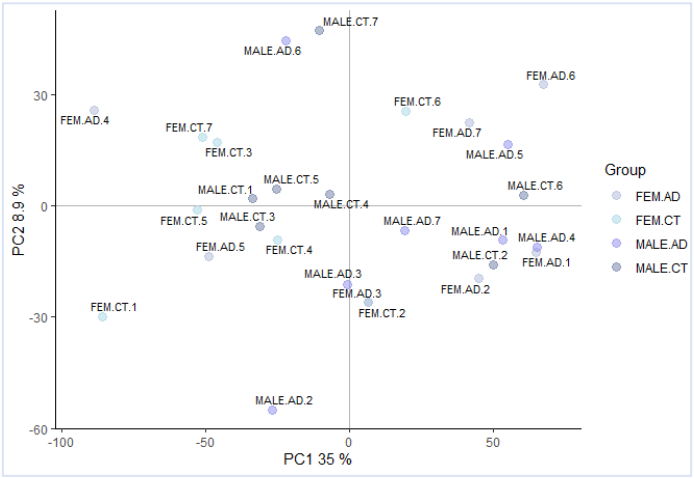

Figura 53. Gráfico PCA de los datos normalizados para la corteza entorrinal.

Las tablas con los genes seleccionados como diferencialmente expresados de cada contraste se encuentran en los archivos "AED/CE/results/topAnnotated_FEMvsMALE.CT" para el contraste de individuos control y "AED/CE/results/topAnnotated_FEMvsMALE.AD" para el contraste de individuos enfermos. La representación gráfica de estas tablas se muestra en las Figuras 54 y 55. Tanto en las tablas como en las figuras puede apreciarse como ningún gen tiene un *p-valor* inferior a 0,05 y un LFC de 1 para ser considerados diferenciales significativamente. A pesar de esto, se realiza la comparación múltiple con los mismos parámetros estadísticos que se tomaron para el hipocampo (*p-valor* ajustado de 0,05 y LFC de 1), pero no se obtiene ningún gen diferencialmente expresado en ninguno de los contrastes. Para intentar rebajar la restricción del filtraje en la comparación, se realiza una segunda comparación múltiple, esta vez con un *p-valor* ajustado de 0,1 y un LFC de 1. En este segundo intento, tampoco se obtienen genes diferencialmente expresados.

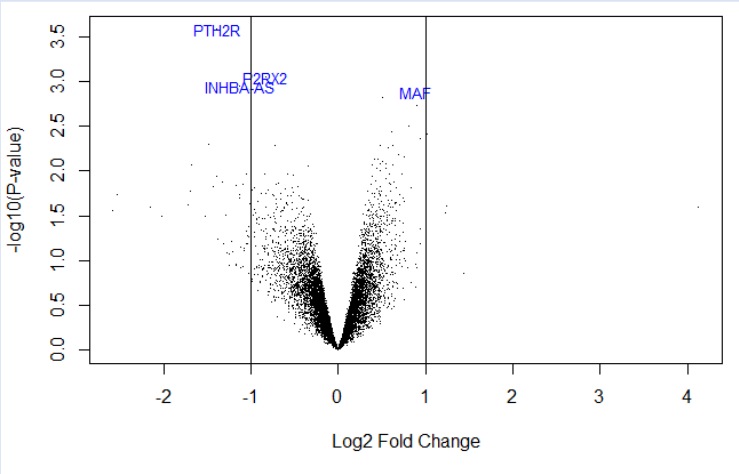

Figura 54. Genes más diferencialmente expresados en pacientes control.

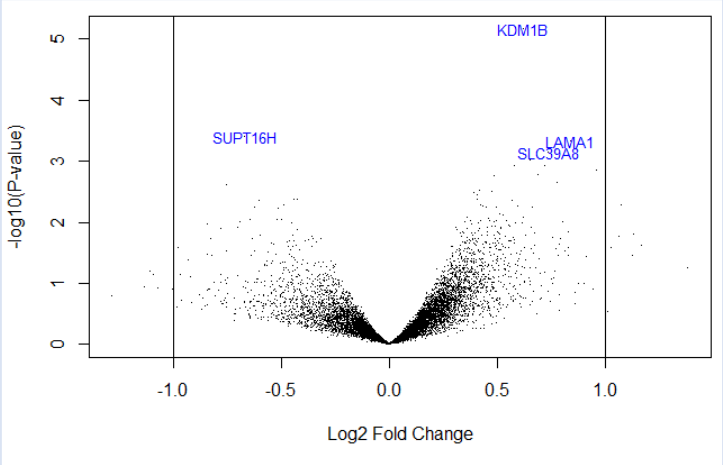

Figura 55. Genes más diferencialmente expresados en pacientes con Alzheimer.

4. Discusión

4.1. Análisis estadístico

La primera evidencia estadística obtenida a partir del análisis descriptivo de los datos es la diferencia de volumen cerebral entre hombres y mujeres con Alzheimer. En este caso, todo indica que los hombres disponen de un mayor volumen cerebral que las mujeres. En concreto, esto se observa tanto en el volumen cerebral completo como en algunas de sus partes: el hipocampo, los ventrículos y la corteza entorrinal. Esta evidencia es apoyada por el análisis estadístico llevado a cabo a través del contraste de hipótesis, demostrando esta diferencia de volumen entre géneros de forma significativa. Esta conclusión también se obtiene en otros estudios donde se demuestra la diferencia entre sexos en individuos con la enfermedad de Alzheimer [22,23].

La segunda evidencia de este proyecto se ha obtenido a partir del análisis de regresión lineal simple, que ha resultado en una relación significativa entre la edad de los pacientes y el volumen del hipocampo. Concretamente se observa una disminución del volumen del hipocampo a medida que aumenta la edad de los individuos con Alzheimer. Esto se ha demostrado también en otros estudios como el realizado por L. Nobis y sus compañeros[22], que aparte de observar una disminución del hipocampo más drástica en mujeres que en hombres, también encuentran esta relación con la edad. Por su parte, L. Apostova y sus compañeros[24], también observaron esta disminución del volumen del hipocampo y de los ventrículos laterales con la edad, pero en este caso, la condición se halló tanto en individuos con Alzheimer como en individuos sanos.

Otras observaciones se han documentado en cuanto a cambios que se dan con la edad tanto en personas sanas como en personas con Alzheimer. En el estudio dirigido por N. Berchtold[25], por ejemplo, se afirma que la expresión de algunos genes sinápticos disminuye a medida que el individuo sano envejece. Además, en personas con la enfermedad, esta respuesta es aún más drástica.

Por último, el análisis de regresión múltiple indica que, a parte de la edad, hay otras variables que influyen en el volumen del hipocampo, como los años de educación, el número de alelos *ApoE4,* el volumen cerebral completo y el volumen de la corteza entorrinal. En este proyecto, estas variables no eran las de especial interés, por lo que estos resultados no se han contrastado con pruebas estadísticas o análisis adicionales. Para comprender y evidenciar mejor estas relaciones sería oportuno realizar un estudio más exhaustivo.

El análisis estadístico ha sido realizado a partir de datos donde todos los individuos incluidos tenían Alzheimer. Hubiera sido conveniente, quizás, incluir también datos de individuos sanos para poder valorar mejor las diferencias y las relaciones que se dan en los pacientes afectados por la enfermedad. Asimismo, se conoce que muchos otros datos clínicos influyen en las diferencias entre géneros para esta enfermedad [8] y no se están teniendo en cuenta en este proyecto, pero siguen siendo de vital importancia para un

correcto entendimiento de las disparidades derivadas de esta enfermedad. Estos aspectos podrían considerarse en futuras investigaciones.

4.2. Análisis de expresión diferencial

La evidencia clave resultante del análisis de expresión ha sido la identificación de genes diferencialmente expresados en el hipocampo. Con esto se ha podido demostrar que, en los datos analizados, los individuos con Alzheimer presentan genes en el hipocampo que se expresan de forma distinta según su género. No obstante, no ocurre lo mismo en el caso de la corteza entorrinal, en la cual no se han detectado genes diferenciales significativos.

Los genes identificados como más diferencialmente expresados se encuentran en la Tabla 15. El primero de ellos es *Sp1.* La proteína que codifica este gen es un factor de transcripción que se expresa en el cerebro. Participa en procesos como el crecimiento celular, la apoptosis, la diferenciación y la respuesta inmune[26]. Este gen ya ha sido también relacionado con la enfermedad de Alzheimer en anteriores estudios[27,28].

MBNL1 es el segundo gen más diferencialmente expresado en el hipocampo según el análisis realizado. La proteína se encarga de mediar la regulación del *splicing* alternativo de pre-mRNAs actuando como represor o activador[29]. Este gen se ha relacionado en otras investigaciones con la distrofia miotónica de tipo I [30,31], que es un tipo de distrofia muscular que causa daño multiorgánico[32]. Esta distrofia tiene ya algunas proteínas afectadas en común con el Alzheimer, como es el caso de la proteína tau[33]. MBNL1 también podría ser una proteína en común entre ambas patologías, aunque no se han encontrado estudios que demuestren esta relación directa de MBNL1 con el Alzheimer.

El tercer gen identificado es *ALCAM,* cuya proteína está involucrada en la neuroinflamación y la homeostasis del cerebro. A pesar de su estrecha relación con la función cerebral, no se ha localizado ningún estudio que relacione este gen con el Alzheimer.

Los principales genes identificados como diferencialmente expresados en este proyecto no son genes con una conocida relación con la enfermedad, por lo que sería necesario más estudios de este tipo para comprobar que realmente están involucrados y que se encuentran expresados diferencialmente en hombres y mujeres con Alzheimer.

La última evidencia surgida de este proyecto son las vías más representativas resultantes del análisis de pathways. Las principales vías detectadas han sido AMPK y EGFR. Ambas vías son conocidas por tener un papel importante en el riesgo a padecer la enfermedad de Alzheimer[34,35]. Sin embargo, los genes más diferenciales no están involucrados en estas vías, por lo que, aunque el resultado obtenido del análisis de pathways tiene sentido según la información ya conocida, no resulta especialmente relevante en este trabajo, puesto que no permite comprender la relación de estas vías con la expresión diferencial según el género de los pacientes.

Este proyecto no es el único que ha encontrado diferencias específicas de sexo a nivel de expresión génica asociadas al Alzheimer. Además, también se han encontrado vías

moleculares sexo-específicas para la enfermedad, así como mecanismos moleculares distintos según el género de los pacientes[11].

La mayoría de las investigaciones llevadas a cabo para valorar las disparidades genéticas entre géneros en la enfermedad de Alzheimer, son realizadas a partir de estudios de asociación de genoma completo o GWAS, con el fin de encontrar polimorfismos de un solo nucleótido (SNPs) que sirvan como posibles dianas para la enfermedad. Aunque para este proyecto se haya decidido llevar a cabo análisis de expresión en lugar de GWAS, con este tipo de estudios también se identifican muchísimos genes involucrados en estas disparidades entre géneros [36,37]. Tanto en estos estudios GWAS como en otros estudios de expresión realizados[38], los genes detectados y relacionados con el Alzheimer y sus disparidades de género son muy variables. Por este motivo es necesario llevar a cabo este tipo de estudios en el que se diferencie el sexo de los individuos para poder comprender mejor la enfermedad y llegar a un consenso de cuáles son los genes principales involucrados en estas diferencias. Además, también es importante valorar la expresión en las diferentes zonas cerebrales, puesto que en este caso únicamente se han tenido en cuenta dos de ellas y solamente en una se han obtenido genes diferenciales.

4.3. Estudio comparativo

Al comparar los dos tipos de análisis llevados a cabo en el proyecto, se puede observar una relación clara en los resultados obtenidos con los datos utilizados. Tanto el análisis de expresión diferencial como el análisis estadístico indican una disparidad clara entre sexos en el hipocampo de individuos con la enfermedad de Alzheimer. Esta disparidad se observa también como resultado del análisis estadístico en la zona de la corteza entorrinal de los individuos, pero no en el análisis de expresión.

Estas diferencias en los resultados obtenidos pueden ser debidas a la utilización de datos distintos para cada estudio, además de por un número bajo de muestras en el análisis de expresión. Lo idóneo hubiese sido utilizar datos provenientes de los mismos individuos en ambos análisis para obtener unos resultados en equilibrio, pero en este proyecto no ha sido posible debido a la escasez de datos públicos de este tipo. Para futuras investigaciones similares, se aconseja a ser posible utilizar datos del mismo origen para evitar estos inconvenientes, así como utilizar un número de muestras o individuos óptimo.

A pesar de estos resultados discordantes para la corteza entorrinal, el análisis estadístico señala una disparidad entre sexos en todas las zonas evaluadas. De nuevo, indicativo de la necesidad de más líneas de investigación enfocadas de esta manera.

5. Conclusión

El objetivo principal de este proyecto era evaluar si en la enfermedad de Alzheimer existe un dimorfismo entre géneros, ya sea clínico o genético, que implique la necesidad de investigación que tenga en cuenta esta diferenciación. Esto se ha podido evidenciar a partir de los resultados obtenidos en los análisis realizados.

A raíz de este trabajo, se han logrado los objetivos planteados inicialmente, ya que se ha podido evidenciar tanto diferencias clínicas como genéticas específicas de sexo para la enfermedad de Alzheimer. Además, se ha podido establecer una relación clara entre los dos tipos de resultados obtenidos.

El seguimiento inicialmente establecido se ha podido seguir cumpliendo las fechas definidas. No obstante, el tiempo dedicado a cada tarea no ha sido el establecido al inicio del proyecto debido a dificultades que se han presentado en algunas de ellas y a la adición de tareas que no estaban previstas al iniciar el trabajo. Por este motivo, el tiempo dedicado al proyecto ha tenido que ser superior para poder cumplir con el plazo de tiempo establecido. A pesar de la aparición de algunos obstáculos durante el proyecto, tanto la metodología como la planificación ha podido cumplirse dentro de los plazos previstos.

Según el análisis estadístico realizado, las mujeres con Alzheimer tienen un volumen cerebral general inferior al de los hombres con Alzheimer. En concreto, la zona del hipocampo presenta estas disparidades entre sexos también en el análisis diferencial, en el que se identifican genes diferencialmente expresados como *Sp1*, *MBLN1* y *ALCAM* entre otros.

Adicionalmente, se refleja en los resultados una disminución del volumen del hipocampo a medida que aumenta la edad de los pacientes con Alzheimer, independientemente de su género.

Como conclusión, a partir de este proyecto se confirma que una de las zonas candidatas donde se originan parte de las disparidades observadas en los pacientes es el hipocampo. En esta zona se presentan diferencias de género tanto clínicas como genéticas, que pueden ser clave para el desarrollo diferencial de la enfermedad. Todo esto sostiene la teoría con la que se iniciaba la investigación: existen disparidades genéticas de género en la enfermedad de Alzheimer, que deben seguir siendo investigadas para conseguir avances en su diagnóstico y su terapia.

Debe tenerse en cuenta, por último, que cualquiera de los genes identificados en este estudio, para considerarse realmente expresado diferencialmente, debería ser verificado mediante técnicas como RT-qPCR. Este estudio es un paso hacia el descubrimiento de genes candidatos y posibles orígenes de las disparidades de sexo en la enfermedad, pero no son los definitivos. Asimismo, con este proyecto se demuestra la importancia de estudiar enfermedades como el Alzheimer diferenciando entre géneros y se anima a futuras investigaciones a adoptar este tipo de enfoque.

6. Glosario

AE:
Análisis estadístico.

AED:
Análisis de expresión diferencial.

ALCAM:
Activated leukocyte cell adhesion molecule.

Alzheimer's Disease Neuroimaging Initiative (ADNI):
Iniciativa de neuroimagen de la enfermedad de Alzheimer.

AMPK:
Proteína Quinasa Activada por Monofosfato de Adenina.

APOE:
Apolipoproteína E.

CE:
Corteza entorrinal.

CLIC4:
Chloride intracellular channel 4.

EGFR:
Receptor del factor de crecimiento epidérmico.

Gene Expression Omnibus (GEO):
Repositorio público de datos de genómica funcional (microarrays y RNAseq).

Gene Ontology (GO):
Ontología genética. Vocabulario controlado que describe el gen y los atributos del producto génico en cualquier organismo.

Genome Wide Association Study (GWAS):
Estudio de asociación de genoma completo.

Hip:
Hipocampo.

Kyoto Encyclopedia of Genes and Genomes (KEGG):
Colección de bases de datos en línea de genomas, rutas enzimáticas, y químicos biológicos.

Log Folg Change (LFC):
Medida que describe cuánto cambia una cantidad entre una mediación original y una posterior.

MBNL1:
Muscleblind like splicing regulator 1.

National Center for Biotechnology Information (NCBI):
Centro Nacional para la Información Biotecnológica.

OMS:
Organización Mundial de la Salud.

PCA:
Análisis de componentes principales.

RAD21:
RAD21 cohesin complex component.

Real-Time qPCR (RT-qPCR):
PCR en tiempo real.

Robust Multi-array Average (RMA):
Algoritmo usado para crear una matriz de expresión.

Single Nucleotide Polymorphism (SNP):
Polimorfismo de un solo nucleótido.

SP1:
Sp1 transcription factor.

7. Bibliografía

1. OMS. Demencia. *Organización Mundial de la Salud* https://www.who.int/es/news-room/fact-sheets/detail/dementia (2020).
2. NIH. Hoja informativa sobre la enfermedad de Alzheimer. *National Institute on Aging* https://www.nia.nih.gov/espanol/hoja-informativa-sobre-enfermedad-alzheimer (2020).
3. Alzheimer's Association. 2020 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **16**, 391–460 (2020).
4. Bellenguez, C., Grenier-Boley, B. & Lambert, J. C. Genetics of Alzheimer's disease: where we are, and where we are going. *Curr. Opin. Neurobiol.* **61**, 40–48 (2020).
5. UniProt. UniProtKB - P02649 (APOE_HUMAN). (2020).
6. Ungar, L., Altmann, A. & Greicius, M. D. Apolipoprotein E, Gender, and Alzheimer's Disease: An Overlooked, but Potent and Promising Interaction. *Brain Imaging Behav.* **8**, 262–273 (2014).
7. Toro, C. A., Zhang, L., Cao, J. & Cai, D. Sex differences in Alzheimer's disease: Understanding the molecular impact. *Brain Res.* **1719**, 194–207 (2019).
8. Nebel, R. A. *et al.* Understanding the impact of sex and gender in Alzheimer's disease: A call to action. *Alzheimer's Dement.* **14**, 1171–1183 (2018).
9. Ruiz Cantero, M. T. *Perspectiva de género en medicina*. (2018).
10. Dumitrescu, L. *et al.* Sex differences in the genetic predictors of Alzheimer's pathology. *Brain* **142**, 2581–2589 (2019).
11. Sun, L. L., Yang, S. L., Sun, H., Li, W. Da & Duan, S. R. Molecular differences in Alzheimer's disease between male and female patients determined by integrative network analysis. *J. Cell. Mol. Med.* **23**, 47–58 (2019).
12. Donohue, M. C. & Sun, C. ADNIMERGE: Key ADNI variables. 20–22 (2013).
13. Wang, Y. ADNIMERGE. *GitHub* https://github.com/CS209-Final-Project/CS209_Final_Project/blob/master/ADNIMERGE.csv (2017).
14. Berchtold, N. & Cotman, C. GSE48350. *NCBI, Gene Expression Omnibus* http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51746 (2019).
15. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).
16. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
17. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

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

19. Yu, G. clusterProfiler: universal enrichment tool for functional and comparative study. http://yulab-smu.top/clusterProfiler-book/chapter1.html.

20. Chen, L. *et al.* Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. *PLoS One* **12**, 1–22 (2017).

21. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).

22. Nobis, L. *et al.* Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage Clin.* **23**, 101904 (2019).

23. Koran, M. E. I., Wagener, M. & Hohman, T. J. Sex differences in the association between AD biomarkers and cognitive decline. *Brain Imaging Behav.* **11**, 205–213 (2017).

24. Apostolova, L. G. *et al.* Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (MCI), and Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **26**, 17–27 (2012).

25. Berchtold, N. C. *et al.* Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. *Neurobiol. Aging* **34**, 1653–1661 (2013).

26. UniProt. UniProtKB - P08047 (SP1_HUMAN). https://www.uniprot.org/uniprot/P08047 (2020).

27. Santpere, G., Nieto, M., Puig, B. & Ferrer, I. Abnormal Sp1 transcription factor expression in Alzheimer disease and tauopathies. *Neurosci. Lett.* **397**, 30–34 (2006).

28. Citron, B. A., Dennis, J. S., Zeitlin, R. S. & Echeverria, V. Transcription factor Sp1 dysregulation in Alzheimer's disease. *J. Neurosci. Res.* **86**, 2499–2504 (2008).

29. UniProt. UniProtKB - Q9NR56 (MBNL1_HUMAN). (2020).

30. Wang, P. Y. *et al.* Reduced cytoplasmic MBNL1 is an early event in a brain-specific mouse model of myotonic dystrophy. *Hum. Mol. Genet.* **26**, 2247–2257 (2017).

31. Carpentier, C. *et al.* Tau exon 2 responsive elements deregulated in myotonic dystrophy type I are proximal to exon 2 and synergistically regulated by MBNL1 and MBNL2. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1842**, 654–664 (2014).

32. orpha.net. Distrofia miotónica de Steinert. (2007).

33. Sergeant, N. *et al.* Dysregulation of human brain microtubule-associated tau mRNA maturation in myotonic dystrophy type 1. *Hum. Mol. Genet.* **10**, 2143–2155 (2001).

34. Cai, Z., Yan, L. J., Li, K., Quazi, S. H. & Zhao, B. Roles of AMP-activated protein kinase in Alzheimer's disease. *NeuroMolecular Med.* **14**, 1–14 (2012).

35. Chen, X., Wang, C., Zhou, S., Li, X. & Wu, L. The impact of EGFR gene polymorphisms on the risk of alzheimer's disease in a chinese han population: A case-controlled study. *Med. Sci. Monit.* **24**, 5035–5040 (2018).

36. Nazarian, A., Yashin, A. I. & Kulminski, A. M. Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex-disparities. *Alzheimer's Res. $ Ther.* **11**, 1–21 (2019).

37. Deming, Y. *et al.* Sex-Specific Genetic Predictors of Alzheimer's Disease Biomarkers. *Acta Neuropathol.* **136**, 857–872 (2018).

38. Dumitrescu, L., Mayeda, E. R., Sharman, K., Moore, A. M. & Hohman, T. J. Sex Differences in the Genetic Architecture of Alzheimer's Disease. *Curr. Genet. Med. Rep.* **7**, 13–21 (2019).