Universitat Oberta de Catalunya

uoc.edu

Análisis de datos NGS para la determinación de nuevos factores moleculares implicados en la adrenoleucodistrofia

Ana María Burgos Ruiz
Máster Bioinformática y Bioestadística
Área 3

María Elena Rojano Rivera
Ferrán Prados Carrasco

Fecha Entrega: 05/01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de datos NGS para la determinación de nuevos factores moleculares implicados en la adrenoleucodistrofia*
Nombre del autor:	*Ana María Burgos Ruiz*
Nombre del consultor/a:	*María Elena Rojano Rivera*
Nombre del PRA:	*Ferrán Prados Carrasco*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 3*
Idioma del trabajo:	*Español*
Palabras clave	*NGS, ARN-seq, adrenoleucodistrofia*

Resumen del Trabajo

La adrenoleucodistrofia ligada al cromosoma X (ALD-X) es una enfermedad bien definida genéticamente, pero cuyo fenotipo clínico posee una gran heterogeneidad hasta ahora inexplicable. Las mutaciones en ABCD1 reflejan uno de los fenotipos centrales observado en los pacientes con X-ALD, la adrenomieloneuropatía (AMN). Sin embargo, estas mutaciones por si solas no explican el fenotipo correspondiente a la adrenoleucodistrofia cerebral (CALD). Diferentes mecanismos desconocidos parecen estar implicados en la enfermedad. Con el fin de determinar los factores genéticos y moleculares implicados en el desarrollo de la CALD se han realizado análisis de expresión, de co-expresión y de enriquecimiento funcional partiendo de los datos de secuenciación. Para ello, se empleó la herramienta DEgenes Hunter para comparar, en un primer momento,18 muestras de procedencia heterogénea y, en segundo lugar, seis muestras procedentes del mismo tipo celular para evitar el sesgo resultante de los diferentes tipos celulares. El análisis reveló que uno de los genes sometidos a una mayor inhibición es CNTN4, el cual codifica para una proteína de adhesión celular implicada en la formación de conexiones en el sistema nervioso en desarrollo. Además, se propone la hipótesis de que la disminución de ácido lisofosfolipídico, molécula señal que participa en la adhesión celular, podría estar participando en la desmielinización característica de la enfermedad.

Abstract:

X-linked adrenoleukodystrophy (X-ALD) is a genetically well-defined disease, but the clinical phenotype of which has hitherto unexplained great heterogeneity. Mutations in ABCD1 reflect one of the central phenotypes seen in patients with X-ALD, adrenomyeloneuropathy (AMN). However, these mutations alone do not explain the phenotype corresponding to cerebral adrenoleukodystrophy (CALD). Different unknown mechanisms appear to be involved in the disease. In order to determine the genetic and molecular factors involved in the development of CALD, expression, co-expression and functional enrichment analyzes have been carried out based on the sequencing data. To this end, the DEgenes Hunter tool was used to compare, firstly, 18 samples of heterogeneous origin and, secondly, six samples from the same cell type to avoid the bias resulting from the different cell types. The analysis revealed that one of the most inhibited genes is CNTN4, which codes for a cell adhesion protein involved in the formation of connections in the developing nervous system. Furthermore, the hypothesis is proposed that the decrease in lysophospholipidic acid, a signal molecule that participates in cell adhesion, could be participating in the demyelination characteristic of the disease.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

1.1.1. Descripción general

La adrenoleucodistrofia (ALD) es una enfermedad rara con base genética que desencadena un incorrecto metabolismo y transporte de los ácidos grasos de cadena larga en los peroxisomas [1], [2]. Esta anomalía produce un daño severo en las vainas de mielina de las neuronas, lo cual provoca en los pacientes trastornos como espasmos musculares, epilepsia, parálisis o hiperactividad, entre otros síntomas patológicos [3]. La ALD está ligada al cromosoma X, específicamente a mutaciones localizadas en el gen *ABCD1*, que se transcribe y traduce en una proteína de transporte de ácidos grasos en peroxisomas [4]. La ALD ligada al cromosoma X posee diferentes fenotipos, uno de ellos es la adrenomieloneuropatía (AMN), cuyo único factor desencadenante es la mutación del gen *ABCD1*. Sin embargo, otro de los fenotipos es la adrenoleucodistrofia cerebral (CALD), con síntomas mucho más severos. Para que tenga lugar este último fenotipo tiene que existir tanto una mutación en el gen *ABCD1* como otros factores genéticos con una directa implicación en el desarrollo de la enfermedad que aumenten la susceptibilidad de los pacientes a ciertos agentes externos o agravando su cuadro fenotípico [5]. La correcta caracterización de estos factores genéticos involucrados en el desarrollo de la CALD es esencial para la comprensión de los mecanismos moleculares implicados en la enfermedad.

En los últimos años, los avances en las tecnologías de secuenciación han permitido su uso como herramienta de apoyo al diagnóstico en pacientes con enfermedades genéticas [6], [7]. Tal es el caso de la tecnología RNA-sequencing (RNA-seq): actualmente, los análisis de transcriptoma usados para caracterizar los genes con expresión diferencial ante determinadas condiciones son asequibles económicamente y su empleo se ha extendido para ayudar a la comprensión de dichas enfermedades [8]. En este caso, el análisis de datos obtenidos con las nuevas tecnologías de secuenciación puede ayudar a comprender las causas que producen la CALD, mejorar su correcto diagnóstico y desarrollar nuevos tratamientos que mejoren la calidad de vida de los pacientes.

Los análisis de RNA-seq permiten cuantificar los niveles de expresión génica y se han aplicado a múltiples campos de la medicina [9]. Existen numerosas herramientas computacionales para analizar los datos de RNA-seq para la determinación de genes expresados diferencialmente. Esto hace que sin una buena base o formación específica en análisis de datos RNA-seq resulte complicado decidir cuál de las distintas opciones es la mejor para llevar a cabo un análisis de transcriptómica. Asimismo, los análisis clásicos de RNA-seq se centran en determinar los perfiles de expresión de genes de manera individual; sin embargo, se conoce que existen módulos de genes que co-expresan juntos [10]. Aquellos genes que no

superan los filtros empleados para los análisis de expresión se descartan de los análisis a pesar de que pueden contener información valiosa. Para ello, llevar a cabo un análisis de co-expresión de genes es una estrategia que puede servir para evitar esta pérdida de información y sirviendo para la comprensión de las causas genéticas de una enfermedad determinada.

El objetivo principal de este trabajo de fin de máster (TFM) se centrará en la determinación de los factores genéticos y moleculares implicados en el desarrollo de la CALD a través del análisis de expresión, co-expresión y funcional de datos de RNA-seq. Para su consecución, se llevará a cabo un reanálisis de muestras de tejidos cerebral comparando la expresión en individuos jóvenes (niños) sanos frente a la de individuos afectados por CALD, a partir de datos publicados en el estudio de Catherine A. Lee y colaboradores en 2018 [11], empleando una versión actualizada del flujo de trabajo puesto a punto por investigadores del grupo de bioinformática BIO276 de la Universidad de Málaga con la herramienta DEgenes Hunter [12].

1.1.2. Justificación del trabajo

Se considera enfermedad rara a aquella que se da en una proporción menor a la de 1 por cada 2000 personas. Debido a su baja prevalencia en la población, su correcto diagnóstico es complicado, teniendo de media una duración de cinco años en hacerse efectivo. Además, su tratamiento no siempre es acorde a las circunstancias del paciente, siendo inexistente o incluso agravando la sintomatología en numerosos casos. Sin embargo, a pesar de tratarse de enfermedades con baja prevalencia, se estima que el número de enfermedades raras supera las 7.000.

Investigar más en profundidad el amplio rango de factores genéticos implicados en el desarrollo de estas enfermedades podría ser de gran ayuda para determinar no solo las consecuencias moleculares implicadas en el desarrollo de estas patologías, sino las de enfermedades más comunes como el cáncer, diabetes, enfermedades cardiovasculares, neurodegenerativas o incluso determinar factores involucrados en la susceptibilidad a patógenos externos como en el caso de la nueva infección producida por el SARS-Cov-2.

En este trabajo llevaremos a cabo el análisis de los posibles genes con expresión diferencial y co-expresados en muestras de pacientes afectados por CALD, así como los factores moleculares implicados en el desarrollo de esta enfermedad. En función de los resultados biológicos obtenidos mediante este análisis, consideraremos su uso como punto de partida para lograr el desarrollo de terapias personalizadas para el tratamiento de este tipo de pacientes.

1.2. Objetivos del Trabajo

1.2.1. Objetivos generales

El objetivo general de este TFM es la determinación de los factores genéticos y moleculares implicados en el desarrollo de la ALD a través del análisis de expresión, co-expresión y funcional de datos de secuenciación, empleando la herramienta DEgenes Hunter [12], a partir del trabajo publicado por Catherine A. Lee y colaboradores en 2018 [11].

1.2.2. Objetivos específicos

1. Implementación del flujo de trabajo que utiliza la herramienta DEgenes Hunter para el análisis de datos NGS en un clúster de supercomputación.
2. Búsqueda de estudios relacionados con ALD y descarga de los conjuntos de datos NGS correspondientes para su análisis.
3. Análisis de datos de expresión para la determinación de genes expresados diferencialmente, co-expresados y enriquecimiento funcional de los conjuntos de datos de ALD.
4. Discusión de los resultados biológicos obtenidos a través de información bibliográfica.

1.3. Enfoque y método seguido

La idea de este trabajo consiste en la puesta a punto de un flujo de trabajo que nos permita conseguir una aproximación fiable de los genes que están siendo diferencialmente expresados. Esto es gracias a la herramienta DEgenes Hunter [12].

A pesar de que los análisis de expresión diferencial a partir de datos de RNA-seq son ampliamente conocidos en el ámbito de la bioinformática, la herramienta DEgenes Hunter ofrece una serie de ventajas para el análisis e interpretación de estos datos. En primer lugar, para la caracterización de los genes expresados diferencialmente (DEG) utiliza la combinación de cuatro algoritmos de análisis de expresión ampliamente empleados en el campo del análisis de RNA-seq (edgeR [13], DEseq2 [14], NOISeq [15] y limma [16]). Asimismo, no se centra solo en el análisis de expresión, sino que su nueva versión incluye un análisis de co-expresión de genes gracias a la implementación del algoritmo WGCNA [17]. Finalmente, DEgenes Hunter incluye herramientas de enriquecimiento funcional empleando la herramienta ClusterProfiler [18], la cual usa términos de la Gene Ontology [19], KEGG [20] y Reactome [21]. El artículo correspondiente a esta nueva versión de DEgenes Hunter está en proceso de revisión (F. M. Jabato y colaboradores, Analysing rare disease model datasets with the ExpHunter Suite, *Clinical and Translational Medicine*).

1.4. Planificación del Trabajo

En este apartado se establecen las tareas que fueron llevadas a cabo durante el proyecto y se concretan los plazos de cada una de ellas, así como los recursos que se han utilizado y los posibles factores que podrían haber afectado de forma negativa al cumplimiento de la planificación.

1.4.1. Tareas

Teniendo en cuenta la temporización y los objetivos marcados, se establecieron tres fases para el desarrollo del proyecto. Las tareas se distribuyeron en las diferentes fases teniendo en cuenta la duración de cada una de ellas.

Fase 1

- Organización y elaboración del Plan de Trabajo
- Descarga de datos para realizar el análisis de expresión diferencial a partir de información almacenada en el SRA.
- Descarga e instalación de software para el análisis y puesta a punto del flujo de trabajo.
- Diseño de los targets para preparar el análisis de los datos de expresión (controles vs. tratamientos/enfermos).

Fase 2

- Pre-procesamiento de los datos: limpieza de las lecturas, indexado de la referencia (*H. sapiens*, GRCh38) y alineamiento de las lecturas.
- Obtención de la tabla de conteo con los genes expresados en las muestras.
- Análisis de la expresión diferencial y co-expresión de los genes en las muestras con la herramienta DEgenes Hunter.
- Enriquecimiento funcional de los datos de expresión y co-expresión con la herramienta DEgenes Hunter.

Fase 3

- Análisis e interpretación biológica de los resultados.
- Redacción del manuscrito.
- Creación de la presentación.

1.4.2. Temporización e hitos

A continuación, se muestra en un diagrama de Gantt la planificación que se ha llevado a cabo durante los meses de octubre, noviembre y diciembre. Se indican todos

los hitos planificadas para la consecución de los objetivos del proyecto, así como da duración temporal de cada una de ellas y la fase en la que se realizó.

En octubre de 2020 se llevó a cabo la primera fase del proyecto. En ella se procedió a la búsqueda bibliográfica de un estudio sobre la ALD con un set de datos que se pudiese emplear para el proyecto. La búsqueda resultó en un proyecto con datos de pacientes con ALD asociada al cromosoma X [11] y públicamente disponibles en el Sequence Read Archive (SRA) y en BioProject (código de acceso: PRJNA422218). Se procedió con la descarga de los archivos crudos de lecturas (fastq) y, paralelamente, se puso a punto el software con el que se llevó a cabo el análisis de expresión diferencial y de enriquecimiento funcional en el clúster de supercomputación Picasso, en el Supercomputing and Bioinnovation Center (SCBI, http://www.scbi.uma.es/site/) de la Universidad de Málaga.

En noviembre de 2020 se procedió con el pre-procesamiento de las lecturas llevando a cabo su limpieza y posterior alineamiento contra la última versión del genoma de referencia humano (GRCh38). Una vez obtenida la tabla de conteos de cada una de las muestras analizadas, se fusionaron en una única tabla que se empleó para llevar a cabo los análisis de expresión diferencial, co-expresión y enriquecimiento genético con la herramienta DEgenes Hunter [12].

En diciembre de 2020 se realizó la interpretación biológica de los datos, se redactó el manuscrito y se diseñó la presentación.

Finalmente, en el mes de enero de 2021 se realizará la presentación del proyecto.

TAREAS	Fecha de inicio	Fecha de finalización	13.10.2020	14.10.2020	15.10.2020	16.10.2020	17.10.2020	18.10.2020	19.10.2020	20.10.2020	21.10.2020	22.10.2020	23.10.2020	24.10.2020	25.10.2020	26.10.2020	27.10.2020	28.10.2020	29.10.2020	30.10.2020	31.10.2020	01.11.2020	02.11.2020	03.11.2020
FASE 1	13/10/20	18/10/20																						
Búsqueda bibliográfica	13/10/20	23/10/20																						
Descarga de datos	23/10/20	27/10/20																						
Instalación de herramientas	28/10/20	31/10/20																						
Exploración de DEGenesHunters	28/10/20	3/11/20																						
Preparación de targets del experimento	1/11/20	6/11/20																						

TAREAS	Fecha de inicio	Fecha de finalización	04.11.2020	05.11.2020	06.11.2020	07.11.2020	08.11.2020	09.11.2020	10.11.2020	11.11.2020	12.11.2020	13.11.2020	14.11.2020	15.11.2020	16.11.2020	17.11.2020	18.11.2020	19.11.2020	20.11.2020	21.11.2020	22.11.2020	23.11.2020	24.11.2020
FASE 2																							
Limpieza	4/11/20	6/11/20																					
Indexado de la referencia	7/11/20	8/11/20																					
Mapeo de lecturas	7/11/20	9/11/20																					
Análisis de expresión diferencial y co-expresión	10/11/20	14/11/20																					
Análisis resultados expresión diferencial y co-expresión	15/11/20	20/11/20																					
Análisis de enriquecimiento funcional	21/11/20	24/11/20																					

TAREAS	Fecha de inicio	Fecha de finalización	25.11.2020	26.11.2020	27.11.2020	28.11.2020	29.11.2020	30.11.2020	01.12.2020	02.12.2020	03.12.2020	04.12.2020	05.12.2020	06.12.2020	07.12.2020	08.12.2020	09.12.2020	10.12.2020	11.12.2020	12.12.2020	13.12.2020	14.12.2020	05.01.2021
FASE 3																							
Interpretación biológica de los datos	25/11/20	10/12/20																					
Redacción de memoria y elaboración ppt	11/12/20	5/1/21																					

1.4.4. Breve sumario de productos obtenidos

Los resultados tangibles que se esperaban obtener de este trabajo se enumeran a continuación:

1. Un plan de trabajo, donde se mostrarán los objetivos, y la forma de alcanzarlos.

2. Una memoria, donde se mostrará tanto el proceso como los resultados del análisis llevado a cabo.
3. Tabla de conteo de expresión genética (enlace con Google drive: https://drive.google.com/file/d/1mjiZ6-xq67WkleH01PuMZ7uIU799Ecz9/view?usp=sharing).
4. Informe con los resultados del análisis de expresión diferencial y de co-expresión con las 18 muestras (enlace con Google drive: https://drive.google.com/file/d/1uBaFuauN6qPJHy0yLqpOWetpN12Y0loJ/view?usp=sharing).
5. Informe con los resultados del análisis de expresión diferencial y de co-expresión con las muestras de queratinocitos (enlace con Google drive: https://drive.google.com/file/d/1UCRULumSau5vg_SfjzFRLBI2JMVB8sb4/view?usp=sharing).
6. Informe de enriquecimiento funcional de los datos de expresión y co-expresión con la herramienta DEgenes Hunter con las 18 muestras (enlace con Google drive: https://drive.google.com/file/d/1iRY_BxgOFClnhxCf36eynSWHHKEIz5Wa/view?usp=sharing).
7. Informe de enriquecimiento funcional de los datos de expresión y co-expresión con la herramienta DEgenes Hunter con las muestras de queratinocitos (enlace con Google drive: https://drive.google.com/file/d/1Qvm43kj8sXqcbx52N-6mRddpH1prPzYw/view?usp=sharing).
8. En el caso de que los resultados fuesen novedosos y de interés en el campo de la medicina, un artículo científico publicado en una revista indexada dentro del Journal Citation Reports (JCR).
9. Una presentación virtual, donde se expondrán cada uno de los apartados de la memoria, explicando la metodología seguida, resultados obtenidos y conclusiones a partir de los mismos.
10. Una autoevaluación del proyecto llevado a cabo.

1.4.5. Breve descripción de los otros capítulos de la memoria

El objeto de este estudio se centra en ver las diferencias entre controles (WT) y pacientes (CALD) para confirmar qué genes cambian de expresión en condiciones normales frente a aquellos que varían de expresión durante el procedimiento patológico. Por ello, en el primer capítulo se explica detalladamente

la enfermedad. Se exponen las cuestiones tanto epidemiológicas como genéticas y moleculares conocidas hasta la fecha. Gracias a esto, se podrá llevar a cabo una interpretación más precisa de los datos que se obtengan. En el siguiente capítulo, se detallan los materiales y métodos utilizados para llevar a cabo el análisis de los genes diferencialmente expresados en CALD, el análisis de co-expresión diferencial y el enriquecimiento funcional. Todos estos resultados se expondrán en otro capítulo, en el que se explicará y desglosará cada uno de los informes con el objetivo de esclarecer los mecanismos moleculares que actúan en la enfermedad. Por último, en el capítulo final, se expondrán las conclusiones a las que se ha llegado tras la realización de este proyecto.

2. Contexto biológico: adrenoleucodistrofia ligada al cromosoma X.

2.1. Aspectos biológicos y bioquímicos de la enfermedad.

La adrenoleucodistrofia ligada al cromosoma X (X-ALD; OMIM, fenotipo MIM número #300100) es la enfermedad peroxisomal hereditaria más común. La incidencia conjunta de hombres hemicigotos y mujeres heterocigotas portadoras es de 1:16.800 recién nacidos [22]. Uno de los síntomas clave durante el desarrollo de la enfermedad es la lenta pero progresiva axonopatía (degeneración de los exones de las neuronas).

La X-ALD puede presentarse de dos formas diferentes: adrenomieloneuropatía (AMN) y adrenoleucodistrofia cerebral (CALD).

La AMN es la forma más leve de ALD. Comienza al final de la adolescencia o al principio de la edad adulta. La axonopatía representa la característica clínica central de la AMN en pacientes varones, su inicio oscila los entre 20 y 30 años, pero en mujeres heterocigotas se inicia entre los 40 y 50 años. Los síntomas con los que la AMN comienza a expresarse son la debilidad y rigidez progresivas de las piernas, alteraciones del esfínter y la alopecia. Además, el 66% de los pacientes masculinos con AMN y menos del 5% de las mujeres tienen insuficiencia adrenocortical (enfermedad de Addison) [2]. Las señales de resonancia magnética que se han detectado en la AMN muestran anomalías de la sustancia blanca en el centro oval, tractos piramidales en el tronco del encéfalo y cápsulas internas. Sin embargo, no hay realce de gadolinio, lo que indica que la barrera hematoencefálica se encuentra intacta y que hay ausencia de un proceso inflamatorio agudo [2]. En esta forma clínica los signos de insuficiencia suprarrenal pueden preceder en muchos años a la aparición de los síntomas neurológicos.

Por su parte, la CALD comienza en la infancia o en la adolescencia, y el cuadro degenerativo neurológico evoluciona hasta una demencia grave con deterioro de la visión, la audición, el habla y la marcha, y un fallecimiento precoz (muerte en 3-5 años). Aproximadamente, en el 60% de los pacientes varones que sufren X-ALD, se produce una rápida y progresiva desmielinización inflamatoria. El inicio de la inflamación es más común en niños (35-40%), antes del inicio de AMN, y menos frecuente (20%) en adolescentes o adultos. La desmielinización inflamatoria suele comenzar en la línea media del cuerpo calloso y progresa hacia afuera como de forma simétrica en ambos hemisferios. Clínicamente, esto coincide con un deterioro neurológico progresivo, que conduce a un estado vegetativo. En ocasiones, se ha observado un paro espontáneo de enfermedad cerebral Por lo que proporcionan un tratamiento eficaz para la CALD. Es importante señalar que la inflamación aguda solo se observa en el sistema nervioso central (SNC) y no en otros tejidos de pacientes con X-ALD [2].

2.2. Genética y bioquímica de X-ALD

En la matriz peroxisomal y gracias a las enzimas de la ruta de la β-oxidación se degradan los ácidos grasos saturados no ramificados de cadena muy larga (VLCFA, del inglés *very long chain fatty acid*). Se considera VLCFA a aquellos ácidos grasos de cadena mayor o igual a 22 carbonos (22:0). Los pacientes con X-ALD, ya sea con AMN como con CALD, acumulan los VLCFA, en particular C26:0, en los tejidos y los fluidos corporales, por lo que sirven como marcador para el diagnóstico de X-ALD [2].

Se han llevado a cabo diversos estudios para identificar tanto los factores genéticos como moleculares implicados en el desarrollo de la X-ALD. Gracias a ellos se han identificado una serie de mutaciones en pacientes de X-ALD que afectan al gen del *trasportador ATP-binding cassette (ABC) miembro 1 de la subfamilia D (ABCD1)* en la localización cromosómica Xq28 [23]. El gen *ABCD1* codifica el semi-transportador ABC peroxisomal ABCD1. El dominio de unión a ATP de ABCD1 se encuentra expuesto al citosol y los sustratos se transportan desde el citosol al peroxisoma consumiendo ATP. Para la proteína ABCD1 humana, los VLCFA activados por CoA, como C26:0-CoA o C24:0-CoA y C22:0-CoA, son sustratos válidos para el receptor. En fibroblastos con ALD-X se ha observado que la degradación de estos ácidos grasos por β-oxidación peroxisomal está fuertemente reducida [24]. Además, también se ha demostrado que en fibroblastos, el trasporte y degradación de C26:0-CoA en los peroxisomas se bloquea utilizando anticuerpos anti-ABCD1 [25].

2.3. Transportadores ABC y su papel en X-ALD

Se ha demostrado que el defecto que subyace a la enfermedad X-ALD es el defecto hereditario en el trasportador peroxisómico. Por lo tanto, el nivel de expresión de *ABCD1* en diferentes tejidos y tipos de células es de suma importancia para comprender esta patología. *ABCD1* se expresa de manera bastante débil en el SNC en comparación con otros tejidos [26].

Existen otros dos transportadores ABC peroxisomales, ABCD2 y ABCD3, que son capaces de asumir funciones de ABCD1. Estos no están mutados en ALD-X [27]. En los tejidos diana de la enfermedad X-ALD, se han descrito patrones de expresión bastante complementarios para los transportadores ABC peroxisomales ABCD1 y ABCD2 [28]. Por ejemplo, en la glándula suprarrenal, *ABCD1* se expresa altamente en la corteza, pero no es detectable en la médula. Por el contrario, *ABCD2* se expresa en la médula, pero no en la corteza. Este patrón de expresión coincide con el hecho de que la patología suprarrenal en X-ALD está restringida a la corteza, donde ABCD2 apenas se expresa.

Igualmente, se ha visto como la sobreexpresión de *ABCD2* y, en menor medida, *ABCD3* puede corregir el defecto metabólico en fibroblastos de pacientes con X-ALD [29]. Por tanto, se puede afirmar que los niveles de expresión de *ABCD2* y *ABCD3*, contribuyen a la manifestación de la deficiencia de ABCD1 en diferentes tipos

celulares. Sin embargo, es poco probable que las variaciones alélicas de *ABCD2* o *ABCD3* colaboren en la heterogeneidad clínica de X-ALD [30].

En un estudio realizado a seis miembros de una misma familia, todos afectados por X-ALD, se observó como la misma mutación (transversión de citosina a guanina en el nucleótido número 1451 del exón cinco de *ABCD1*) resultaba en todos los fenotipos clínicos de X-ALD posibles [31]. Además se han encontrado pacientes con todo el espectro clínico de X-ALD que tenían el mismo defecto: una pérdida completa de la proteína ABCD1 debida o bien a cambios en el marco de lectura, a mutaciones sin sentido o a grandes delecciones [32]. Todo esto indica una falta de correlación entre el genotipo y el fenotipo para esta patología [33].

2.4. Factores genéticos implicados en CALD.

Las mutaciones en el gen *ABCD1* se han asociado a la AMN; sin embargo, la CALD no puede explicarse solo por mutaciones en este gen. Esta forma inflamatoria de X-ALD se manifiesta en los pacientes (la gran mayoría niños menores de diez años) por su incapacidad de degradar VLCFA, así como la incorporación de estos en diferentes lípidos complejos. Diferentes mecanismos desconocidos parecen estar implicados en distintos procesos patológicos, entre los que se incluyen: (1) la desmielinización cerebral inicial, (2) la conversión esporádica de esta desmielinización en desmielinización inflamatoria rápidamente progresiva, (3) la conversión AMN puro a desmielinización inflamatoria rápidamente progresiva, y (4) la incapacidad de detener la inflamación mediante mecanismos intrínsecos o estrategias terapéuticas antiinflamatorias.

Unos de los factores que puede producir desestabilización progresiva de las vainas de mielina y la posterior desmielinización es la acumulación de VLCFA [2]. Por tanto, esta fase inicial de inicio espontáneo de la desmielinización podría estar directamente relacionada con el nivel de VLCFA en la vaina de mielina. Aunque los niveles de VLCFA en sangre o fibroblastos cultivados son indistinguibles en pacientes con X-ALD con diferentes fenotipos clínicos, se encontró que las cantidades de VLCFA eran más altas en la sustancia blanca de los pacientes con CALD en comparación con los pacientes con AMN [34]. Además, se ha visto como los oligodendrocitos derivados de células madre pluripotentes inducidas de pacientes con CALD acumularon más VLCFA que los derivados de pacientes con AMN [35]. En la Figura 1 se representa un modelo hipotético que muestra los posibles eventos secuenciales que conduzcan a la desmielinización inflamatoria en CALD [22].

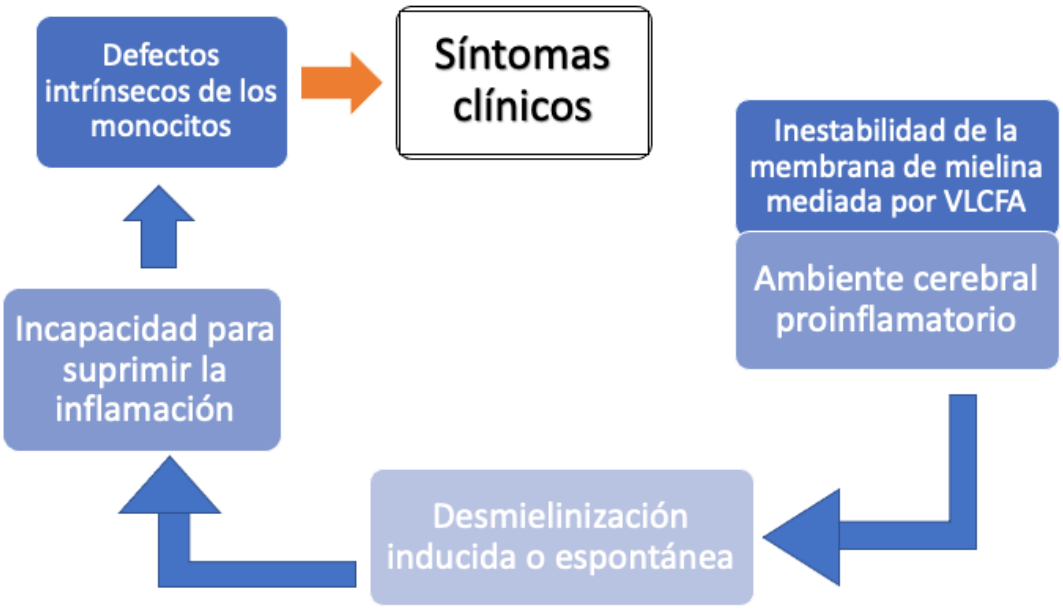

Figura 1. Modelo hipotético de CALD. Se muestra los eventos secuenciales que conducen a la desmielinización inflamatoria en CALD, que desemboca en la aparición de los síntomas clínicos.

Para la CALD, *ABCD1* sigue siendo un gen de susceptibilidad necesario, pero no suficiente, para que ocurra la desmielinización inflamatoria. Por lo tanto, para comprender por qué se produce este fenotipo concreto de X-ALD es necesario conocer todos los posibles factores genéticos involucrados en su desarrollo. Esto permitiría desarrollar tratamientos o fármacos específicos que palien las consecuencias de esta enfermedad en los pacientes que la padecen.

2.5. Análisis de datos transcriptómicos para la determinación de genes expresados diferencialmente en CALD.

Los nuevos avances en las tecnologías de secuenciación han permitido su utilización para fines clínicos como herramientas de apoyo al diagnóstico de pacientes con enfermedades raras [36]. Por ejemplo, las tecnologías de RNA-seq permiten la obtención de datos para determinar los perfiles de expresión génica en distintos tipos celulares ante diferentes condiciones [8]. Para el trabajo presentado en esta memoria, emplear datos obtenidos a partir de este tipo de tecnología de secuenciación podría ser útil para determinar los genes que se expresan de manera diferencial en los pacientes de CALD [11]. La determinación de estos genes permitiría ahondar en el conocimiento de los mecanismos moleculares implicados en este fenotipo específico de X-ALD, además del gen *ABCD1* como se ha descrito previamente.

En este trabajo proponemos un análisis de datos de expresión a partir de datos obtenidos previamente en un estudio de secuenciación de pacientes con CALD [11] empleando una herramienta que combina distintos paquetes de análisis de datos de expresión [12] y que su última versión incluye además un análisis de co-expresión de genes no llevado a cabo en los datos del estudio del que parten nuestro re-análisis. Este análisis de co-expresión se llevará a cabo para la determinación de módulos de genes con perfiles de expresión similares, de manera que puedan arrojar información

sobre qué otros posibles genes puedan estar dando lugar al fenotipo de los pacientes con CADL. Finalmente, con el objetivo de conocer qué mecanismos moleculares están implicados en el desarrollo de la enfermedad, se llevará a cabo un análisis de enriquecimiento funcional sobre los genes expresados diferencialmente como de los genes en los módulos de co-expresión determinados. Para llevar a cabo este análisis, se empleará un flujo de trabajo que llevará a cabo las funciones de pre-procesamiento de las lecturas de manera automática, así como el análisis de expresión diferencial, co-expresión y funcional con la herramienta DEgenes Hunter [12].

3. Objetivos del trabajo

El objetivo principal de este trabajo fin de máster (TFM) consiste en la determinación de los genes que pueden dar lugar a la adrenoleucodistrofia cerebral (CALD) mediante el re-análisis de datos de expresión (RNA-seq) con un novedoso protocolo de análisis de datos desarrollado por investigadores de la Universidad de Málaga. A continuación, se enumeran los sub-objetivos para la consecución de este TFM:

1. Búsqueda de datos de RNA-seq de pacientes con adrenoleucodistrofia cerebral (CALD) para la determinación de los genes implicados en su desarrollo.

2. Preparación del entorno de trabajo en un sistema de supercomputación para la limpieza de lecturas, alineamiento contra el genoma de referencia y análisis de expresión diferencial, co-expresión y enriquecimiento funcional con la herramienta DEgenes Hunter.

3. Comparación de las muestras de pacientes CALD versus controles para determinar los genes expresados diferencialmente en la enfermedad, así como la comparación de muestras para muestras de pacientes con el mismo tipo celular (queratinocitos) para confirmar que la expresión no depende del tipo celular sino de la propia enfermedad.

4. Análisis de los módulos de co-expresión para la determinación de genes con perfiles de expresión similares que puedan ser la causa del fenotipo de los pacientes con CALD.

5. Análisis de enriquecimiento funcional de los genes expresados diferencialmente tanto en el análisis general como en el de módulos de co-expresión para la determinación de los factores moleculares implicados en el desarrollo de la enfermedad.

4. Material y métodos

El objetivo principal de este trabajo de fin de máster consiste en la determinación de los genes implicados en el desarrollo de la adrenoleucodistrofia cerebral (CALD), un subtipo de adrenoleucodistrofia ligada al cromosoma X (X-ALD). Para ello, examinaremos las diferencias entre individuos sanos y pacientes con CALD para confirmar qué genes cambian de expresión en condiciones normales frente a aquellos que varían de expresión durante el procedimiento patológico. Todo este trabajo fue llevado a cabo gracias a los recursos de supercomputación del Centro de Supercomputación y Bioinnovación de la Universidad de Málaga (http://www.scbi.uma.es/site).

4.1. Descarga de datos

Para este trabajo se seleccionó un experimento cuyo set de datos incluía 18 muestras [11]: nueve de ellas pertenecientes a niños con adrenoleucodistrofia cerebral (CALD) y las nueve muestras restantes correspondientes a individuos sanos o controles (WT). En total se tomó muestra de tres pacientes y tres controles, y se obtuvieron tres réplicas de cada uno de ellos. Los datos pertenecen al estudio se pueden encontrar en el repositorio público de datos genómicos funcionales *Gene Expression Omnibus* (GEO, código de acceso: GSE108012) [11]. En concreto, se eligió este estudio por realizar análisis de RNA-seq sobre células de la barrera hematoencefálica. Como se describió anteriormente, uno de los mecanismos que se desconocen sobre la CALD es cómo y porqué se produce un incremento en permeabilidad del endotelio cerebral que constituye la barrera hematoencefálica, hecho que no sucede en los pacientes con adrenomielopatía, el otro fenotipo de X-ALD.

Según se detalla en el estudio [11], las muestras se obtuvieron de forma no invasiva. Esto significa que no proceden directamente de la barrera hematoencefálica, sino de diversos tejidos accesibles, entre los que se incluyen fibroblastos, queratinocitos y células urinarias. Por las características de la enfermedad y como afecta a los pacientes, el tipo de células cuya expresión génica con mayor interés para analizar fueron las células endoteliales de la barrera hematoencefálica. Este tipo de células se consiguieron reprogramando otros tipos celulares a su condición de pluripotencia, para posteriormente reprogramarlas como células endoteliales de la barrera hematoencefálica [11]. Por lo tanto, existen una seria de variables externas propia de cada muestra (Tabla 1) que se usaron también en el análisis de expresión diferencial.

Tabla 1. Tabla con las variables de cada muestras. Las variables de las muetsras incluyen el sexo, la procedencia de la célula, el método de reprogramación y los factores reprogramadores. Traducido de [11].

Grupo	Sexo	Tipo de célula derivada	Método	Factores reprogramadores
ccALD1	Masculino	Fibroblastos	Retrovirus	OCT4, SOX2, KLF4, c-MYC
ccALD2	Masculino	Fibroblastos	Retrovirus	OCT4, SOX2, KLF4, c-MYC
ccALD3	Masculino	Queratinocitos	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT1	Femenino	Queratinocitos	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT2	Masculino	Células de la orina	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT3	Masculino	CD34+ células de la médula ósea	Sendai virus	OCT4, SOX2, KLF4, c-MYC

El ARN utilizado para el análisis se aisló utilizando el RNeasy Mini Kit (Qiagen), las librerías se generaron con TruSeq Stranded mRNA Sample Preparation kit (Illumina). Se generaron lecturas de 150 pb de longitud de extremos emparejados (*paired-end*) utilizando Illumina MiniSeq [11].

4.2. Flujo de trabajo, fase I: limpieza y alineamiento de lecturas

El proceso de análisis de las lecturas se llevó a cabo mediante un flujo de trabajo creado con la herramienta AutoFlow [37] en el que se incluyen las etapas de pre-procesamiento de las lecturas, así como el alineamiento contra el genoma de referencia, análisis de expresión diferencial, co-expresión y funcional presentados en este trabajo de fin de máster. A continuación, se detallarán las herramientas empleadas en dicho flujo de trabajo para las distintas etapas del análisis de las lecturas descargadas.

4.2.1. Limpieza de lecturas

Sobre los archivos de lecturas descargadas se llevó a cabo el proceso de limpieza empleando la herramienta SeqTrimBB, una adaptación de la herramienta SeqTrim [38] para lecturas Illumina y basada en BBMap (https://jgi.doe.gov/data-and-tools/bbtools/). Se emplearon los parámetros por defecto de la herramienta para poder llevar a cabo la limpieza de las lecturas a excepción del tamaño mínimo de las lecturas, que fue ajustado a 65. Del mismo modo, se llevó a cabo un análisis de las lecturas antes y después de la etapa de limpieza con la herramienta FastQC [39].

4.2.2. Alineamiento contra genoma de referencia y obtención de tabla de conteo

Una vez realizada la limpieza de las lecturas se procedió a su alineamiento frente al genoma de referencia empleando la herramienta STAR [40]. La versión del

ensamblaje humano para llevar a cabo esta etapa de alineamiento fue la GCRh38. Los parámetros para llevar a cabo el alineamiento se mantuvieron por defecto.

A partir de estos resultados se obtuvieron las tablas de conteo para cada una de las muestras. Esas tablas de conteo se fusionan en una sola que incluye por cada gen el número de conteos obtenidos en cada muestra (archivo final_counts.txt) y que se empleará para el análisis de expresión diferencial y de co-expresión. La tabla se encuentra accesible a través del siguiente enlace de Google Drive: https://drive.google.com/file/d/1mjiZ6-xg67WkleH01PuMZ7uIU799Ecz9/view . Asimismo, el flujo de trabajo empleado en este trabajo de fin de máster incluye una ejecución de la herramienta Qualimap [41] para confirmar la calidad de las lecturas de las muestras una vez llevado a cabo el alineamiento. Para facilitar una visualización general del estado de todas las muestras se genera un archivo HTML que engloba todos estos resultados de calidad. El archivo se encuentra disponible a través del siguiente enlace: https://drive.google.com/file/d/140XYoZCH6ce6irEMTsX9UzUSotqGVNhg/view?usp=sharing.

4.3. Flujo de trabajo, fase II: DEgenes Hunter

Una vez llevada la limpieza y alineamiento de lecturas frente al genoma de referencia hasta la obtención de la tabla de conteos, se procedió a la determinación de genes expresados diferencialmente (en inglés *differentially expressed genes*, DEG).

La suite de herramientas ExpHunterSuite (F. Jabato y colaboradores, *Clinical and Translational Medicine*, en revisión), es una adaptación y mejora de la herramienta DEgenes Hunter [12] y consiste en una serie de programas diseñados para el análisis de expresión, co-expresión y funcional de datos transcriptómicos obtenidos mediante la tecnología RNA-seq. Se encuentra disponible en GitHub (https://github.com/seoanezonjic/ExpHunterSuite) para su descarga e instalación. Para este trabajo haremos uso de los dos programas principales que tiene dicha suite (y que llamaremos DEgenes Hunter para simplificar): degenes_Hunter.R (análisis de expresión diferencial y co-expresión) y functional_Hunter.R (análisis de enriquecimiento funcional).

DEgenes Hunter precisa de una tabla de conteo y un archivo de *targets* para poder llevar a cabo su función. En este caso, la tabla de conteo se obtiene mediante la fusión de las distintas tablas de conteo obtenidas para las muestras analizadas como se describió en el apartado 4.2.2. Por otro lado, el archivo de *targets* corresponde con un archivo tabulado donde se especifica el diseño experimental para que las herramientas de DEgenes Hunter puedan llevar a cabo el análisis diferencial correctamente. Este diseño experimental consiste en definir qué muestras son controles y diferenciarlas de cuáles son tratamientos (o mutantes), así como cualquier variable externa que sea susceptible de estar relacionada con la expresión diferencial de los genes de las muestras. En la Figura 2 se detalla el archivo de *targets* empleado

en este TFM, utilizando para su diseño la información disponible en el artículo de partida en el que se basa este trabajo [11].

sample	group	sex	derived_cell_type	delivery_method
WT1_rep1	ctr	female	ketarinocytes	retrovirus
WT1_rep2	ctr	female	ketarinocytes	retrovirus
WT1_rep3	ctr	female	ketarinocytes	retrovirus
WT2_rep1	ctr	male	urine_cells	retrovirus
WT2_rep2	ctr	male	urine_cells	retrovirus
WT2_rep3	ctr	male	urine_cells	retrovirus
WT3_rep1	ctr	male	CD34_bone_marrow_cells	sendai_virus
WT3_rep2	ctr	male	CD34_bone_marrow_cells	sendai_virus
WT3_rep3	ctr	male	CD34_bone_marrow_cells	sendai_virus
ccALD1_rep1	treat	male	fibroblasts	retrovirus
ccALD1_rep2	treat	male	fibroblasts	retrovirus
ccALD1_rep3	treat	male	fibroblasts	retrovirus
ccALD2_rep1	treat	male	fibroblasts	retrovirus
ccALD2_rep2	treat	male	fibroblasts	retrovirus
ccALD2_rep3	treat	male	fibroblasts	retrovirus
ccALD3_rep1	treat	male	ketarinocytes	retrovirus
ccALD3_rep2	treat	male	ketarinocytes	retrovirus
ccALD3_rep3	treat	male	ketarinocytes	retrovirus

Figura 2. Archivo de targets. Archivo que utiliza DEgenes Hunter para llevar a cabo su función.

4.3.1. Análisis de calidad, expresión diferencial y co-expresión

En primer lugar, DEgenes Hunter analiza las muestras que se emplean para el análisis a partir de la información almacenada en la tabla de conteo. Se incluye un cálculo de la correlación entre las muestras para determinar si alguna presenta diferencias frente a las demás. Igualmente, el análisis devuelve unos mapas de calor (*heatmaps*) y dendogramas para observar el agrupamiento de las muestras previa normalización de los datos de la tabla de conteo y después de su normalización.

Con respecto a la normalización de los datos, DEgenes Hunter emplea la función *rlog* del paquete DESeq2 [42]. Este método normaliza empleando el total de lecturas de la muestra, lo que minimiza las diferencias entre muestras que tienen pocos conteos para un gen.

Asimismo, se lleva a cabo un análisis de componentes principales (ACP) para ver cómo se agrupan las muestras. El ACP es una técnica estadística que permite simplificar la complejidad de las muestras dada su alta cantidad de variables, con el fin de condensar estas variables en solo una serie de componentes. Esto hace que sea empleado, sobre todo, en el análisis exploratorio de datos. En este caso, el ACP convierte las variables (genes) en un número mucho más reducido de nuevas

variables (componentes) fácilmente representados mediante un gráfico. Las dos primeras componentes suelen ser las que expliquen la mayor parte de la varianza en los datos y gracias su representación gráfica se pueden observar las diferencias principales entre las muestras y cómo se agrupan según su grado de similitud.

Todos los resultados del control de calidad se engloban en un informe HTML junto con los resultados del análisis de expresión diferencial y de co-expresión que se explica a continuación.

Para llevar a cabo el procedimiento de análisis de expresión, DEgenes Hunter integra cuatro algoritmos (o paquetes) de análisis de expresión diferencial: edgeR [43], lima [16], NOISeq [44] y DESeq2 [42]. La definición de parámetros se dejó por defecto y para este análisis se activaron los cuatro algoritmos de expresión. Estos paquetes requieren tanto la tabla de conteo como el archivo de targets donde se especifique las muestras control (*ctrl*, individuos sanos) y tratamiento (*treat*, pacientes ALD) para llevar a cabo el análisis de expresión diferencial. Los genes se determinan como prevalente/posible DEG dependiendo del número de herramientas que hayan detectado al mismo gen como DEG. Por ejemplo, si un gen es detectado por los cuatro paquetes de expresión, se clasificará como prevalente DEG, y en el caso de que haya sido determinado por menos de cuatro, se clasificará como posible DEG. Del mismo modo, para que un gen se considere diferencialmente expresado, tiene que cumplir que su p-valor ajustado sea menor de 0.05 y que su logaritmo de tasa de cambio (logFC) sea mayor o igual que 1.

Los genes son etiquetados por DEgenes Hunter de la siguiente forma: (1) descartados, si no superan los umbrales límite de expresión; (2) prevalentes, si su expresión ha sido significativamente detectada por los cuatro paquetes de análisis de expresión diferencial; (3) posible, si su expresión ha sido significativamente detectada por al menos uno de los paquetes de análisis de expresión diferencial.

Con respecto al análisis de co-expresión, DEgenes Hunter implementa la librería WGCNA (*weighted gene co-expression network analysis*) [17] para localizar módulos de genes que presentan un perfil de expresión similar. Este análisis sirve para determinar genes que, en análisis canónicos de expresión diferencial, podrían haber sido descartados por no superar los umbrales de corte y que, sin embargo, su cambio de expresión sí pudiera ser significativo. Los parámetros para llevar a cabo este análisis se dejaron por defecto.

Todos los resultados, tanto el del análisis de expresión diferencial como el de co-expresión, y el de calidad de los resultados, se engloban en un archivo HTML que se discutirá ampliamente en los resultados de este trabajo.

4.3.2. Análisis de enriquecimiento funcional

Finalmente, sobre los DEG determinados tanto por el análisis de expresión diferencial como de co-expresión se ejecuta un análisis de enriquecimiento funcional gracias a los algoritmos implementados en DEgenes Hunter. Este análisis de

enriquecimiento funcional devuelve un archivo HTML general con el análisis de enriquecimiento funcional para los DEG determinados del análisis de expresión diferencial, otro para el enriquecimiento funcional del análisis de co-expresión y otros por cada uno de los módulos de genes detectados. Las anotaciones funcionales empleadas para este procedimiento incluyen términos de las tres sub categorías en las que se divide la Gene Ontology (GO, Biological Process, Molecular Functions y Cellular Component) [19], en términos de la Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] y de Reactome [46].

El análisis de enriquecimiento funcional nos permite realizar una interpretación más precisa de los resultados tanto de expresión diferencial como de co-expresión, como se expone en el siguiente capítulo.

4.4. Tratamiento de las muestras.

Para determinar que los genes expresados diferencialmente en el experimento en el que se basa el presente TFM, se van a analizar las muestras del estudio publicado en [11], creando dos grupos de análisis. Por un lado, se seleccionarán para su estudio todas las 18 muestras pertenecientes a nueve controles y nueve enfermos, enfrentando controles contra pacientes para ver qué genes se expresan diferencialmente en ambos tipos de muestras sin tener en consideración a qué tipo celular pertenecen las muestras. Igualmente, con el fin de confirmar que los genes expresados diferencialmente no son exclusivos del tipo celular del que se toma la muestra, se hará el mismo análisis, pero seleccionando aquellas muestras pertenecientes a queratinocitos (tres muestras control contra tres muestras paciente). Se escogió este tipo celular al ser el único con muestras representativas de ambos grupos (controles y pacientes).

5. Resultados

En este capítulo se expondrán de forma detallada y se explicarán los resultados obtenidos del análisis de los datos, con el objetivo de esclarecer los mecanismos moleculares que subyacen a la CALD y que hacen que se manifieste de una forma mucho más severa que la AMN, el otro fenotipo de la X-ALD. En primer lugar, se mostrarán los resultados correspondientes al análisis de alineamiento de las muestras para confirmar que este procedimiento se ha llevado a cabo satisfactoriamente. A continuación, se mostrarán los resultados obtenidos para el análisis de controles contra pacientes de todos los tipos celulares y en segundo lugar se explicarán los resultados para la comparativa de controles contra pacientes del mismo tipo celular (queratinocitos). Las condiciones analizadas están disponibles en la Tabla 1 (tipo celular, método de reprogramación celular, condición y sexo).

5.1. Análisis de calidad de las lecturas

Se llevó a cabo un análisis de calidad de las lecturas tras realizar el pre-procesamiento (limpieza) y alineamiento contra el genoma de referencia para confirmar si hubo pérdidas significativas de información tras llevar a cabo ambos prodecimientos. Debido a su tamaño, no se ha considerado oportuno incluir las figuras del antes y después de la limpieza, pero en el informe en https://drive.google.com/file/d/140XYoZCH6ce6irEMTsX9UzUSotqGVNhg/view?usp=sharing se observa que se mantiene un tamaño de lectura homogeneizado no inferior a 100 pb tras el procedimiento (apartado *Length distribution after trimming*). Del mismo modo, no se observan pérdidas de lecturas significativas una vez llevada a cabo la limpieza de las mismas (Figura 3).

Figura 3. Número total de lecturas antes y después de la limpieza. Representación del total de lecturas antes (rojo) y después (azul) del proceso de limpieza llevado a cabo con SeqTrimBB.

En cuanto al alineamiento de las lecturas (Figura 4), se observa que la gran mayoría de las mismas alineaban de forma única contra la referencia. Todas aquellas lecturas que mapearon contra múltiples localizaciones cromosómicas fueron

descartadas de este análisis. Asimismo, del total de lecturas por muestra que se filtraron tras el proceso de limpieza, la gran mayoría alinearon contra una única región del genoma, y de ellas a su vez contra un gen (Figura 5). Esto nos permitió no descartar ninguna muestra para su empleo en el análisis de expresión diferencial.

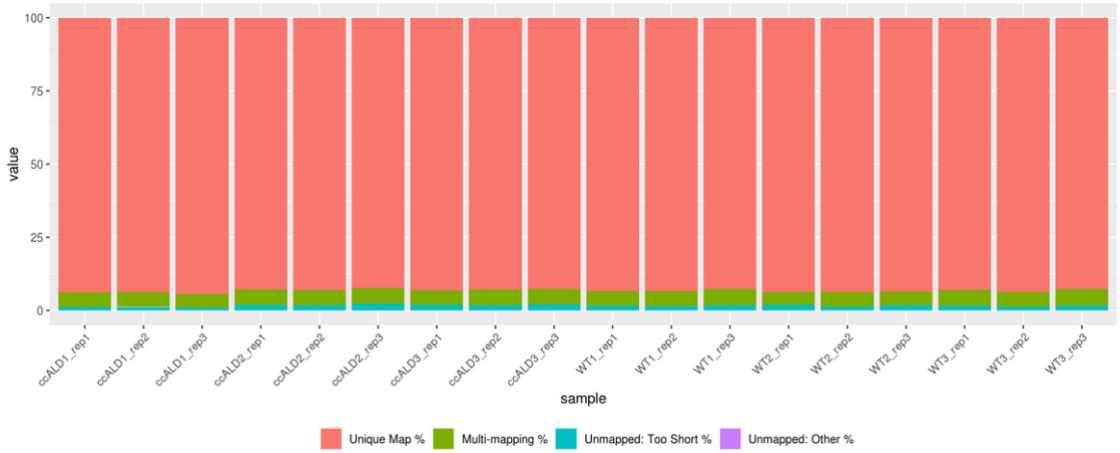

Figura 4. Representación de las lecturas alineadas contra la referencia con la herramienta STAR. Todo el porcentaje de lecturas en rojo fueron empleadas para el análisis para cada una de las muestras (eje X). Las lecturas que alinearon contra más de una región cromosómica fueron descartadas (verde), así como las que no superaron el umbral de tamaño mínimo (azul).

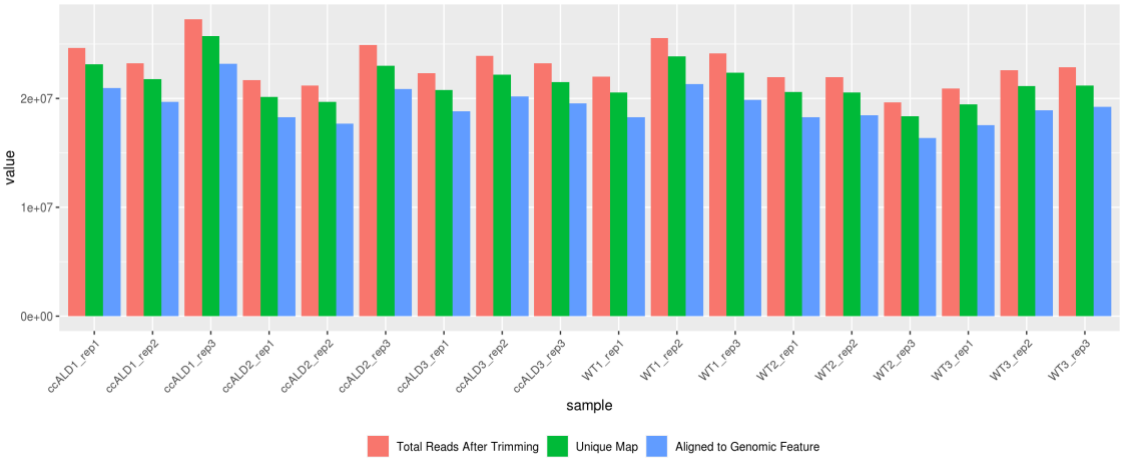

Figura 5. Representación del total de lecturas por muestra tras el proceso de limpieza (rojo), las que alinearon contra una única región cromosómica (verde) y de ellas las que alinearon contra un elemento genómico o gen (azul).

.

5.2 Exploración de los datos: controles contra pacientes (todos los tipos celulares)

El informe HTML de análisis de expresión diferencial por DEgenes Hunter (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1uBaFuauN6qPJHy0yLqpOWetpN12Y0loJ/view?usp=sharing) nos arroja la información necesaria para realizar una exploración de los datos

previa a la búsqueda de genes diferencialmente expresados. Gracias a ello, podemos hacernos una idea de la calidad de las muestras y de su agrupamiento según sean controles o pacientes para llevar a cabo un correcto análisis de expresión diferencial.

5.2.1. Control de calidad

A continuación, se muestran los diagramas de correlación que comparan los niveles de expresión de todos los genes entre las diferentes muestras, para todos los controles (Figura 6) y todas las muestras de tratamiento (Figura 7). Los valores situados a la derecha muestran el coeficiente de correlación de Pearson que sirve para comparar las similitudes entre dos grupos de muestras. Por norma general, cuando dos muestras comparten un coeficiente de correlación (R) mayor o igual a 0,96 pertenecen al mismo grupo. Valores bajos de correlación (R < 0.8) para todos los valores una muestra en concreto indica posibles problemas con la misma. Para estas muestras no se han detectado valores de R significativamente bajos para ninguna de las muestras, por lo que podemos descartar que ninguna comprometa los resultados del análisis de expresión diferencial.

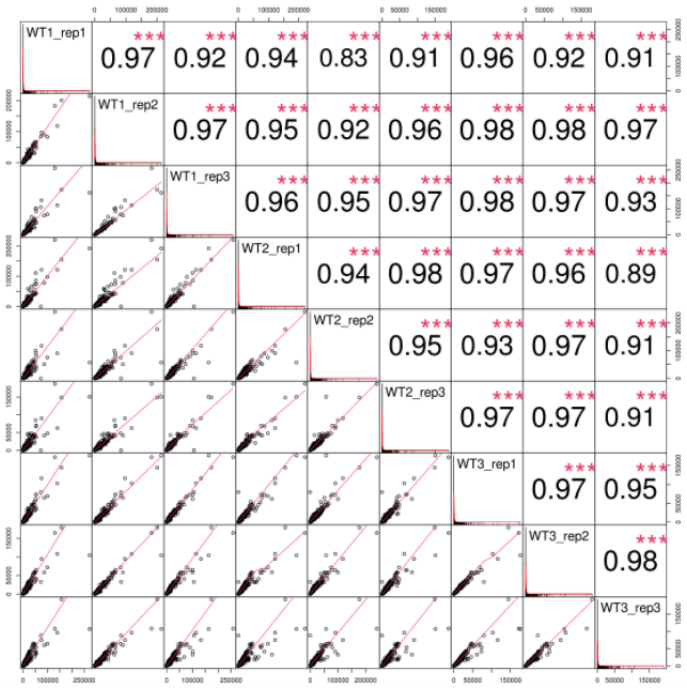

Figura 6. Gráficos de correlación del control de calidad para las muestras de controles. En esta figura se representan los valores de correlación para las distintas muestras de controles (WT) comparadas. Las réplicas (rep) que están dentro del mismo grupo tienden a tener coeficientes de correlación de Pearson igual o mayor a 0.96.

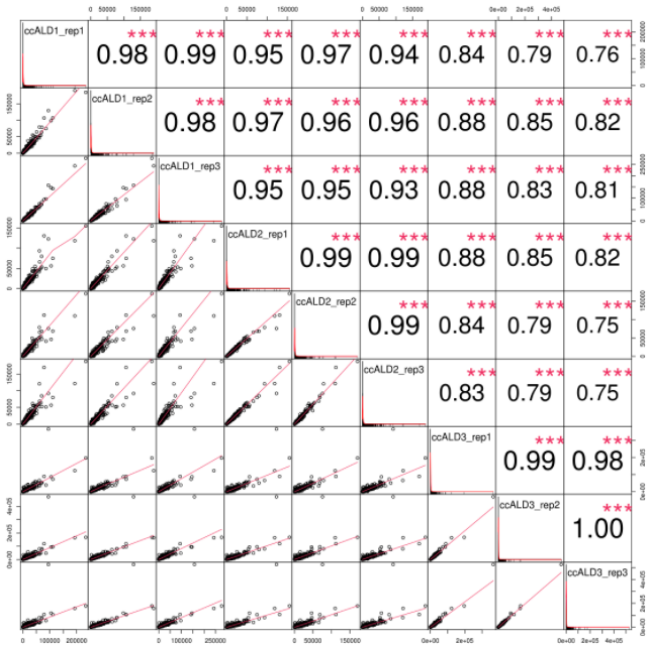

Figura 7. Gráficos de correlación del control de calidad para muestras de pacientes. Se representan los valores de correlación para las distintas muestras de pacientes (ccALD) comparadas. Las réplicas (rep) que están dentro del mismo grupo tienden a tener coeficientes de correlación de Pearson igual o mayor a 0.96.

Otra forma de visualizar la relación entre muestras es el uso combinado de mapas de calor (*heatmaps*) y dendrogramas (*clusters*). DEgenes Hunter genera dos mapas de calor previa y post normalización (Figura 8 y 9, respectivamente) de los valores de la tabla de conteo.

Con respecto a los resultados obtenidos para el ACP, en el estudio se observó una separación evidente entre las muestras WT y CALD apoyada por la componente con mayor porcentaje de la varianza (eje x, 58%) (Figura 10). Asimismo, hay que mencionar que las muestras que han sido tomadas en queratinocitos (WT1 y ccALD3) están separadas de los demás tipos celulares como se refleja en el porcentaje de la varianza del segundo componente (eje y, 13%).

Las muestras analizadas en el estudio proceden de tipos celulares diferentes (Tabla 1). Además, el vector que utilizaron para conseguir la des-diferenciación varía según las muestras. Por lo que tenemos disparidad en las variables. De este modo podemos suponer que muchos de los cambios en la expresión de los genes son fruto del tipo celular del que procedía la muestra y no de la enfermedad.

En este punto debemos mencionar que estos resultados supusieron la modificación reflejada en el plan del proyecto, la cual ha consistido en la repetición del análisis con estas muestras tomadas en queratinocitos (WT1 y ccALD3), siguiendo el mismo protocolo aplicado para el total de las muestras. Se sabe que el uso de un solo control y tratamiento no es representativo para poder realizar un análisis de expresión diferencial significativo, pero en este caso y como se mencionó en la sección de Material y Métodos, la información obtenida se empleará para determinar si los genes

diferencialmente expresados en el análisis de todas las muestras se deben a la diferencia entre controles y pacientes y no al tipo celular.

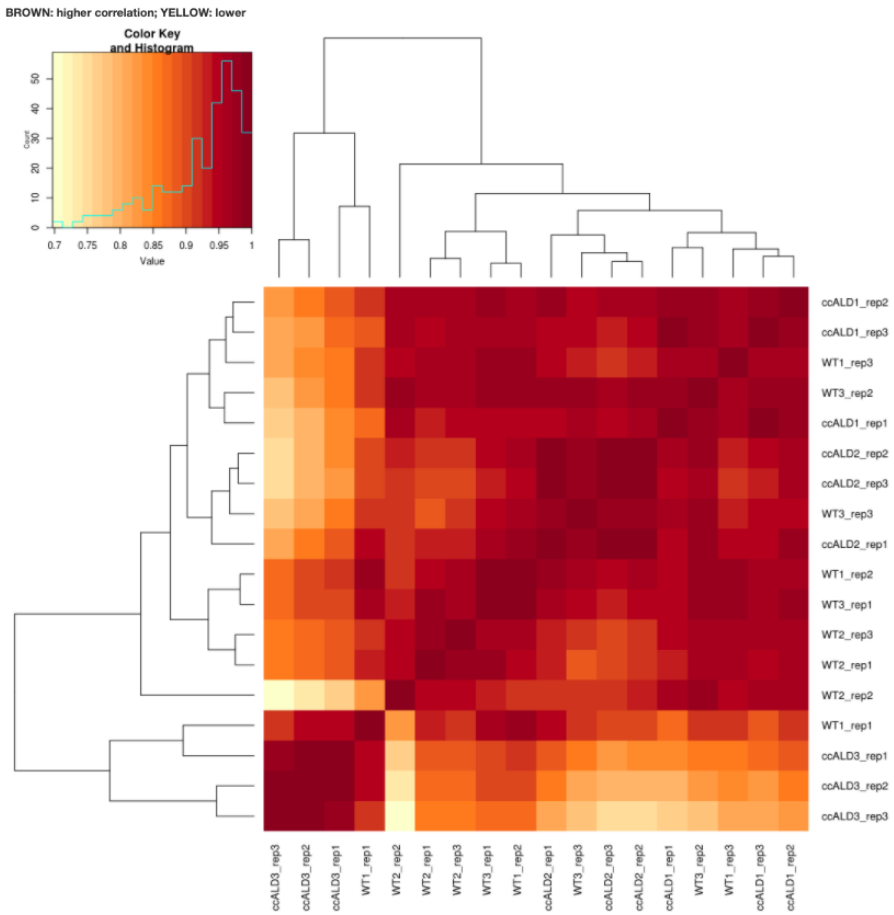

Figura 8. Mapa de calor y dendograma de agrupamiento de las muestras previa normalización. Se observa que antes de normalizar los valores de la tabla de conteo, las muestras se agrupan según réplicas (rep) pero no se consigue una clara distinción entre grupos control (WT) y paciente (ccALD). Los tonos más oscuros indican una mayor correlación entre las muestras, mientras que los más claros una menor correlación.

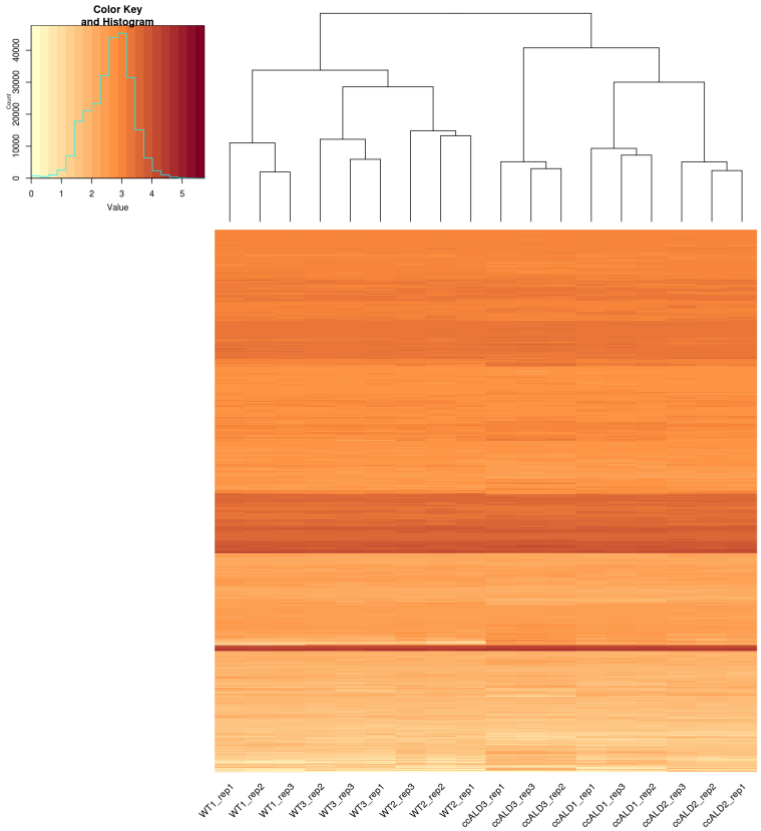

Figura 9. Mapa de calor y dendograma de agrupamiento de las muestras tras normalización. Se observa que después de normalizar los valores de la tabla de conteo, las muestras se agrupan según réplicas (rep) y se observa una clara distinción entre grupos control (WT) a la izquierda del gráfico frente a las muestras paciente (ccALD), agrupadas a la derecha. Los tonos más oscuros indican una mayor correlación entre las muestras, mientras que los más claros una menor correlación.

.

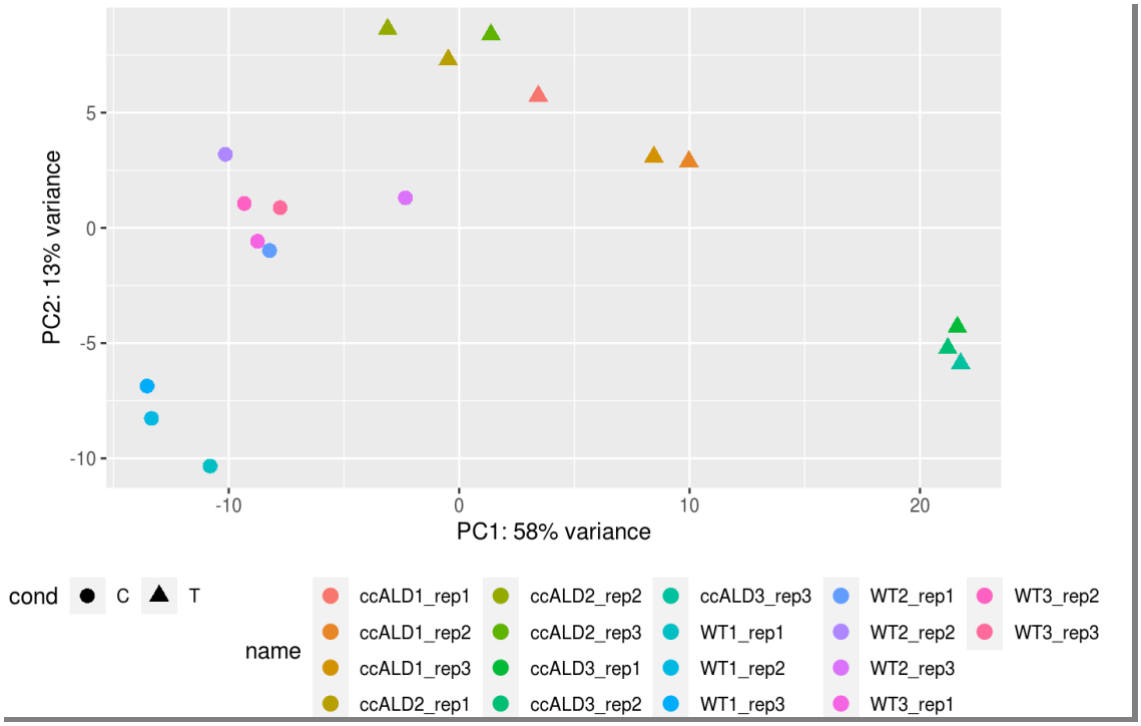

Figura 10. Análisis de componentes principales. DEgenes Hunter lleva a cabo este análisis usando los valores de recuento después de la normalización de rlog del paquete DESeq2. Se observa una clara separación de los grupos control (C, círculos) y pacientes (T, triángulos), apoyada por la componente con el mayor porcentaje de la varianza (PC1, 58%).

5.2.2. Análisis de expresión diferencial y co-expresión

Del análisis de las nueve muestras WT contra las correspondientes nueve muestras ccADL, de un total de 60.656 genes anotados para la versión GCRh38 del ensamblaje del genoma de *H. sapiens*, DEgenes Hunter detectó expresión diferencial en 990 genes con al menos uno de los paquetes de análisis de expresión diferencial (genes etiquetados como posibles DEG) y 667 con los cuatro paquetes al mismo tiempo (etiquetados como DEG prevalentes). Cabe destacar que los paquetes que detectaron un mayor número de DEG por separado fueron limma (79), seguido de NOISeq2 (14). Los gráficos de análisis de expresión (MA y volcano plots) del informe generado tras el análisis de expresión diferencial (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1uBaFuauN6qPJHy0yLqpOWetpN12Y0loJ/view?usp=sharing) no muestran anomalías significativas en los resultados de los distintos análisis.

Con respecto al análisis de co-expresión llevado a cabo mediante el algoritmo WGCNA implementado en DEgenes Hunter, se han obtenido 39 clústeres distintos, entre los cuales caben citar los módulos 21, 29, 35, 23, 37 y 33. Esta información aparece disponible en el dendograma de correlación absoluta correspondiente al apartado *Eigen values clustering* del informe de análisis de expresión general (Figura 11).

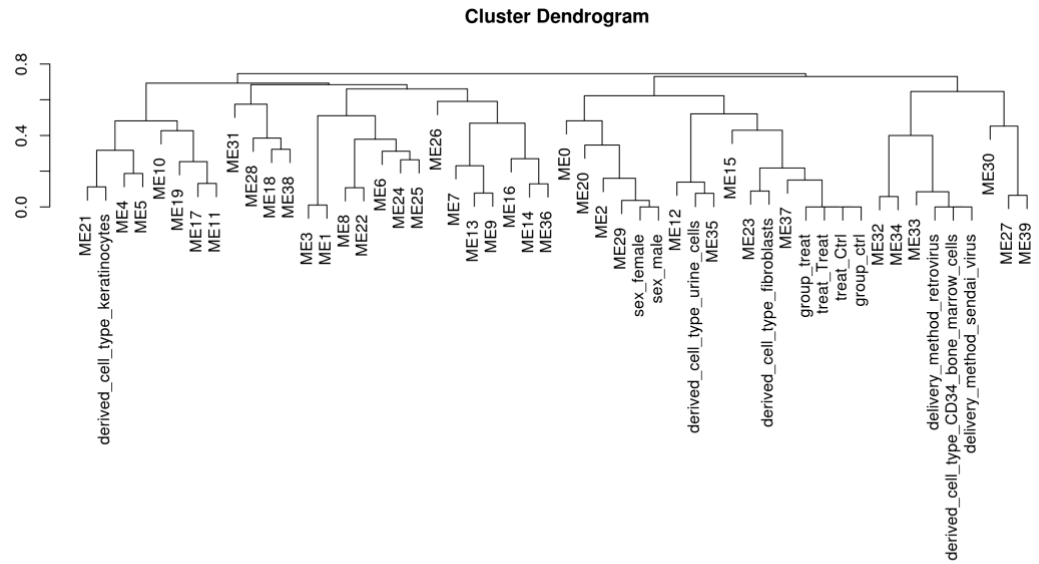

Figura 11. Dendrograma de correlación absoluta. Muestra las distancias entre estos módulos junto con los factores del experimento (tipo celular, sexo, grupo…). Las distancias se han calculado utilizando correlación absoluta, por lo que cuanto más elementos cercanos, mayor correlación absoluta entre elementos.

En el módulo 21, el patrón de expresión de los genes coincide con el patrón de expresión de los genes del tipo celular correspondiente a queratinocitos. Lo mismo sucede para los módulos 35 y 23 para células del tracto urinario y fibroblastos, los dos otros tipos celulares empleados en el experimento en el que se basa este TFM [11]. En el módulo 29, el patrón de expresión de sus genes coincide con el del agrupamiento por sexo de los individuos. Por su parte, el patrón de expresión de los genes del módulo 37 coincide con el de expresión de los genes según las condiciones tratadas (controles y pacientes) y el del módulo 33 con el tipo de método utilizado para la reprogramación celular.

5.2.3. Análisis de enriquecimiento funcional.

Las anotaciones funcionales empleadas para este análisis funcional incluyen términos de las tres subcategorías en las que se divide la Gene Ontology (GO, Biological Process, Molecular Functions y Cellular Component) [19], en términos de la Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] y de Reactome [46]. Todas las categorías funcionales se calculan con CluterProfiler y los términos GO se calculan también con TopGo. Cada categoría se analiza mediante el análisis de sobrerrepresentación (ORA) y el análisis de conjuntos de genes (GSEA). En el caso del método GSEA, todos los genes se clasifican el cambio en su tasa de expresión y el algoritmo explora qué genes con tasa de expresión similar comparten un término de la categoría funcional seleccionada.

Los informes generados por DEgenes Hunter al ejecutar su módulo funcional devuelve varios archivos HTML con la información recopilada a partir de los programas de enriquecimiento funcional implementados. En este caso, se describirán los resultados obtenidos para el análisis de enriquecimiento funcional (https://drive.google.com/file/d/1iRY_BxgOFClnhxCf36eynSWHHKEIz5Wa/view?usp=sharing) obtenidos a partir del análisis de expresión diferencial general y el de los módulos de co-expresión que contienen genes cuyo patrón de expresión coincide con el de las condiciones analizadas en este experimento (tipo celular, método de reprogramación celular, condición y sexo): módulos 21 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1rikDw9-_J9WO4VmPjgLC6Jurd48-wAk_/view?usp=sharing), 29 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1huUZTnctaVSgawzSdNxVK567lBsI3aJP/view?usp=sharing), 35 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1PSEck0qPOhXCz_3ZVVbaQrB38sfbW5X-/view?usp=sharing), 23 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1iZgIw0jE347XKun_VlIoJS02oRvPDSS6/view?usp=sharing), 37 (disponibles para su descarga en el siguiente enlace: https://drive.google.com/file/d/1cS5Dq_Esrpx2A2c1BU3apgv17QS89MNi/view?usp=sharing) y 33 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1bvk9eKFHYJrQxJ1JmXzUWfWsNGYBlYst/view?usp=sharing).

Los resultados correspondientes al análisis de enriquecimiento funcional para los genes detectados con expresión diferencial muestran el top de genes con valores tanto positivos como negativos de expresión diferencial para las muestras analizadas. En el caso de los cinco genes con un mayor nivel de sobreexpresión (ordenados según la media calculada de valores de logFC positivos por los paquetes de expresión diferencial empleados, parámetro *mean_logFCs* en la tabla *Top positive* en el informe de enriquecimiento funcional), se encuentran los genes *CYP4F29P* (11.313), *CTSF (*9.741), *ZNF208 (*9.136)*, ZNF560 (*7.881) y *ZNF728 (*7.171). Con respecto a los genes con un mayor nivel de inhibición (*Top negative* en el informe de enriquecimiento funcional), se encuentran *CNTN4* (-9.136), *COL22A1* (-5.819), *USP6* (-5.803), *TCERG1L* (-5.420) y *C5orf17* (-5.233).

El informe incluye gráficos construidos a partir del método de análisis de sobrerrepresentación (ORA). Estos análisis emplean grupos de genes expresados diferencialmente (DEG) significativos (DEG sobreexpresados o inhibidos) y realiza una prueba hipergeométrica para cada término de la categoría funcional seleccionada. El gráfico representado en la Figura 11 muestra los términos significativos funcionales en orden ascendente por p-valor ajustado para las categorías pertenecientes a los términos GO dentro de la subcategoría *Molecular functions*.

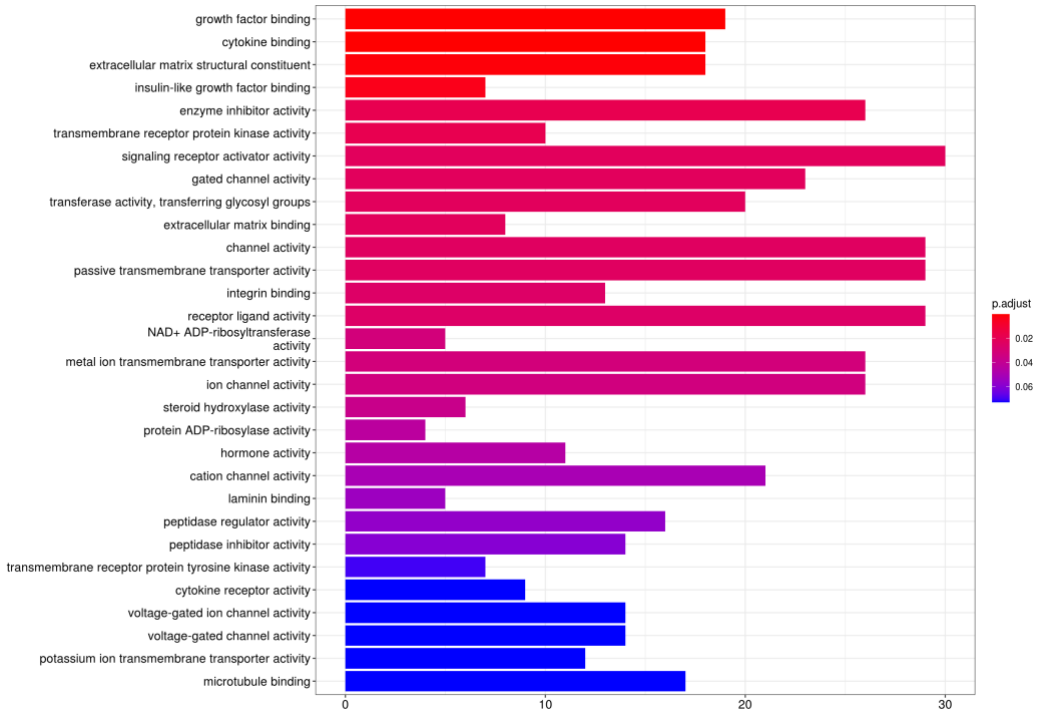

Figura 12. Gráfico ORA correspondiente a los datos GO (subcategoría Molecular functions). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

En los análisis de grupos funcionales según GO, para la categoría *Molecular functions* (GOMF), las categorías más significativas corresponden con los términos *growth factor binding*, *cytokine binding*, *extracellular matrix structural consistuent* e *insulin-like growth factor binding*, con unos p-valores próximos a 0.02.

El análisis también devuelve una red donde se conectan los términos funcionales (nodos) a través de sus genes asociados (líneas grises, cuyo el grosor representa el número de genes compartidos entre ambos términos funcionales). El tamaño de los términos funcionales muestra el número de genes conectados y el color el p-valor ajustado del término funcional. En este caso, para la categoría GOMF se observan varios módulos, entre los que están uno de términos relacionados con la actividad de canales dependientes de voltaje, otro de regulación de la actividad peptidasa, otro de recepción de ligandos, otro con actividad dependiente de ADP y un último módulo con términos relacionados con receptores transmembrana (Figura 12, módulos descritos señalados en círculos verdes).

Con respecto a la subcategoría de GO *Biological process* (GOBP), para estas muestras se han encontrado genes enriquecidos con los módulos funcionales correspondientes a los términos *epithelial cell proliferation, extracellular matrix organization, regulation of epithelial cell proliferation* y *extracellular structure organization* con el p-valor más significativo (Figura 13). La CALD produce una desmielinización intensa que produce daño axonal, modificando la morfología y estructura de la neurona para llevar a cabo correctamente su función. Por su parte, en la red de términos funcionales (Figura 14) se observan dos módulos funcionales

correspondientes a categorías relacionadas con procesos de respuesta inmune y procesos dependientes de angiogénesis, estrechamente relacionados con los genes enriquecidos para esta categoría funcional.

Por otra parte, del análisis con términos de la subcategoría GO *Cellular component* (GOCC), se puede observar en la Figura 15 que las categorías más significativas en cuanto al p-valor corresponden con los términos *collagen-containing extracellular matrix*, *membrane microdomain*, *integral component of synaptic membrane*, *synaptic membrane* e *ion channel complex*. No es extraño encontrar genes con expresión diferencial para los pacientes con CALD enriquecidos con estos términos para la subcategoría GOCC, ya que esta patología afecta al correcto funcionamiento de las neuronas. Con respecto al análisis de red (Figura 16), se observan dos módulos de términos claramente diferenciados relacionados con componentes de la membrana sináptica y regiones específicas de la membrana celular.

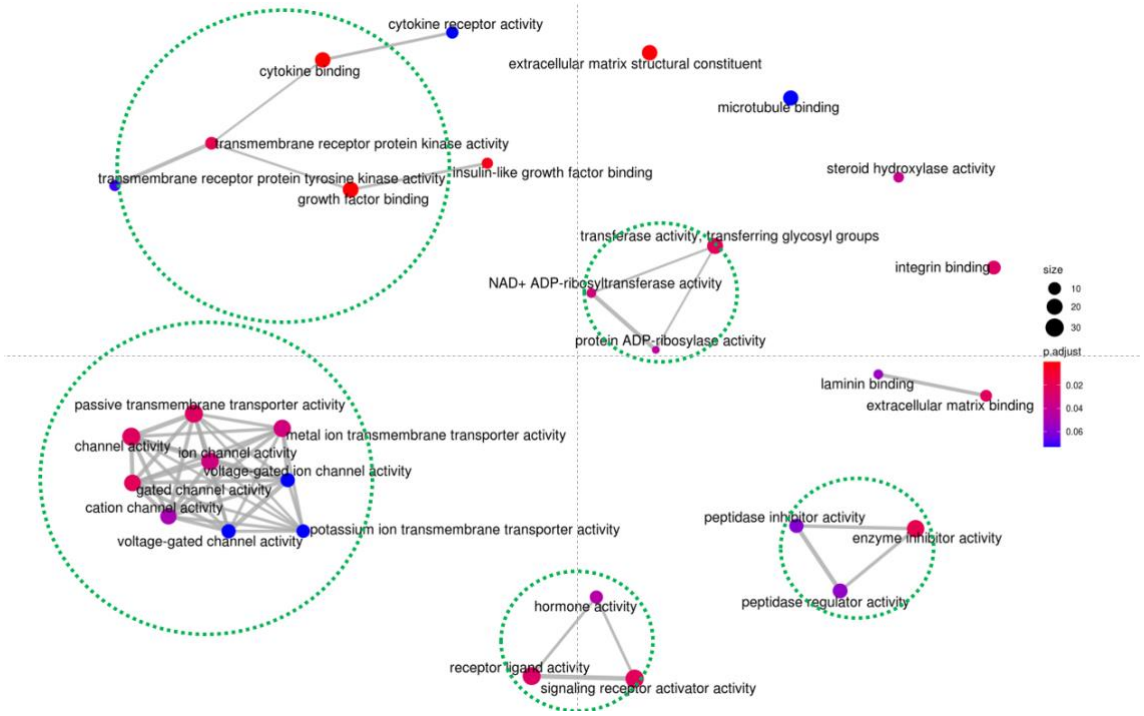

Figura 13. Mapa gráfico de redes de enriquecimiento funcional para GO Molecular functions. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional. Se ha remarcado en círculos discontinuos verdes los módulos descritos en el texto.

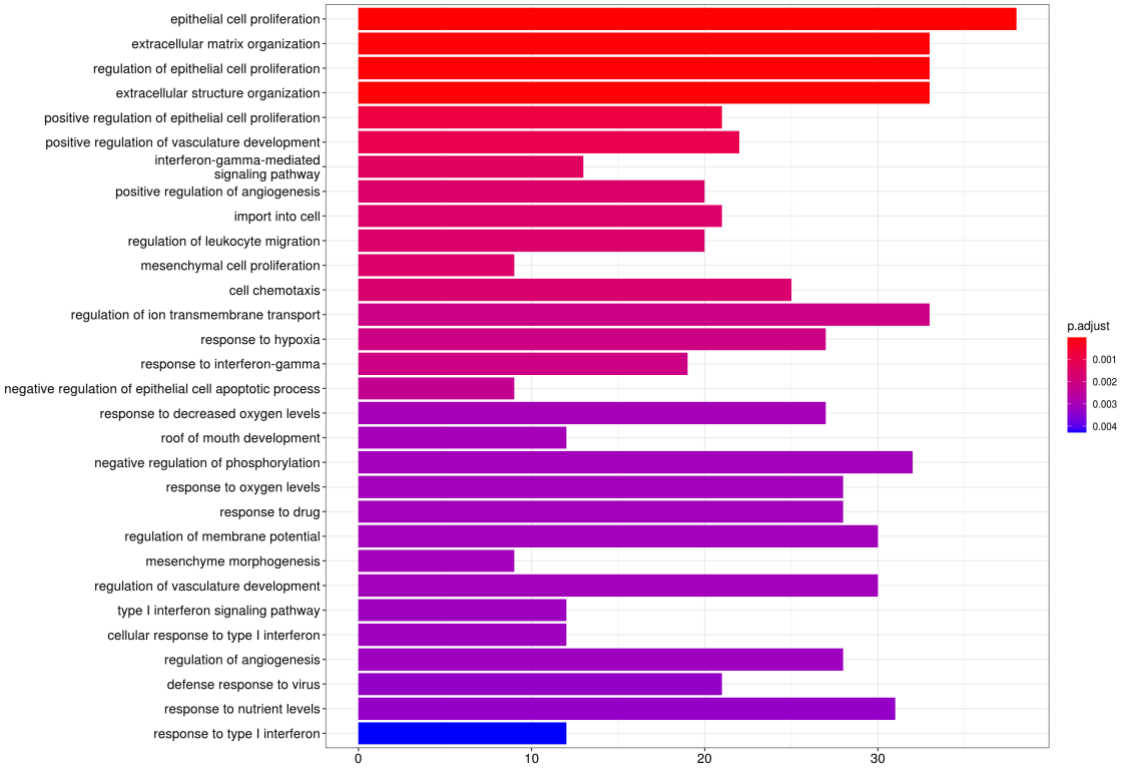

Figura 14. Gráfico ORA correspondiente a los datos GO (subcategoría Biological process). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

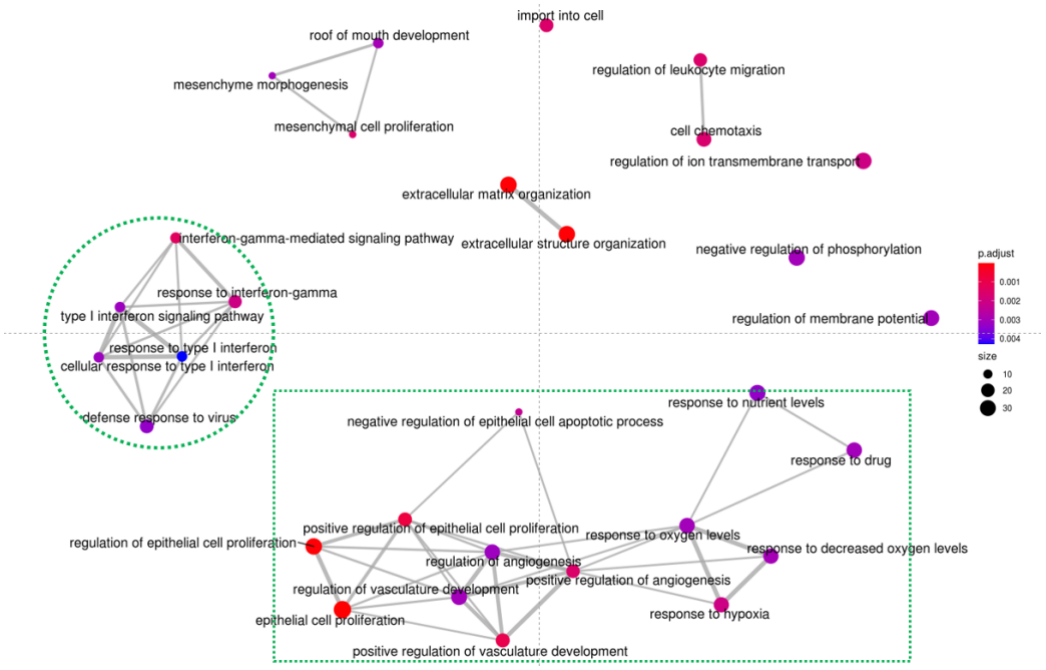

Figura 15. Mapa gráfico de redes de enriquecimiento funcional para GO Biological process. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional. Se ha remarcado en líneas discontinuas verdes los módulos descritos en el texto.

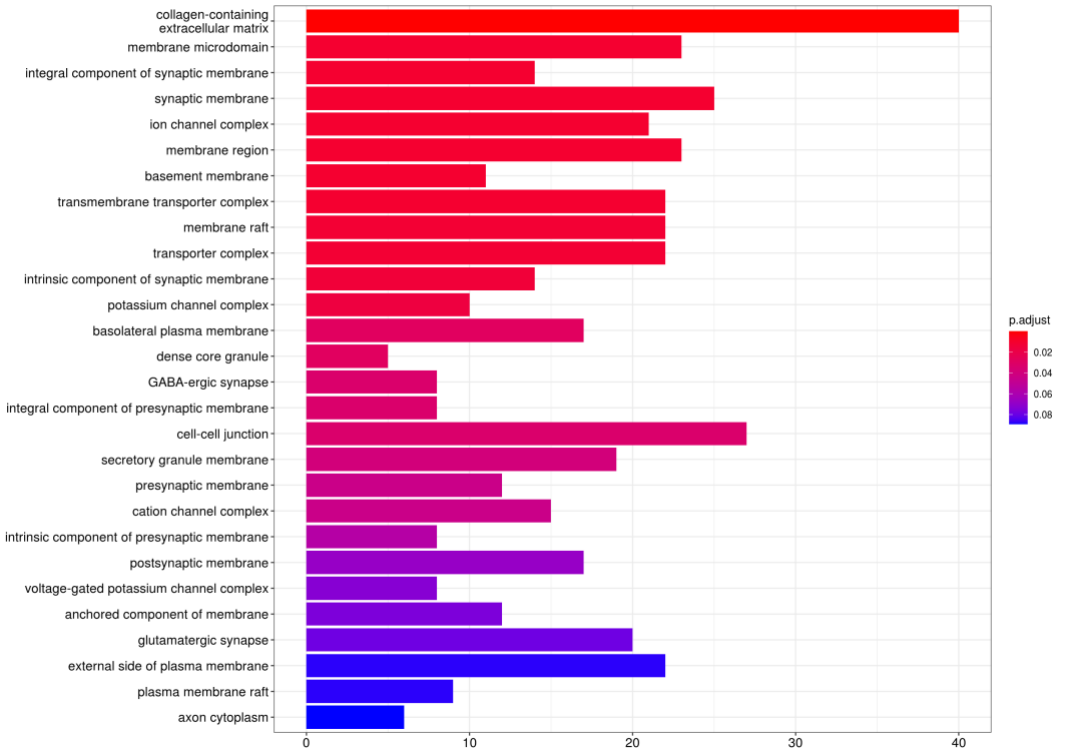

Figura 16. Gráfico ORA correspondiente a los datos de GO (subcategoría Cellular component). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

.

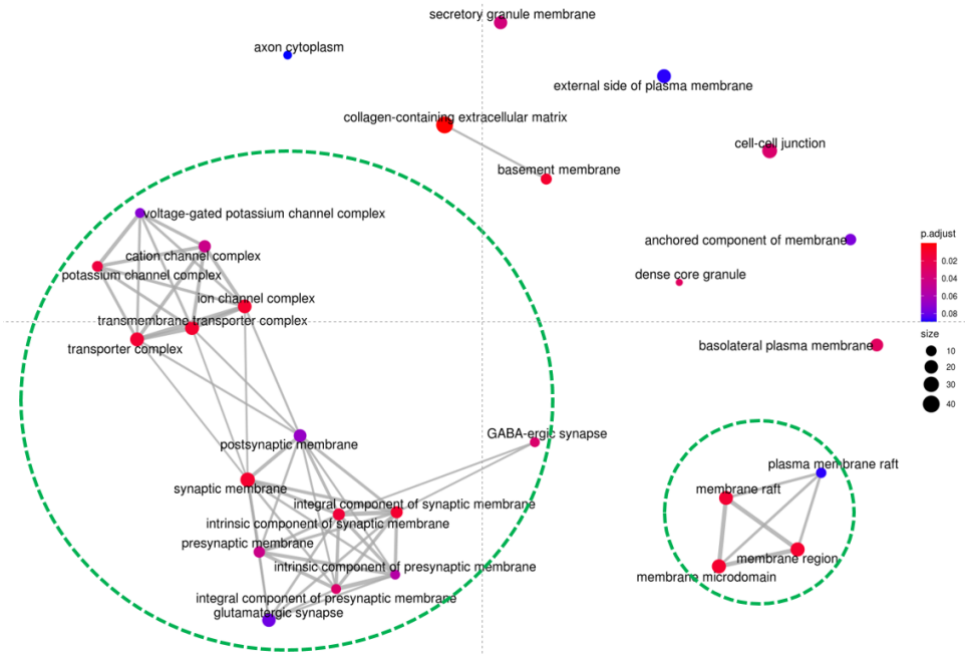

Figura 17. Mapa gráfico de redes de enriquecimiento funcional para GO Cellular component. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace

referencia al número de genes que hay asociados a esa categoría funcional. Se ha remarcado en círculos discontinuos verdes los módulos descritos en el texto.

Con respecto al análisis funcional llevado a cabo para los términos Reactome para estas muestras, en la Figura 17 se han encontrado genes enriquecidos con los módulos funcionales correspondientes a los términos *Glycoprotein hormones*, *Peptide hormone biosynthesis* e *Interferon alpha/beta signaling* como los más significativos en términos de p-valor. De ellos, el término *Interferon alpha/beta signaling* parece estar más asociado con el proceso patológico observado en pacientes CALD puesto que las muestras fueron tomadas para observar los genes expresados en la barrera hematoencefálica, y esta familia de citoquinas (interferón) ha sido estrechamente relacionada con enfermedades del SNC [47]. En cuanto al análisis de red (Figura 18), se observan dos módulos correspondientes a señalización por interferón y hormonas peptídicas.

En cuanto al análisis funcional llevado a cabo para los términos KEGG, cabe destacar que el único termino con un p-valor próximo a 0.02 es *Steroid hormone biosynthesis*, el cual es muy similar al término *Peptide hormone biosynthesis* del análisis en Reactome. No se incluyen los gráficos correspondientes al análisis de red para este análisis al no conseguirse clústeres diferenciados por el escaso número de genes enriquecidos en estos términos KEGG.

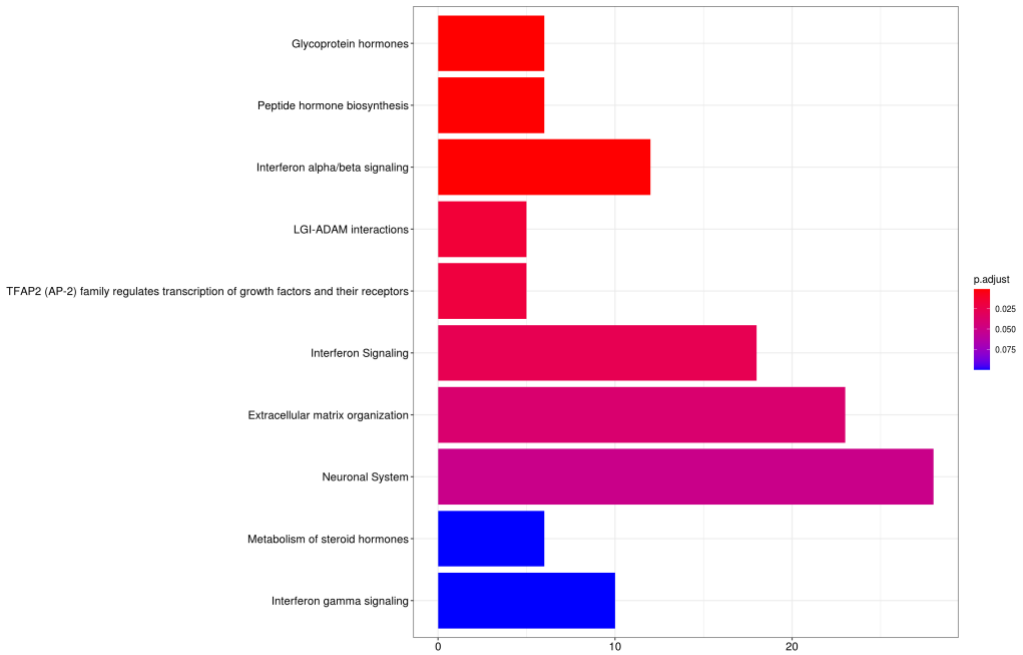

Figura 18. Gráfico ORA correspondiente a los datos de Reactome. El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

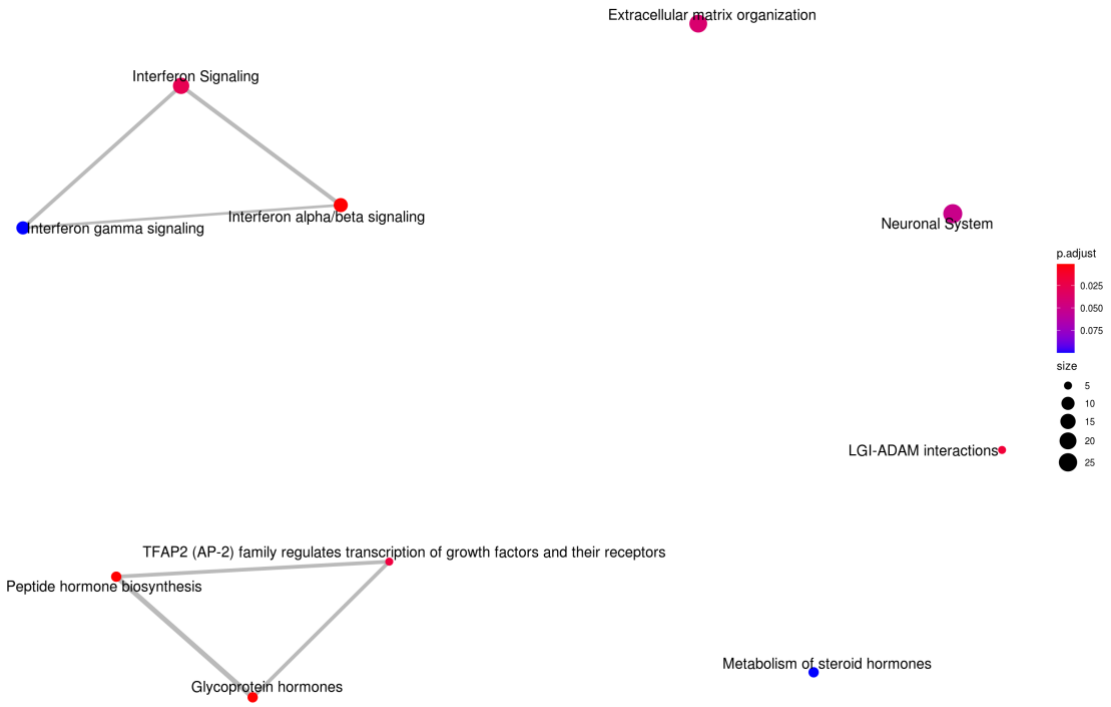

Figura 19. Mapa gráfico de redes de enriquecimiento funcional para Reactome. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional.

En resumen, en la Tabla 2 se representan los tres primeros términos de cada una de las tres categorías del análisis GO (procesos biológicos, funciones moleculares y componentes celulares) así como los del análisis KEGG, con el objetivo de observar a grandes rasgos los mecanismos moleculares que subyacen a la CALD.

Tabla 2. Resumen de los términos con mayor significancia del análisis. De izquierda a derecha: términos KEGG, términos GO de función molecular (FM), términos GO de procesos biológicos (PB), términos GO de componentes celulares (CC) y reactome.

KEGG	FM	PB	CC	Reactome
Biosíntesis de hormonas esteroideas	Unión al factor de crecimiento	Proliferación de células epiteliales	Matriz extracelular que contiene colágeno	Hormonas glicoproteicas
Ruta de señalización PPAR	Unión de citocinas	Organización de la estructura extracelular	Microdominios de membrana	Biosíntesis de hormonas peptídicas
Rutas en cáncer	Constituyente estructural de la matriz extracelular	Regulación de la proliferación de células epiteliales	Componentes integrales de la membrana sináptica	Señalización del interferón alfa/beta

Respecto a los módulos de co-expresión, podemos ver como en el módulo 23 los términos más representados GO de componentes celulares representados tienen que ver con la membrana (*outer membrane*, *organelle outer membrane* y *mitocondrial outer membrane*), en el módulo 35 con el mecansimo de *splicing* (*spliceosomal complex, U2-type presplceosome* y *presplieosome*), en el módulo 37 con los términos *interstitial matrix*, *nuclear outer membrane* y *growth cone* y por último, en el módulo 21 con términos relacionados con la ARN polimerasa.

5.3. Análisis de controles contra pacientes en queratinocitos

Durante la realización de este TFM, se observó que el diseño experimental del estudio en el que se basa [11] impide comparar muestras provenientes del mismo tipo celular para controles y pacientes, exceptuando el caso de las muestras tomadas para queratinocitos. Ante esto, y para confirmar que los genes expresados diferencialmente en el análisis de las 18 muestras es fruto de la enfermedad y no del tipo celular, se llevó a cabo la comparación entre muestras pertenecientes al mismo tipo celular.

Hay que considerar que esta comparación se llevó a cabo entre tres réplicas de dos individuos (WT1 y ccALD3) y recordar que este análisis se lleva a cabo con la única finalidad de comparar los resultados funcionales con los obtenidos para el resto de individuos del experimento con el fin de determinar si la expresión diferencial de los genes es debida a la enfermedad o al tipo celular.

5.3.1. Análisis de calidad

El análisis de calidad de las muestras de queratinocitos (WT1 y ccALD3) arrojó unos valores de correlación significativos para la comparación de todas ellas (Figura 19). Del mismo modo y como era esperable según los resultados del ACP para todas las muestras, el resultante de la comparación de WT1 y ccALD3 separó las muestras en dos grupos claramente diferenciables en su componente con mayor porcentaje de la varianza (Figura 20). El resto de análisis de calidad de los datos está disponible en el archivo HTML correspondiente al análisis de expresión para esta comparación, disponible para su descarga a través del siguiente enlace: https://drive.google.com/file/d/1uBaFuauN6qPJHy0yLqpOWetpN12Y0loJ/view?usp=sharing.

5.3.2. Análisis de expresión diferencial y co-expresión

Para esta comparación, se detectaron un total de 1900 genes expresados diferencialmente por los cuatro algoritmos empleados. Se debe remarcar que este resultado es fruto de la poca variabilidad entre ambas muestras, surgiendo como diferencialmente expresados genes propios de la enfermedad y del fondo genético del propio individuo.

Con respecto al análisis de co-expresión, se detectaron hasta 23 módulos distintos. En los módulos 1 y 2, el patrón de expresión de los genes coincide con el

patrón de expresión de los genes del tipo celular correspondiente al grupo de comparación analizado (Figura 20).

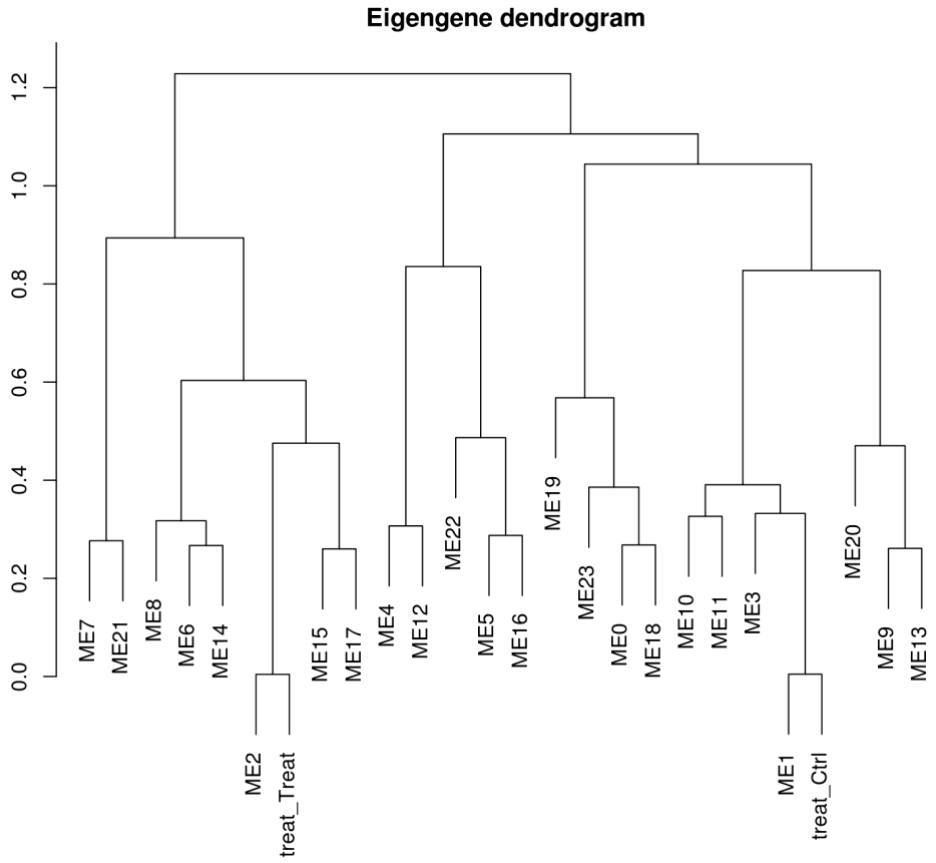

Figura 20. Dendrograma de correlación absoluta. Muestra las distancias entre estos módulos junto con los factores del experimento (tipo celular, sexo, grupo…). Las distancias se han calculado utilizando correlación absoluta, por lo que cuantos más elementos cercanos, mayor correlación absoluta entre elementos.

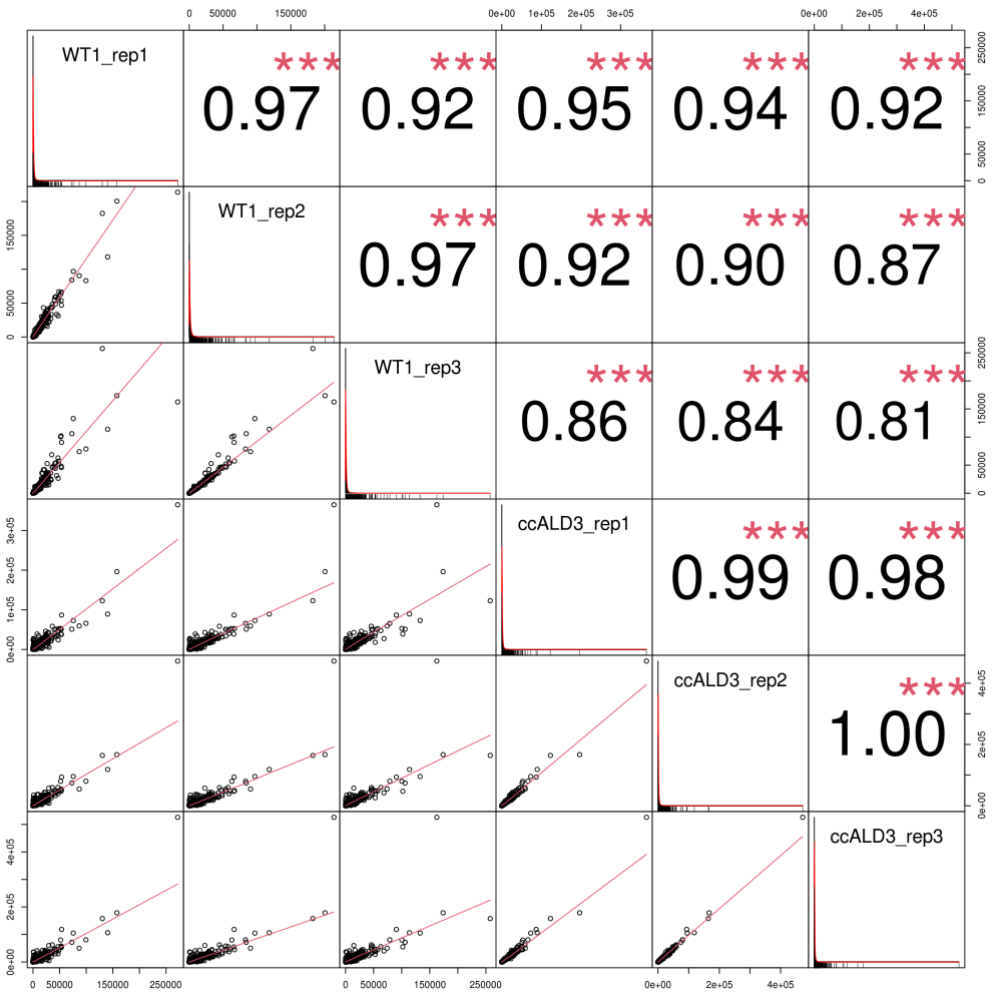

Figura 21. Gráficos de correlación del control de calidad para las muestras de queratinocitos. En esta figura se representan los valores de correlación para las distintas muestras de controles (WT1) y pacientes (ccALD3) comparadas. Las réplicas (rep) que están dentro del mismo grupo tienden a tener coeficientes de correlación de Pearson igual o mayor a 0.96.

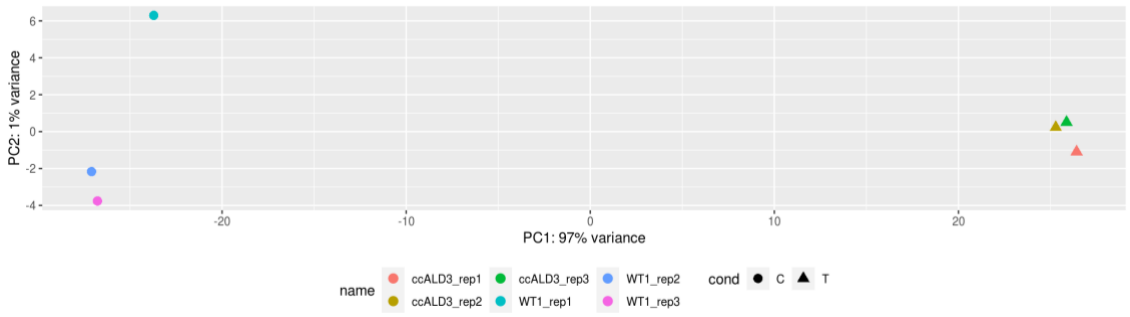

Figura 22. Análisis de componentes principales. DEgenes Hunter lleva a cabo este análisis usando los valores de recuento después de la normalización de rlog del paquete DESeq2. Se observa una clara separación de los grupos control (C, círculos) y pacientes (T, triángulos), apoyada por la componente con el mayor porcentaje de la varianza (PC1, 97%).

5.3.3. Análisis de enriquecimiento funcional

En este apartado, se describirán los resultados obtenidos para el análisis de enriquecimiento funcional obtenidos a partir del análisis de expresión diferencial general de las muestras de queratinocitos (https://drive.google.com/file/d/1UCRULumSau5vg_SfjzFRLBI2JMVB8sb4/view?usp=sharing) y el de los módulos de co-expresión de queratinocitos que contienen genes cuyo patrón de expresión coincide con el de las condiciones analizadas en este experimento (tipo celular, método de reprogramación celular, condición y sexo): módulos 1 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1U2tHevUqZ0weece22Dg_0oETlmx0TDOU/view?usp=sharing) y 2 (disponible para su descarga en el siguiente enlace: https://drive.google.com/file/d/1bZn2KtdmKqPNB5M7OxFCAqR5l6bbffym/view?usp=sharing).

Los cinco genes con un mayor nivel de sobreexpresión (ordenados según la media calculada de valores de logFC positivos por los paquetes de expresión diferencial empleados, parámetro *mean_logFCs* en la tabla *Top positive* en el informe de enriquecimiento funcional), se encuentran los genes *RPS4Y1* (12.676), *PSG4* (12.533), *KDM5D (*12.429*), TXLNGY (*11.865) y *UTY (*11.700). Con respecto a los genes con un mayor nivel de inhibición (*Top negative* en el informe de enriquecimiento funcional), se encuentran *CNTN4* (-11.115), *COL22A1* (-10.840), *C5orf17* (-7.800), *USP6* (-7.642) y *TAFA5* (-7.503).

El gráfico representado en la Figura S1 (Anexo I) muestra los términos significativos funcionales en orden ascendente por p-valor ajustado para las categorías pertenecientes a los términos GO dentro de la subcategoría *Molecular functions*. Las categorías más significativas corresponden con los términos *growth factor binding*, *extracellular matrix structural consistuent*, *cytokine binding*, *integrin binding*, *extracelular matrix binding* y *extracelular matrix structural constituent conferring tensile strangth* con unos p-valores próximos a 0.02.

La red devuelta por el análisis que conecta los términos funcionales (nodos) a través de sus genes asociados (líneas grises, cuyo el grosor representa el número de genes compartidos entre ambos términos funcionales) para la categoría GOMF expone varios módulos, entre los que destacan uno de términos relacionados con la actividad de canales dependientes de voltaje y uno con términos relacionados con receptores transmembrana (Figura S2, Anexo I).

Con respecto a la subcategoría de GO *Biological process*, para estas muestras se han encontrado genes enriquecidos con los módulos funcionales correspondientes a los términos *extracellular matrix organization, extracellular structure organization, reproductive structure developmen, reproductive system developmen, ossification, epithelial cell proliferation, regulation of epithelial cell proliferation* y con el p-valor más significativo (Figura S3, Anexo I). Por su parte, en la red de términos funcionales (Figura S4, Anexo I) se observan tres módulos funcionales correspondientes a categorías relacionadas con procesos de desarrollo del sistema reproductivo (esto

puede deberse a que las células a comparar pertenecen a individuos de diferente sexo), procesos dependientes de angiogénesis y procesos relacionado con la respuesta a BMP.

Por otra parte, del análisis con términos de la subcategoría GO *Cellular component*, se puede observar en la Figura S5 (Anexo I) que las categorías más significativas en cuanto al p-valor corresponden con los términos *collagen-containing extracellular matrix*, cell-substrate junction, *basement membrane, focal adhesión, endoplasmatic reticulum lumen, cell-cell junction, ion channel complex y membrane microdomain*. Se puede ver como la adhesión de la célula tanto al substrato como a otras células se está viendo afectada. Con respecto al análisis de red (Figura S6, Anexo I), se observan dos módulos de términos claramente diferenciados relacionados con componentes de la membrana sináptica y regiones específicas de la membrana celular.

En la Tabla 3 que se expone a continuación se representan los seis primeros términos de cada una de las tres categorías del análisis GO (procesos biológicos, funciones moleculares y componentes celulares) así como los del análisis KEGG, con el objetivo de observar a grandes rasgos los mecanismos moleculares que subyacen a la CALD.

Tabla 3. Resumen de los términos con mayor significancia del análisis. De izquierda a derecha: términos KEGG, términos GO de función molecular (FM), términos GO de procesos biológicos (PB), términos GO de componentes celulares (CC) y reactome.

KEGG	FM	PB	CC	Reactome
Rutas del cáncer	Unión al factor de crecimiento	Organización de la matriz extracelular	Matriz extracelular que contiene colágeno	Organización de la matriz extracelular
Adhesiones focales	Constituyente estructural de la matriz extracelular	Organización de la estructura extracelular	Unión célula-substrato	Formación de fibras elásticas
Moléculas celulares de adhesión (CAMs)	Unión de citocinas	Desarrollo de estructura reproductiva	Membrana basal	Ensamblaje de fibras de colágenos y otras estructuras multimétricas
Ruta de señalización MAPK	Unión de integrina	Desarrollo de sistema reproductivo	Adhesiones focales	Moléculas asociadas con las fibras elásticas
Interacción de la membrana extracelular y el receptor	Unión de matriz extracelular	Osificación	Lumen del retículo endoplasmático	Interacción de la integrina de la superficie celular
Digestión y absorción de proteínas	Constituyente estructural de la matriz extracelular que confiere resistencia a la tracción	Proliferación de células epiteliales	Unión célula-célula	Señalización de interferón alfa

Respecto a los módulos de co-expresión, podemos ver como en el módulo 1 (Figura S7) los términos GO de componentes celulares representados tienen que ver con la adhesión celular y la vía endocitica (*cell-sbstrate junction, focal adhesion, vacuolar membrane, lysosomal membrane, lytic vacuole membrane, early endosome*…)

6. Discusión

La adrenoleucodistrofia ligada al cromosoma X (ALD-X) es una enfermedad bien definida genéticamente, pero cuyo fenotipo clínico posee una gran heterogeneidad hasta ahora inexplicable [48]. Las mutaciones en *ABCD1* reflejan el fenotipo observado en los pacientes con adrenomieloneuropatía (AMN). Sin embargo, estas mutaciones por si solas no explican el fenotipo de correspondiente a la adrenoleucodistrofia cerebral (CALD) [2]. Como se comentó en el Capítulo 2, diferentes mecanismos desconocidos parecen estar implicados en la enfermedad, entre ellos el inicio de la desmielinización que sufren los pacientes y precede a la axonopatía, responsable de los síntomas clínicos de la enfermedad [2], [3], [22]. Existen varias hipótesis de por qué puede estar sucediendo este acontecimiento. Sin embargo, en este trabajo y gracias al preciso análisis de datos NGS utilizando la herramienta DEgenes Hunter [12] se plantea un nuevo mecanismo que podría estar siendo el causante de la desmielinización inicial.

Las células de pacientes con CALD en las que se ha basado el estudio de DEG que se expone en este proyecto demostraron contener acúmulos de gotas lipídicas en mayor medida que las células carentes de la enfermedad [11]. Esto es así porque la CALD, al igual que la AMN, se caracteriza por el acúmulo de VLCFA debido a su incapacidad para degradarlos en el peroxisoma, ya que el trasportador (ABCD1) que se encarga de su incorporación a la matriz extracelular se encuentra mutado. Por ello, resultaba interesante consultar que genes estaban siendo expresados de forma diferencial bajo esas condiciones.

Gracias al análisis de los genes diferencialmente expresados entre células de pacientes (CALD) y células de individuos sanos (WT) se determinó el pseudogén *CYP4F29P* como el más sobreexpresado, seguido de los genes *CTSF* y tres genes pertenecientes a la superfamilia de proteínas de dedos de zinc (ZNF208, ZNF560 y ZNF728). El psudogén *CYP4F29P* pertenece a la superfamilia del citocromo P450 y no se han encontrado referencias que asocien su sobreexpresión con esta patología. Con respecto al gen *CTSF* se conoce que da lugar a una proteína implicada en la degradación intracelular de proteínas y que se ha asociado a diferentes patologías neuronales, así como a epilepsia visual [49]. Con respecto a los tres restantes genes sobreexpresados en este estudio, se sabe que sus productos están involucrados en múltiples procesos moleculares, incluyendo el desarrollo y diferenciación celular. De los genes con un mayor silenciamiento cabe destacar *CNTN4*, perteneciente a la superfamília de inmunoglobulinas derivadas del cerebro y que actúan como moléculas de adhesión entre los axones neuronales. El hecho de aparecer con un nivel de inhibición tan grande podría estar directamente implicado en los fallos de transmisión de señales en el cerebro de estos pacientes, dando lugar al fenotipo observado. Por lo tanto, podría ser un gen candidato interesante a analizar y que pudiera explicar lo que sucede en pacientes con este tipo de adrenoleucodistrofia. En cuanto al posterior análisis funcional se puede ver como existe una sobrerrepresentación de los siguientes términos de rutas de Reactome: *glycoprotein hormones, peptide hormone biosynthesis, interferon alpha / beta signaling e LGI-ADAM interaction*. Los dos

primeros términos hacen referencia a la síntesis de hormonas, se conoce que ciertos tipos celulares con función endocrina que muestran una sobreexpresión de *ABCD1* también producen proopiomelanocortina (POMC), el precursor de ACTH y de otras hormonas peptídicas [50], por lo que pudiera ser que la ausencia de este gen *ABCD1* active otros genes que desencadenen la activación de síntesis de estas hormonas. La señalización del interferón alfa/beta puede deberse al proceso de respuesta inflamatoria secundaria típico de CALD [11]. Por último, el siguiente término más representado es *LGI-ADAM interaction*, se ha demostrado que estas interacciones juegan un papel crucial en el desarrollo y la función del sistema nervioso de los vertebrados, principalmente mediando la transmisión sináptica y la mielinización [51]. Como se ha comentado en el Capítulo 2, los pacientes con CALD sufren una rápida y progresiva desmielinización, la sobrerrepresentación del termino *LGI-ADAM interaction* concuerda con este hecho.

El análisis de los módulos de co-expresión para la comparación entre controles y pacientes presentó una alta correlación con el módulo 37, con resultados significativos para términos de la categoría funcional GOCC, tales como *intersticial matrix*, *nuclear outer membrane*, *growth cone*. Se conoce que las alteraciones en estas funciones moleculares dan lugar al fenotipo observado en pacientes ALD-X [52], por lo que un análisis exhaustivo de los genes que pertenecen a este módulo podría arrojar información sobre las causas potenciales que den lugar a la enfermedad. Debido a la extensión de este trabajo y su duración temporal, esta tarea será llevada a cabo en un trabajo futuro. Como se ha comentado en el Capítulo 4, el análisis se realizó con muestras procedentes de diferentes tipos celulares. Esta heterogeneidad en la procedencia de las células podría ser un factor de sesgo en el análisis, ya que muchos de los cambios observados vendrían dados por el tipo celular y no por la enfermedad en sí. Con el fin de solucionar este problema, se realizó un análisis de nuevo con seis de las 18 muestras, ambas procedentes de queratinocitos, siendo tres de las muestras réplicas de un paciente con la enfermedad y tres pertenecientes a un control. Lo más destacable de este segundo análisis es la visible sobrerrepresentación de términos relacionados con la adhesión celular y con los cambios estructurales. Si volvemos a fijarnos en el primer análisis (el que abarca todas las muestras) vemos que uno de los términos en los cuales se ha enriquecidoun mayor número de genes con altos niveles de expresión en la categoría de componentes celulares de los términos GO es *cell-cell junction.* Sin embargo, estos resultados junto con el análisis de los módulos 1 y 2 de co-expresión (cuyos genes presentaban un perfil de correlación muy similar al de los genes del vector de muestras empleado para este análisis) no arrojó suficiente información como para determinar si los cambios de expresión eran propios del tipo celular empleado. Necesitaríamos realizar de nuevo el experimento con un mayor número de muestras, donde se pudieran comparar en distintos individuos el perfil de expresión de estos genes para que pudiera ser significativo. En trabajo futuro se buscarán más análisis de esta enfermedad que tengan datos de RNA-seq públicamente disponibles y con un diseño experimental más sólido para poder realizar análisis como el planteado en este TFM.

Unos de los mecanismos desconocidos en esta enfermedad es el inicio de la desmielinización [2]. Las células gliales (células de Schwann y oligodendrocitos) son las encargadas de proteger los axones neuronales recubriéndolos con su propio

cuerpo celular o sus prolongaciones. El desarrollo de las células de Schwann depende fundamentalmente de la regulación de la función de adhesión tanto a los componentes de la matriz extracelular de la lámina basal [53] como a los propios axones y otras a células de Schwann [54].

El ácido lisofosfolipídico (LPA del inglés, *lysophospholipidic acid*) (2-hydroxy-3-phosphonooxypropyl), es un es un derivado de fosfolípidos que puede actuar como molécula de señalización [55] (Figura 23). Se ha demostrado que es un regulador de las reorganizaciones citoesqueléticas de actina en varios tipos celulares [50, 52, 53]. Con la identificación del gen receptor LPA (*pA1*) [58], y la posterior demostración de su expresión por células de Schwann [59] y oligodendrocitos [60], se propusieron funciones de señalización de LPA en la regulación de la adhesión y morfología de las células mielinizantes [61]. Se ha demostrado que los mecanismos de señalización citoesquelética inducidos por LPA posiblemente podrían contribuir no solo al control de la mielinización [62] sino también a la supervivencia de las propias células de Schwann [61].

Existen varias especies de LPA tanto de ácidos grasos saturados (16:0-LPA y 18:0-LPA) como de ácido graso mono y poliinsaturado (18:1-LPA, 18:2-LPA y 20:4-LPA). En un estudio reciente, en el que se comparaba el perfil lipídico celular de una mujer con X-ALD con el de un individuo sano, se comprobó que una de las nueve moléculas lipídicas que se veía disminuida era 16:0-LPA [63]. Esto encaja con el hecho de existen estudios en los que se demostró que el alargamiento de los VLCFA saturados y monoinsaturados se ve incrementado en los fibroblastos de pacientes con defectos de β-oxidación peroxisómica, incluido X-ALD [64]. Por lo que la cantidad 16:0-LPA podría estar siendo disminuida como consecuencia de la actividad de las elongasas sobre la cadena lipídica de estas moléculas.

Analizando los términos más representativos obtenidos tras el análisis funcional (unión célula-célula, adhesiones focales, unión célula-substrato, organización de la matriz extracelular, entre otros) se puede sugerir que el principal cambio en los mecanismos moleculares que se está dando a causa de la enfermedad son cambios de adhesión y estructurales de la célula, lo que encaja con el hecho de que la función de LPA es, entre otras, regular la adhesión celular en las células gliales [62].

La pérdida de adhesión de las células gliales puede explicar el inicio de la desmielinización ya que supondría la desestructuración de la mielina que envuelve los axones. Además, este estudio se realizó con células de la barrera hematoencefálica, que veían alterados sus mecanismos de adhesión. Esto sugiere que, al igual que afecta a las células gliales, también afecta de esta misma forma a las células endoteliales de la barrera hematoencefálica, alterando sus adhesiones y provocando un aumento en la permeabilidad de esta y, por ende, un aumento de la inflamación cerebral (ambos sucesos observados en pacientes con CALD).

Por ello, en este trabajo de fin de máster se plantea la hipótesis de que la disminución de 16:0-LPA juega un rol central en el inicio de la desmielinización que se da en X-ALD. Además, podría estar promoviendo el aumento en la permeabilidad de la barrera hematoencefálica con el consecuente inicio del proceso de inflamación

cerebral típico de CALD. Para confirmar estos resultados, se debería proceder con la validación experimental de los mismos, lo cual se planteará como trabajo futuro para seguir con esta línea de investigación.

7. Conclusiones

1. Se han analizado datos pertenecientes a muestras de pacientes con CALD con el fin de analizar los factores genéticos implicados en su desarrollo. Existen pocos conjuntos de datos RNA-seq con información sobre esta enfermedad para poder obtener conclusiones sólidas, y en el empleado para el presente trabajo, el diseño experimental debería incluir un mayor número de individuos, así como una heterogeneidad en los factores analizados.

2. El procedimiento de instalación de la herramienta DEgenes Hunter así como el empleo del flujo de trabajo para el análisis de datos RNA-seq ha demostrado ser útil para la obtención de datos biológicos en el tiempo establecido para la realización de este trabajo.

3. La obtención de genes expresados diferencialmente en todos los tipos celulares empleados en el estudio del que parte este trabajo y el tipo específico (queratinocitos) no arrojan suficiente información como para concluir que la expresión de los genes se debe a un tipo celular específico.

4. Se ha caracterizado el gen *CNTN4* como altamente silenciado en la comparación para los pacientes CADL a partir de los datos analizados, y el módulo de co-expresión 37 presenta genes cuyo perfil de correlación podría explicar con más detalles los mecanismos moleculares implicados en el desarrollo de la enfermedad.

5. La sobrerrepresentación de términos relacionados con la adhesión observados en el análisis, así como las evidencias que existen de que el 16:0-LPA se ve reducido en pacientes con X-ALD, sugiere que está molécula de señalización que regula los procesos de adhesión, está jugando un papel central en el inicio de la desmielinización que se da en X-ALD.

El objetivo general de este TFM consistía en la determinación de los factores genéticos y moleculares implicados en el desarrollo de la CALD a través del análisis de expresión, co-expresión y funcional de datos de secuenciación, empleando la herramienta DEgenes Hunter [6]. Podemos afirmar que este objetivo ha sido cumplido de forma exitosa gracias a la generación de una nueva hipótesis sobre los mecanismos moleculares de la enfermedad.

8. Glosario

X-ALD: acrónimo de adrenoleucodistrofia ligada al cormosoma X. Enfermedad rara peroxisomal que conduce a una desmielinización cerebral, disfunción axonal de la médula espinal, insuficiencia suprarrenal y, en algunos casos, insuficiencia testicular. Es debida a mutaciones en el gen *ABCD1.*

AMN: acrónimo de adrenomieloneuropatía. Es el fenotipo central de la adrenoleucodistrofia ligada al cromosoma X.

CALD: por su acrónimo en inglés *Cerebral Adrenoleucodistrophy*. Es el fenotipo más severo de la X-ALD.

VLCFA: por su acrónimo en inglés *Very Long Chain Fatty Acid*. Se denominan así los ácidos grasos de cadena muy larga (los que tienen 22 carbonos o más).

LPA: por su acrónimo en inglés *Lysophospholipidic Acid*. Derivado de fosfolípidos que puede actuar como molécula de señalización.

NGS: por su acrónimo en inglés *Next-Generation Sequencing.* Se refiere a la tecnología de secuenciación de ADN a gran escala que permite consultar el genoma completo, los exones dentro de todos los genes conocidos o solo exones de genes seleccionados (panel objetivo)

DEG: por su acrónimo en inglés *Differentially Expressed Genes.* Hace referencia los genes que se encuentran expresados de forma diferente entre dos muestras distintas.

GO: por su acrónimo del inglés *Gene Ontology*. Se refiere a la ontología genética que describe productos génicos con tres categorías independientes: proceso biológico, componente celular y función molecular

WT: por su acrónimo del inglés *Wild Type.* Significa 'natural', 'salvaje', 'tipo silvestre', hace referencia al fenotipo sin tratamiento o enfermedad.

Heatmap: su traducción es “mapa de calor” es una técnica de visualización de datos que expone la magnitud de un fenómeno como color en dos dimensiones. La variación de color proporciona pistas visuales obvias sobre cómo el fenómeno se agrupa o varía en el espacio.

Dendrograma: Diagrama en forma de árbol que muestra la relación jerárquica existente entre objetos.

8. Referencias

[1] J. Sepúlveda, F. Gutiérrez, M. Moreno, and A. Hernández, “Peroxisomal Proliferation Induced by Treatment with Clofibrate in a Patient with a Peroxisomal Disease,” *Cell Biochem. Biophys.*, 2000.

[2] H. W. Moser, “X-linked adrenoleukodystrophy,” in *Treatment of Pediatric Neurologic Disorders*, 2005.

[3] S. Kemp, J. Berger, and P. Aubourg, “X-linked adrenoleukodystrophy: Clinical, metabolic, genetic and pathophysiological aspects,” *Biochimica et Biophysica Acta - Molecular Basis of Disease*. 2012.

[4] F. Kallabi *et al.*, “Molecular characterization of X-linked adrenoleukodystrophy in a tunisian family: Identification of a novel missense mutation in the ABCD1 gene,” *Neurodegener. Dis.*, 2013.

[5] C. Wiesinger, F. S. Eichler, and J. Berger, “The genetic landscape of X-linked adrenoleukodystrophy: Inheritance, mutations, modifier genes, and diagnosis,” *Application of Clinical Genetics*. 2015.

[6] K. Lohmann and C. Klein, “Next Generation Sequencing and the Future of Genetic Diagnosis,” *Neurotherapeutics*, vol. 11, no. 4, pp. 699–707, 2014.

[7] M. Beigh, “Next-Generation Sequencing: The Translational Medicine Approach from ‘Bench to Bedside to Population,’” *Medicines*, vol. 3, no. 2, p. 14, 2016.

[8] S. A. Byron, K. R. Van Keuren-Jensen, D. M. Engelthaler, J. D. Carpten, and D. W. Craig, “Translating RNA sequencing into clinical diagnostics: Opportunities and challenges,” *Nature Reviews Genetics*, vol. 17, no. 5. Nature Publishing Group, pp. 257–271, May-2016.

[9] K. O. Mutz, A. Heilkenbrinker, M. Lönne, J. G. Walter, and F. Stahl, “Transcriptome analysis using next-generation sequencing,” *Current Opinion in Biotechnology*. 2013.

[10] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, “Gene co-expression analysis for functional classification and gene-disease predictions,” *Brief. Bioinform.*, 2018.

[11] C. A. A. Lee, H. S. Seo, A. G. Armien, F. S. Bates, J. Tolar, and S. M. Azarin, “Modeling and rescue of defective blood-brain barrier function of induced brain microvascular endothelial cells from childhood cerebral adrenoleukodystrophy patients,” *Fluids Barriers CNS*, 2018.

[12] I. González Gayte, R. Bautista Moreno, P. Seoane Zonjic, and M. G. Claros, “DEgenes Hunter - A Flexible R Pipeline for Automated RNA-seq Studies in Organisms without Reference Genome,” *Genomics Comput. Biol.*, vol. 3, no. 3, p. e31, 2017.

[13] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.

[14] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, vol. 15, no. 550, 2014.

[15] S. Tarazona *et al.*, “Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package,” *Nucleic Acids Res.*, 2015.

[16] M. E. Ritchie *et al.*, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, 2015.

[17] P. Langfelder and S. Horvath, “WGCNA: An R package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 559, 2008.

[18] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, “ClusterProfiler: An R package for comparing biological themes among gene clusters,” *Omi. A J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, 2012.
[19] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*. 2000.
[20] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 27, no. 1. pp. 29–34, 1999.
[21] D. Croft *et al.*, “The Reactome pathway knowledgebase,” *Nucleic Acids Res.*, 2014.
[22] J. Berger, S. Forss-Petter, and F. S. Eichler, “Pathophysiology of X-linked adrenoleukodystrophy,” *Biochimie*. 2014.
[23] J. Mosser *et al.*, “Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters,” *Nature*, vol. 361, no. 6414, pp. 726–730, 1993.
[24] C. W. T. Roermund *et al.*, “The human peroxisomal ABC half transporter ALDP functions as a homodimer and accepts acyl–CoA esters,” *FASEB J.*, vol. 22, no. 12, pp. 4201–4208, Dec. 2008.
[25] C. Wiesinger, M. Kunze, G. Regelsberger, S. Forss-Petter, and J. Berger, “Impaired very long-chain acyl-CoA β-oxidation in human X-linked adrenoleukodystrophy fibroblasts is a direct consequence of ABCD1 transporter dysfunction,” *J. Biol. Chem.*, vol. 288, no. 26, pp. 19269–19279, Jun. 2013.
[26] F. Fouquet *et al.*, “Expression of the adrenoleukodystrophy protein in the human and mouse central nervous system,” *Neurobiol. Dis.*, vol. 3, no. 4, pp. 271–285, 1997.
[27] T. Matsukawa *et al.*, “Identification of novel SNPs of ABCD1, ABCD2, ABCD3, and ABCD4 genes in patients with X-linked adrenoleukodystrophy (ALD) based on comprehensive resequencing and association studies with ALD phenotypes,” *Neurogenetics*, vol. 12, no. 1, pp. 41–50, Feb. 2011.
[28] N. Troffer-Charlier, N. Doerflinger, E. Metzger, F. Fouquet, J. L. Mandel, and P. Aubourg, “Mirror expression of adrenoleukodystrophy and adrenoleukodystrophy related genes in mouse tissues and human cell lines,” *Eur. J. Cell Biol.*, vol. 75, no. 3, pp. 254–264, Mar. 1998.
[29] A. Netik, S. Forss-Petter, A. Holzinger, B. Molzer, G. Unterrainer, and J. Berger, “Adrenoleukodystrophy-related protein can compensate functionally for adrenoleukodystrophy protein deficiency (X-ALD): Implications for therapy,” *Hum. Mol. Genet.*, vol. 8, no. 5, pp. 907–913, 1999.
[30] E. M. Maier *et al.*, “X-linked adrenoleukodystrophy phenotype is independent of ABCD2 genotype,” *Biochem. Biophys. Res. Commun.*, vol. 377, no. 1, pp. 176–180, Dec. 2008.
[31] J. Berger, B. Molzer, I. Faé, and H. Bernheimer, “X-linked adrenoleukodystrophy (ALD): A novel mutation of the ALD gene in 6 members of a family presenting with 5 different phenotypes,” *Biochem. Biophys. Res. Commun.*, vol. 205, no. 3, pp. 1638–1643, 1994.
[32] K. D. Smith *et al.*, “X-linked adrenoleukodystrophy: Genes, mutations, and phenotypes,” *Neurochem. Res.*, vol. 24, no. 4, pp. 521–535, 1999.
[33] G. N. O’Neill, M. Aoki, and R. H. Brown, “ABCD1 translation-initiator mutation demonstrates genotype-phenotype correlation for AMN,” *Neurology*, 2001.
[34] C. M. Kassmann and K. A. Nave, “Oligodendroglial impact on axonal function and survival - A hypothesis,” *Current Opinion in Neurology*, vol. 21, no. 3. Curr Opin Neurol, pp. 235–241, Jun-2008.
[35] J. Jang *et al.*, “Induced pluripotent stem cell models from X-linked adrenoleukodystrophy patients,” *Ann. Neurol.*, vol. 70, no. 3, pp. 402–409, Sep. 2011.
[36] Z. Liu, L. Zhu, R. Roberts, and W. Tong, “Toward Clinical Implementation of

Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We?," *Trends in Genetics*. 2019.
[37] P. Seoane *et al.*, "AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (Vicia faba)," *Curr. Bioinform.*, vol. 11, 2016.
[38] J. Falgueras, A. J. Lara, N. Fernández-Pozo, F. R. Cantón, G. Pérez-Trabado, and M. G. Claros, "SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read.," *BMC Bioinformatics*, vol. 11, no. 38, pp. 1–12, 2010.
[39] S. Andrews, F. Krueger, A. Seconds-Pichon, F. Biggins, and S. Wingett, "FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics," *Babraham Institute*, 2015. .
[40] A. Dobin *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.
[41] K. Okonechnikov, A. Conesa, and F. García-Alcalde, "Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data," *Bioinformatics*, vol. 32, no. 2, pp. 292–294, 2015.
[42] J. Costa-Silva, D. Domingues, and F. M. Lopes, "RNA-Seq differential expression analysis: An extended review and a software tool," *PLoS One*, vol. 12, no. 12, p. e0190152, 2017.
[43] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 2009.
[44] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: A matter of depth," *Genome Res.*, 2011.
[45] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*. 1999.
[46] A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.*, 2018.
[47] S. C. Nallar and D. V. Kalvakolanu, "Interferons, signal transduction pathways, and the central nervous system," *Journal of Interferon and Cytokine Research*. 2014.
[48] M. Engelen, S. Kemp, and B. T. Poll-The, "X-linked adrenoleukodystrophy: Pathogenesis and treatment," *Current Neurology and Neuroscience Reports*, vol. 14, no. 10. Current Medicine Group LLC 1, pp. 1–8, 2014.
[49] J. Van Der Zee *et al.*, "Mutated CTSF in adult-onset neuronal ceroid lipofuscinosis and FTD," *Neurol. Genet.*, 2016.
[50] R. Höftberger *et al.*, "Peroxisomal localization of the proopiomelanocortin-derived peptides β-lipotropin and β-endorphin," *Endocrinology*, 2010.
[51] L. Kegel, E. Aunin, D. Meijer, and J. R. Bermingham, "LGI Proteins in the Nervous System," *ASN Neuro*, vol. 5, no. 3, p. AN20120095, Jun. 2013.
[52] P. L. Musolino *et al.*, "Brain endothelial dysfunction in cerebral adrenoleukodystrophy," *Brain*, 2015.
[53] M. B. Bunge, R. P. Bunge, N. Kleitman, and A. C. Dean, "Role of peripheral nerve extracellular matrix in schwann cell function and in neurite regeneration," *Dev. Neurosci.*, vol. 11, no. 4–5, pp. 348–360, 1989.
[54] P. C. Letourneau, T. A. Shattuck, F. K. Roche, M. Takeichi, and V. Lemmon, "Nerve growth cone migration onto Schwann cells involves the calcium-dependent adhesion molecule, N-cadherin," *Dev. Biol.*, vol. 138, no. 2, pp. 430–442, 1990.
[55] E. J. van Corven, A. Groenink, K. Jalink, T. Eichholtz, and W. H. Moolenaar, "Lysophosphatidate-induced cell proliferation: Identification and dissection of signaling pathways mediated by G proteins," *Cell*, vol. 59, no. 1, pp. 45–54, Oct. 1989.
[56] W. H. Moolenaar, "Lysophosphatidic acid signalling," *Curr. Opin. Cell Biol.*, vol. 7, no. 2, pp. 203–210, 1995.

[57] A. Gohla, R. Harhammer, and G. Schultz, “The G-protein G13 but not G12 mediates signaling from lysophosphatidic acid receptor via epidermal growth factor receptor to Rho,” *J. Biol. Chem.*, vol. 273, no. 8, pp. 4653–4659, Feb. 1998.

[58] J. H. Hecht, J. A. Weiner, S. R. Post, and J. Chun, “Ventricular zone gene-1 (vzg-1) encodes a lysophosphatidic acid receptor expressed in neurogenic regions of the developing cerebral cortex,” *J. Cell Biol.*, vol. 135, no. 4, pp. 1071–1083, Nov. 1996.

[59] J. A. Weiner and J. Chun, “Schwann cell survival mediated by the signaling phospholipid lysophosphatidic acid,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 9, pp. 5233–5238, Apr. 1999.

[60] J. A. Weiner, J. H. Hecht, and J. Chun, “Lysophosphatidic acid receptor gene vzg-1/lp(A)1/edg-2 is expressed by mature oligodendrocytes during myelination in the postnatal murine brain,” *J. Comp. Neurol.*, vol. 398, no. 4, pp. 587–598, Sep. 1998.

[61] J. A. Weiner, N. Fukushima, J. J. A. Contos, S. S. Scherer, and J. Chun, “Regulation of Schwann cell morphology and adhesion by receptor-mediated lysophosphatidic acid signaling,” *J. Neurosci.*, vol. 21, no. 18, pp. 7069–7078, Sep. 2001.

[62] C. Fernandez-Valle, L. Gwynn, P. M. Wood, S. Carbonetto, and M. B. Bunge, “Anti-β1 integrin antibody inhibits schwann cell meylination,” *J. Neurobiol.*, vol. 25, no. 10, pp. 1207–1226, 1994.

[63] I. C. Huffnagel *et al.*, “Disease progression in women with X-linked adrenoleukodystrophy is slow,” *Orphanet J. Rare Dis.*, vol. 14, no. 1, p. 30, Dec. 2019.

[64] S. Kemp *et al.*, “Elongation of very long-chain fatty acids is enhanced in X-linked adrenoleukodystrophy,” *Mol. Genet. Metab.*, vol. 84, no. 2, pp. 144–151, Feb. 2005.

9. Anexo I: material suplementario

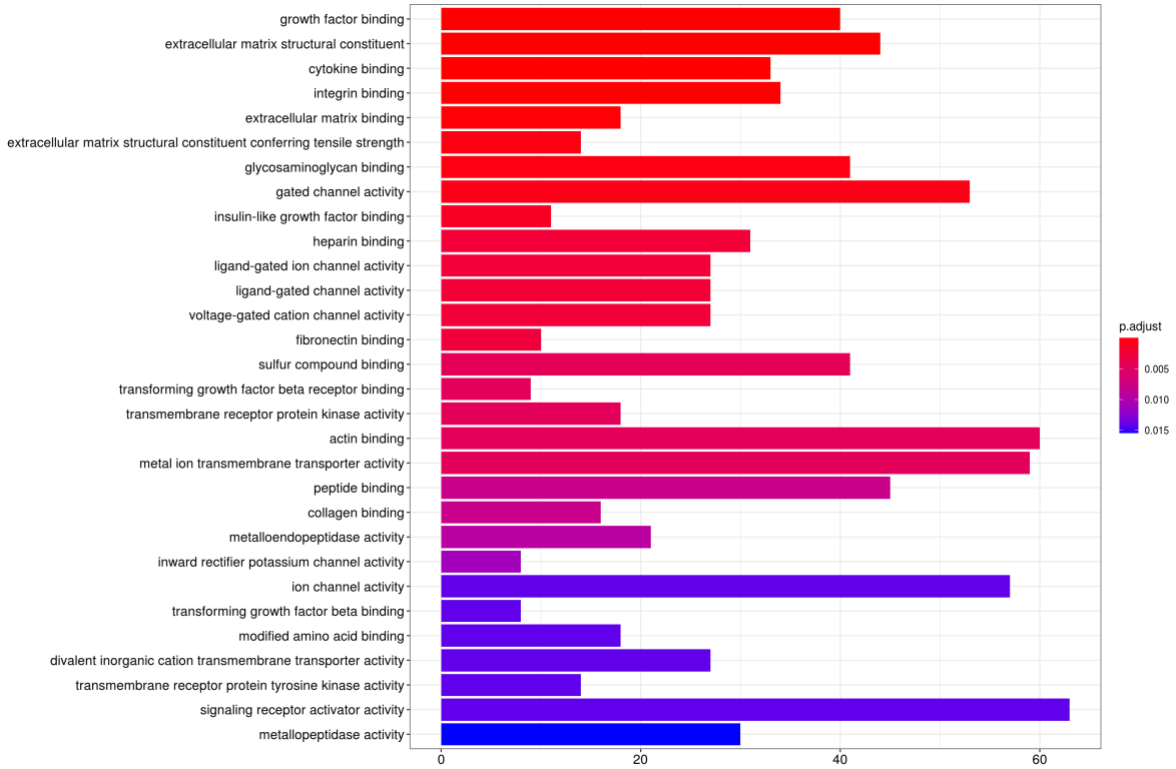

Figura S1. Gráfico ORA correspondiente a los datos GO (subcategoría *Molecular functions*). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

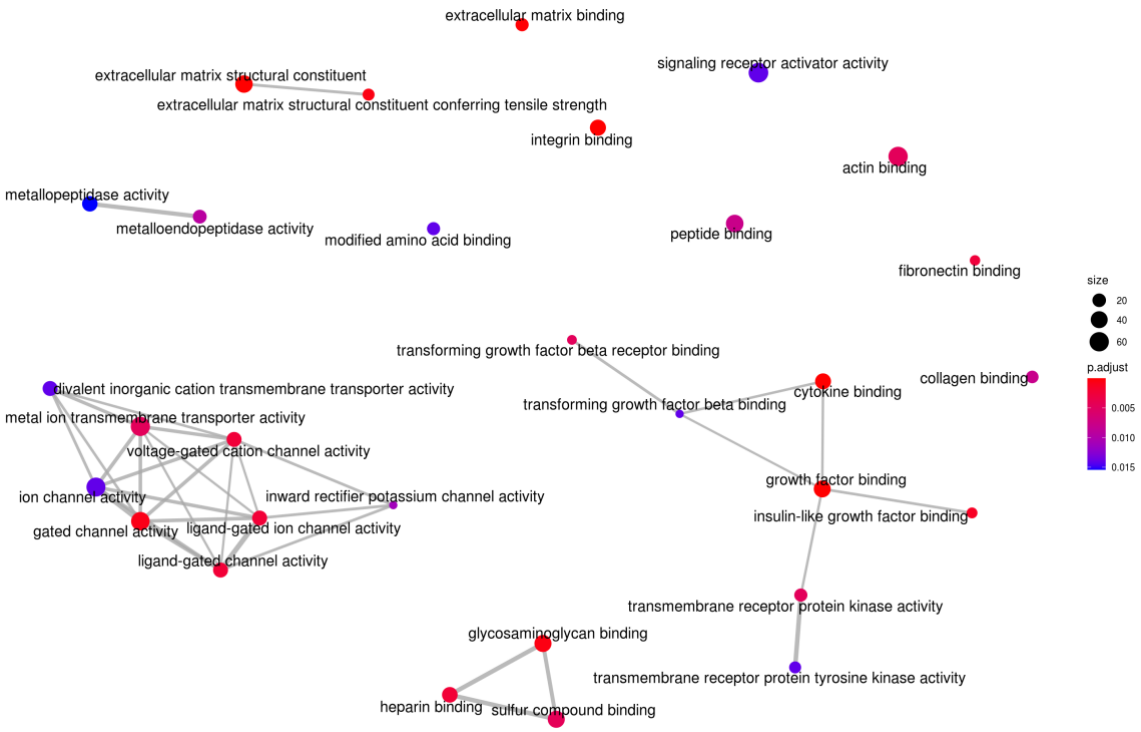

Figura S2. Mapa gráfico de redes de enriquecimiento funcional para GO *Molecular functions*. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional.

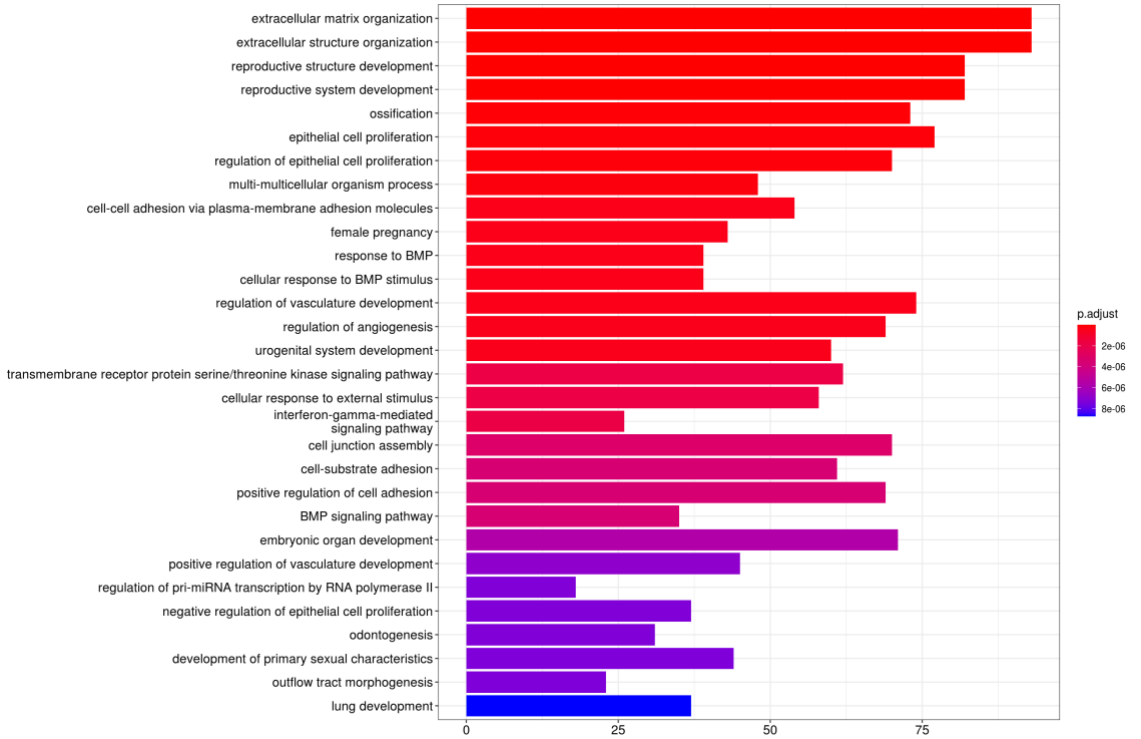

Figura S3. Gráfico ORA correspondiente a los datos GO (subcategoría *Biological process*). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para

cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

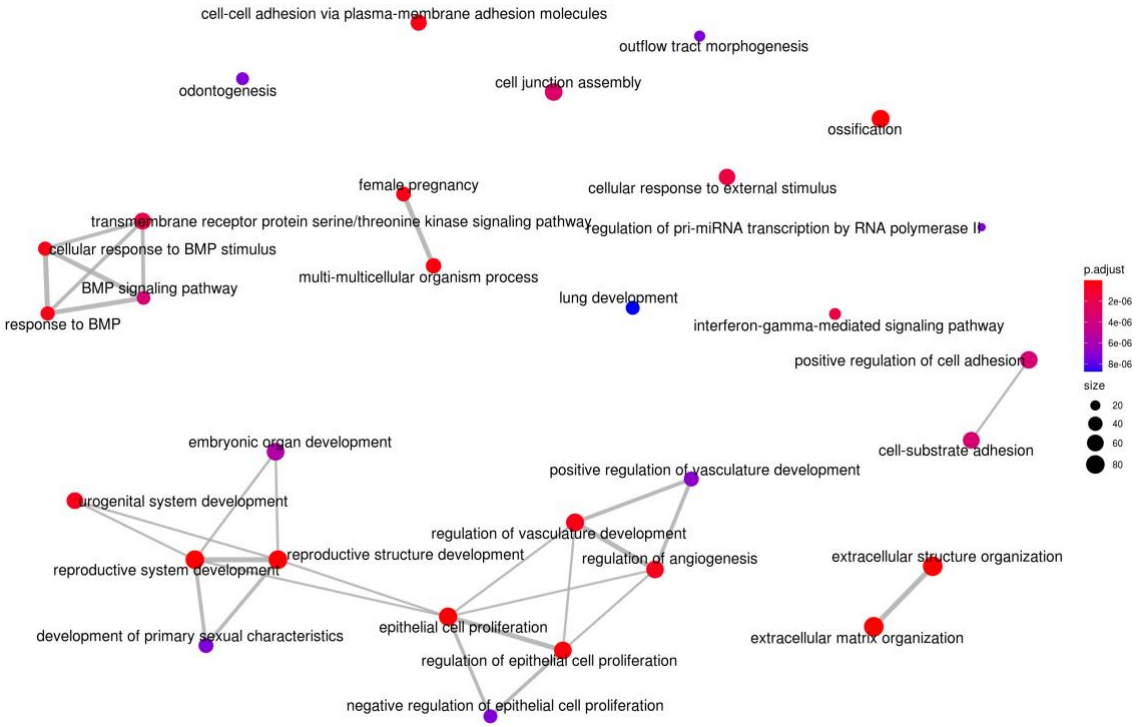

Figura S4. Mapa gráfico de redes de enriquecimiento funcional para GO *Biological process*. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional.

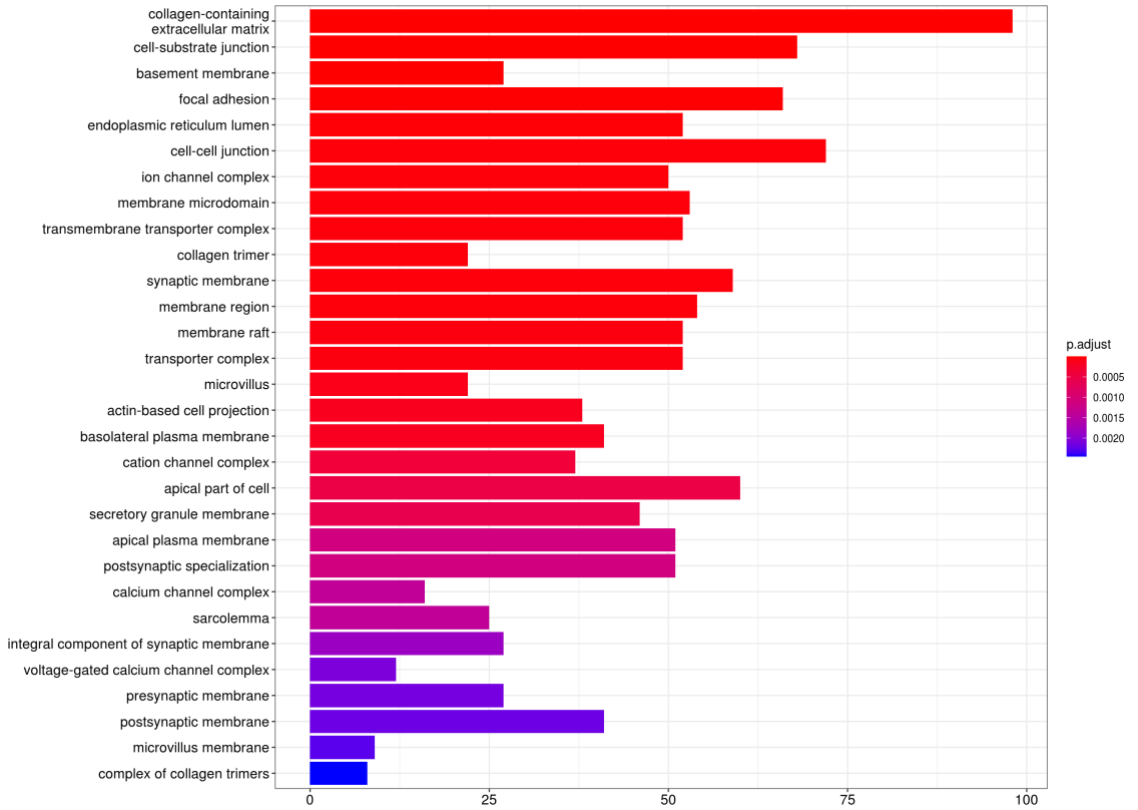

Figura S5. Gráfico ORA correspondiente a los datos GO (subcategoría *Cellular component*). El color de la escalar representa el p-valor ajustado asociado, siendo rojo más significativo y azul menos para cada una de las categorías funcionales. El eje X representa la proporción de genes conocidos para un término funcional dado.

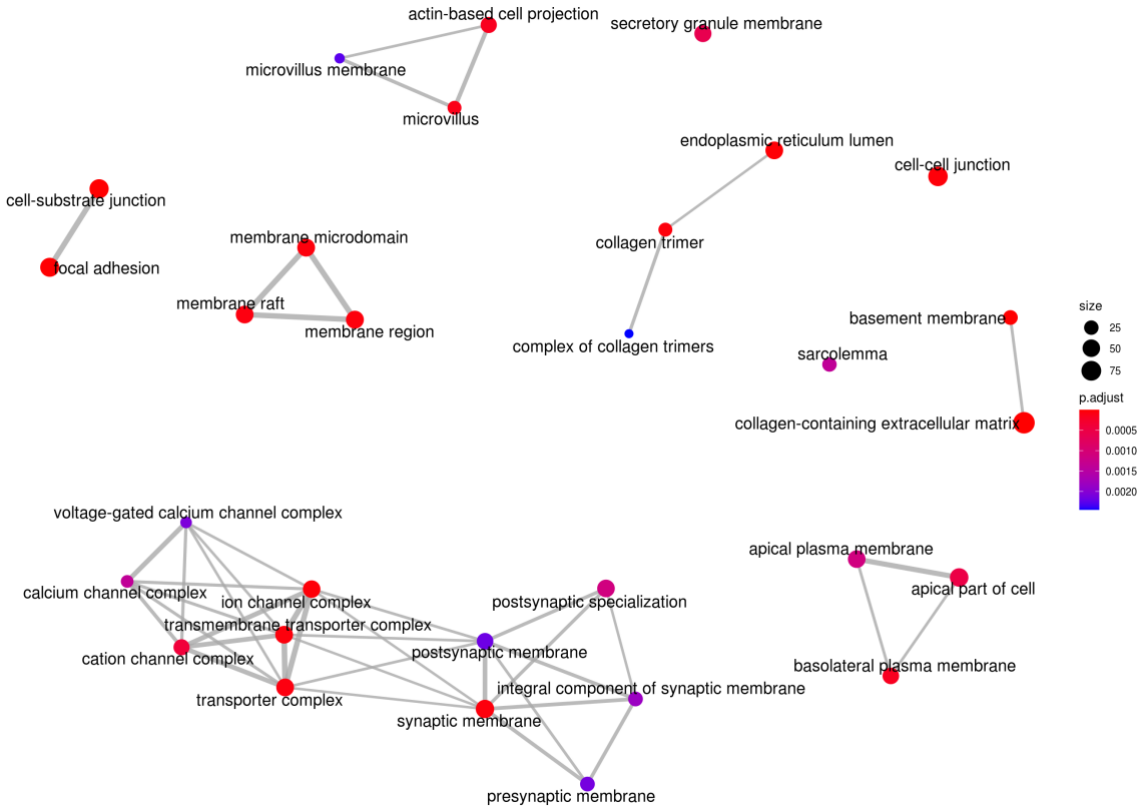

Figura S6. Mapa gráfico de redes de enriquecimiento funcional para GO *Cellular component*. El color de las barras representa el p-valor asociado. El grosor de las líneas grises equivale a la cantidad de genes que comparten dos términos. El diámetro de los puntos hace referencia al número de genes que hay asociados a esa categoría funcional.