Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Grado en Biotecnología

ETSIAMN - *Universitat Politècnica de València*

Trabajo Fin de Grado – Curso 2020/2021

ANÁLISIS DE DATOS ÓMICOS PARA DETERMINAR DIFERENCIAS ENTRE NACIMIENTOS PREMATUROS Y A TÉRMINO A PARTIR DE EXOSOMAS DE LECHE MATERNA

Valencia, septiembre 2021

AUTOR

Antonio Porlán Miñarro

COTUTORAS

Sonia Tarazona Campos

María Teresa Rubio Martínez-Abarca

TUTOR UPV

José Javier Forment Millet

Título – Análisis de datos ómicos para determinar diferencias entre nacimientos prematuros y a término a partir de exosomas de leche materna

Autor – D. Antonio Porlán Miñarro

Cotutoras – Dña. Sonia Tarazona Campos y Dña. María Teresa Martínez-Abarca

Tutor UPV – D. José Javier Forment Millet

Localidad y fecha – Valencia, septiembre de 2021

Resumen

El parto prematuro está asociado a resultados adversos en el desarrollo del neonato. En este aspecto, la leche materna y su composición juegan un papel importante en el correcto desarrollo del neonato, ya que es un alimento que está presente en sus primeros meses de vida. La leche materna tiene una composición variable, en la que aparte del perfil de nutrientes podemos analizar otro tipo de moléculas que nos ayuden a entender las diferencias de dicha composición en nacimientos prematuros.

En concreto, en este trabajo se realiza un estudio transcriptómico y lipidómico de exosomas de 10 muestras de leche materna, 5 de madres de bebés prematuros y 5 de madres de bebés nacidos a término. El análisis transcriptómico se basa en el estudio de la expresión de microRNAs (miRNAs), que son una clase de RNAs pequeños, endógenos y no codificantes que regulan la expresión de los genes humanos, mientras que el análisis lipidómico se basa en el estudio de los lípidos. El objetivo de este estudio es identificar miRNAs y lípidos con cambios significativos entre ambos estados y estudiar los procesos biológicos en los que están implicados. Para ello, se realizó el procesamiento y la normalización de los datos para proceder con el análisis estadístico.

Como resultados de estos análisis, se obtuvieron 30 lípidos con concentraciones alteradas entre ambos grupos. A partir de ellos, se determinó que los niveles de triacilglicéridos (TGs) y diacilglicéridos (DGs) son más altos en las muestras de madres de bebés prematuros, mientras que los niveles de esfingomielinas (SM) y ceramidas (Cer) son más altos en las muestras de las madres con bebés a término y están relacionados con efectos pro-apoptóticos. En transcriptómica, se obtuvieron 2358 genes diana que se encontraban enriquecidos en los 56 miRNAs diferencialmente expresados entre los dos grupos comparados. A su vez, estos genes diana estaban relacionados con diferentes procesos biológicos como la respuesta inmune, la ruta de señalización de la prostaglandina o la ruta de la señalización de la apoptosis, las rutas de señalización Jak/STAT y PI3K, la subfamilia de quimiocinas CXC, la ruta de señalización del IFN I o la ruta de activación del sistema del complemento.

Palabras clave – exosomas, leche materna, microRNA, ómicas, transcriptómica, lipidómica, mapeo, secuenciación.

Abstract

Preterm delivery is associated with adverse outcomes in the development of the neonate. In this regard, breast milk and its composition play an important role in the proper development of the newborn, since it is a food that is present in the first months of life. Human milk has a variable composition, in which apart from the nutrient profile we can analyze other types of molecules that help us understand the differences in said composition in premature births.

Specifically, in this work a transcriptomic and lipidomic study of exosomes is carried out from 10 samples of human milk, 5 of mothers of premature babies and 5 of mothers of full-term babies. The transcriptomic analysis is based on the study of the expression of microRNAs (miRNAs), which are a class of small, endogenous and non-coding RNAs that regulate the expression of human genes, while lipidomic analysis is based on the study of lipids. The objective of this study is to identify miRNAs and lipids with significant changes between both states and to study the biological processes in which they are involved. In order to do this, the data was processed and normalized to proceed with the statistical analysis.

As results of these analyses, 30 lipids with altered concentrations were obtained between both groups. From them, it was determined that the levels of triacylglycerides (TGs) and diacylglycerides (DGs) are higher in samples from mothers of premature babies, while the levels of sphingomyelins (SM) and ceramides (Cer) are higher in samples from mothers with full-term babies and are associated with pro-apoptotic effects. In transcriptomics, 2358 target genes were obtained that were enriched in the 56 miRNAs differentially expressed between the two groups compared. In turn, these target genes were related to different biological processes such as the immune response, the prostaglandin signaling pathway or the apoptosis signaling pathway, the Jak/STAT and PI3K signaling pathways, the CXC chemokine subfamily, the IFN I signaling pathway, or the complement system activation pathway.

Key words – exosomes, human milk, microRNA, omics, transcriptomics, lipidomics, mapping, sequencing.

Agradecimientos

En primer lugar, me gustaría dar las gracias a mi familia, mis padres Juan y María y mi hermano Pedro, que me han apoyado en todo momento, fuera cual fuera la decisión que quisiera tomar y de los que nunca he recibido un "No" por respuesta. Ellos siempre han estado donde los necesitaba, en casa cuando volvía a Lorca o en Valencia cuando había un acontecimiento especial o cuando simplemente, este acontecimiento era reunirnos después de algún tiempo.

Por otro lado, agradecer a Valencia y a la UPV por las relaciones y las experiencias de las que me ha hecho participe en estos 4 años. En esta época, no sólo he creado vínculos con personas de mi carrera que van a ser muy importantes durante el resto de mi vida, sino que también me ha hecho darme cuenta de que estaba rodeado de personas geniales y no era consciente de ello. Me llevo para siempre las clases en el 3P, los almuerzos en la cafetería de Bellas Artes, los viajes de Biobros sin ningún tipo de planificación y las fiestas, que fueran como fueran, siempre nos dejaban un recuerdo inolvidable.

En el ámbito académico, no he sido consciente durante todo este tiempo de que mis conocimientos hubieran alcanzado áreas que, al empezar la carrera, ni siquiera sabía de su existencia. Esto ha sido posible gracias al esfuerzo propio, a las horas de trabajos, clases, prácticas y estudio, pero he de confesar que no habría sido posible sin encontrarme por el camino con profesores que te hacían la vida mucho más fácil. Me siento afortunado por haber podido conocer un poco del lado humano de varios de ellos y esto, es algo que me llevo guardado.

Y cómo no, hacer una mención especial a mis cotutoras Sonia y Teresa. Sinceramente no sé qué habría sido de mí de no caer en vuestras manos. Me habéis ayudado sin ningún tipo de excusa, sacando huecos para mis peticiones dentro de vuestras ajetreadas agendas y, lo mejor de todo, siempre con una amabilidad y una calidad humana que no se encuentra en cualquier lugar. Por ello, quiero agradeceros de corazón todo vuestro esfuerzo y deciros que nunca me olvidaré de lo que habéis hecho por mí.

Agradecimientos a la Unidad de Bioinformática y Bioestadística del Centro de Investigación Príncipe Felipe (CIPF) por proporcionar acceso al clúster, cofundado por el Fondo Europeo de Desarrollo Regional (FEDER) en la Comunidad Valenciana 2014-2020.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

1.1. Consecuencias y trastornos relacionados con el nacimiento precoz de neonatos

Cada año, una cifra estimada de 15 millones de bebés en todo el mundo, lo que es más de 1 de cada 10, nacen de forma prematura y esta cifra continúa creciendo. Se considera prematuro a un nacimiento previo a las 37 semanas de gestación, pero a su vez hay subcategorías basadas en la edad gestacional, pudiendo ser extremadamente prematuro si el nacimiento se produce en menos de 28 semanas de gestación, muy prematuro si se sitúa entre 28 y 32 semanas, y moderadamente prematuro si se produce entre las semanas 32 y 37. Estas subdivisiones son de gran importancia, ya que la disminución de la edad gestacional está relacionada con un incremento de la mortalidad, discapacidad, intensidad del cuidado requerido por el neonato y por lo tanto, el aumento de los costes sanitarios que esto genera (Blencowe et al., 2012; WORLD HEALTH ORGANIZATION, 2018).

Estos nacimientos prematuros son debidos a numerosas razones. Muchos de ellos ocurren espontáneamente, pero algunos están ligados a la inducción del parto o al nacimiento por cesárea, ya sea por razones médicas o no médicas. Otras causas comunes del nacimiento prematuro incluyen embarazos múltiples, infecciones y afecciones crónicas como la diabetes; estrés, isquemias o hemorragias uteroplacentarias, o incluso podría deberse a una influencia genética, aunque a menudo la causa no llega a ser identificada (Blencowe et al., 2012).

Las complicaciones derivadas de los nacimientos prematuros lideran las causas de muerte entre niños de menos de 5 años, siendo responsables de aproximadamente 1 millón de muertes en el año 2015 (Figura 1) (Liu et al., 2016). A pesar de que la mayoría de los bebés de nacimientos prematuros sobreviven, tienen un mayor riesgo de alteraciones del desarrollo neurológico y complicaciones respiratorias y gastrointestinales. Además, pueden llegar a enfrentarse a una incapacidad de por vida, que incluye discapacidades cognitivas y problemas visuales y auditivos (Blencowe et al., 2012; Goldenberg et al., 2008).

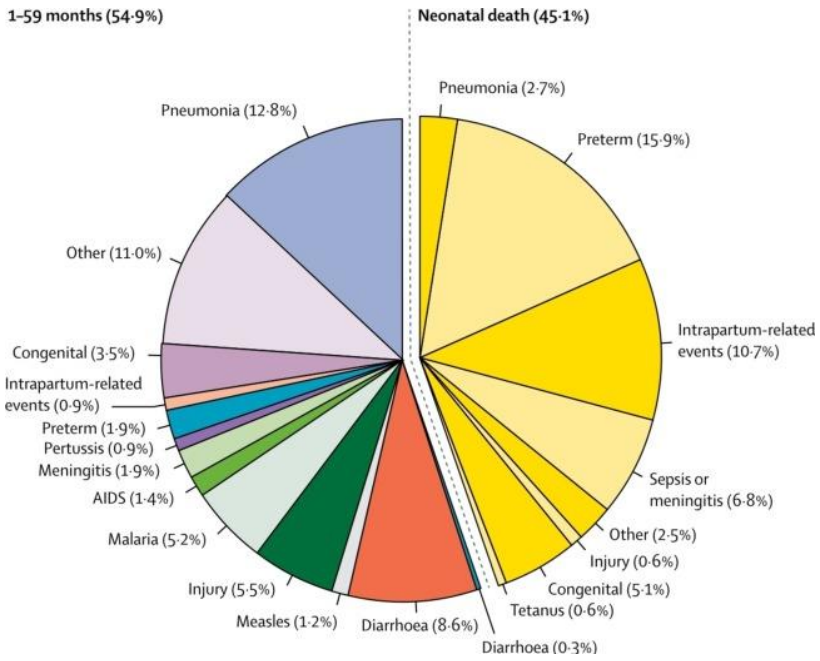


Figura 1. Causas globales de muertes de niños menores de 5 años en 2015. La zona amarilla del gráfico corresponde al porcentaje de muertes de neonatos y al conjunto de enfermedades responsables (Figura adaptada de Liu et al., 2016).

El cuidado de bebés extremadamente prematuros tiene complicaciones a muchos niveles y la alimentación no es una excepción. Algunos problemas derivados de un nacimiento antes de término requieren especial atención en cuanto a la alimentación se refiere. En cuanto a este aspecto, cobra especial importancia la leche materna, siendo un alimento del cual se benefician durante los primeros meses e incluso años de vida.

1.2. Leche materna: composición, características y aplicaciones

La leche humana proporciona una nutrición ideal para niños proporcionando todos los macronutrientes necesarios para el crecimiento y desarrollo neonatal. Comprender la composición de la leche materna proporciona una herramienta importante para la mejora de la alimentación infantil, particularmente de los bebés frágiles y de alto riesgo, como lo son muchos de los nacidos en estado de pretérmino (Ballard & Morrow, 2013). La leche materna probablemente sea el alimento funcional con mayor importancia hasta el momento, siendo un alimento dinámico que proporciona beneficios tanto nutricionales como de salud para infantes, y ayuda a su supervivencia y desarrollo saludable (Ballard & Morrow, 2013; Oftedal, 2012). Que la leche sea un fluido dinámico significa que su composición varía a lo largo del tiempo en función de diversos factores, tales como la alimentación, el periodo de lactancia, las poblaciones, el periodo de gestación o incluso el momento del día. Asimismo, podemos distinguir entre diferentes estados de leche materna en función del momento en el que se encuentre, como es el calostro (primera leche producida durante la lactancia), la leche de transición y la leche madura (Ballard & Morrow, 2013).

La composición de leche materna incluye una mezcla de diversos componentes tales como nutrientes (macronutrientes y micronutrientes), glóbulos grasos, hormonas, factores de crecimiento, células del sistema inmune, anticuerpos, citoquinas, péptidos antimicrobianos y vesículas extracelulares (EVs) que juegan un papel de gran importancia en el desarrollo del neonato y que también generan una gran variabilidad en su composición (Ballard & Morrow, 2013). En relación a esta composición, el contenido de lactosa y lípidos aumenta durante el período posparto inmediato, mientras que el contenido de proteína disminuye gradualmente (Ballard & Morrow, 2013). Se ha observado que la leche de madres que dieron a luz a bebés prematuros contiene una mayor cantidad de proteínas y lípidos que la leche perteneciente a madres que dieron a luz a bebés a término, disminuyendo esta cantidad con el aumento de la edad gestacional del infante.

La leche materna también contiene cientos o miles de moléculas bioactivas distintas que protegen contra la infección. Dentro de esta amplia variedad de componentes funcionales encontramos lactoferrina, lisozima, oligosacáridos e IGF-I (Factor de crecimiento insulínico tipo 1) han mostrado tener influencia en el desarrollo del intestino y del sistema inmune de los neonatos (Ballard & Morrow, 2013; Kahn et al., 2018). Además, protegen a los niños de enfermedades respiratorias, infecciones del oído medio, y enfermedades gastrointestinales, y generan efectos protectores contra la diabetes mellitus, la obesidad, hiperlipidemia, hipertensión, enfermedades cardiovasculares, autoinmunidad y asma.

Debido a la amplia variedad de compuestos que podemos encontrar y a pesar de la variabilidad debida a los factores nombrados, la leche materna se sitúa como una fuente candidata de la cual extraer materia biológica. Uno de los componentes de la leche materna que pueden ser utilizados como fuente de materia biológica son los exosomas, debido al gran contenido de moléculas que portan en su interior, para llevar a cabo estudios de ómicas y concluir una firma molecular que nos permita distinguir entre los tipos de leche materna que hay en función del carácter de término o pretérmino del nacimiento del bebé.

1.3. Los exosomas y su función biológica

Uno de los componentes de la leche son las EVs, un grupo heterogéneo de vesículas, entre ellas los exosomas, que son pequeñas vesículas de doble membrana lipídica de entre 30 y 150 nm de diámetro, liberadas al entorno extracelular desde diversas células. Inicialmente, se observó que la liberación de EVs formaba parte de un mecanismo de eliminación de desechos que tenía como objetivo descartar materiales inservibles en las células. Sin embargo, posteriores investigaciones mostraron que la liberación de EVs también forma parte de un mecanismo de comunicación intercelular que permite una conexión indirecta célula-célula a través de interacciones específicas con las células diana, englobado tanto en los procesos fisiológicos normales como en los de progresión de patologías (Abels & Breakefield, 2016; Ontoria-Oviedo et al., 2018). A su vez, los exosomas pueden modular tanto respuestas celulares como la actividad metabólica y, dado que cambian su composición según el estado fisiológico de la célula productora, pueden ejercer distintos efectos funcionales en las células diana (Garcia et al., 2016). Los exosomas tienen la estructura que se muestra en la Figura 2. Poseen unas proteínas de superficie que están asociadas con la adhesión a la membrana celular y el transporte, así como con la presentación de antígenos. Estas proteínas de membrana modulan el desarrollo infantil, ya que ejercen sus efectos reguladores a través de numerosos mecanismos (Kahn et al., 2018).

La biogénesis de los exosomas se inicia a través de una gemación hacia dentro de la membrana plasmática de las células (endocitosis) y continúa con los eventos de reconocimiento, clasificación y abscisión de la carga que involucra el complejo endosómico requerido para el transporte (ESCRT). Estos eventos conducen a la formación de cuerpos multivesiculares (MVB), que albergan los exosomas y que pueden secretarse al espacio extracelular o degradarse en lisosomas (Abels & Breakefield, 2016; Hurley & Odorizzi, 2012). Una vez formados los MVBs, se transportan a la membrana plasmática celular y se liberan las vesículas mediante exocitosis (Skotland et al., 2017). Sin embargo, la inactivación de la ruta ESCRT no inhibe la formación de MVBs, ya que hay otros mecanismos por los que los exosomas pueden ser generados en vías distintas independientes de ESCRT y que pueden operar en paralelo a esta ruta. Estos mecanismos varían dependiendo del tipo celular y del contenido de la vesícula. Entre otras, encontramos vías alternativas que dependen de distintos tipos de lípidos como las ceramidas de esfingolípidos (Abels & Breakefield, 2016; Zempleni et al., 2017).

Se ha visto que los exosomas de leche materna son capaces de encapsular diferentes tipos de moléculas incluyendo citoquinas, moléculas de transporte de membrana, quimiocinas, proteínas, lípidos y también, diversos tipos de moléculas de RNA como small RNAs y mRNAs, tal y como se muestra en la Figura 2, otorgándoles protección frente a la degradación enzimática y no enzimática y proporcionando una vía para la captación de cargas por endocitosis de exosomas (Garcia et al., 2016; Ontoria-Oviedo et al., 2018; Zempleni et al., 2017). Por otro lado, la composición lipídica de los

exosomas comparte características comunes con las células de origen, aunque investigaciones más específicas han mostrado que algunos lípidos pueden estar específicamente asociados a diferentes tipos de EVs (Abels & Breakefield, 2016). En cuanto a los lípidos enriquecidos en EV se encuentran la esfingomielina, el colesterol, el ganglíosido GM3, los lípidos disaturados, la fosfatidilserina y la ceramida. En contraste, la fosfatidilcolina y el diacil-glicerol aparecen en menor cantidad comparado con la composición de la membrana lipídica de la célula de origen (Laulagnier et al., 2004; Llorente et al., 2013).

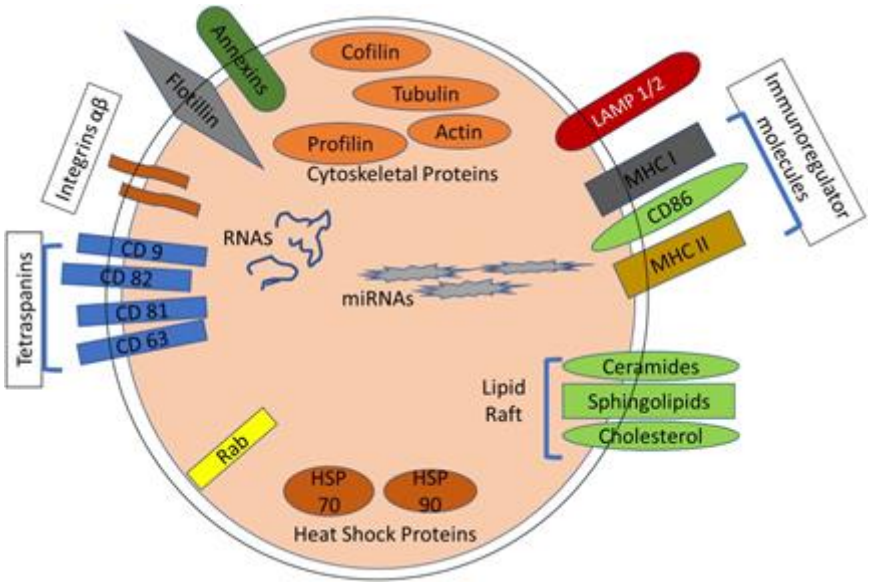


Figura 2. Composición estructural y contenido de los exosomas. Se observan diferentes moléculas de membrana, entre ellas lípidos; y componentes nucleicos internos tales como miRNAs o RNAs en general (Figura adaptada de Farooqi et al., 2018).

Muchos tipos celulares, como las células del sistema inmune, células epiteliales o células cancerosas pueden internalizar exosomas secretados por otras células. Además, se ha visto que los exosomas de leche materna son resistentes a los procesos de digestión y son internalizados por células epiteliales del intestino del bebé (Liao et al., 2017), lo que ocurre de la misma manera con los miRNAs que se encuentran encapsulados en los exosomas de la leche (Kahn et al., 2018). Numerosos estudios han propuesto que la transferencia de miRNAs procedentes de exosomas puede modular las funciones biológicas de las células aceptoras. Los exosomas son de particular interés en este aspecto porque su carga de RNAs no es un proceso aleatorio, sino que implica mecanismos de clasificación que favorecen algunas cargas sobre otras (Abels & Breakefield, 2016).

1.4. miRNAs y lípidos en exosomas

Las concentraciones de las moléculas que transportan los exosomas se pueden medir utilizando técnicas de alto rendimiento para poder estudiar el conjunto completo de todas ellas. En concreto, en este trabajo se analiza la expresión de microRNAs (miRNAs) medida mediante técnicas de secuenciación (miRNA-Seq) y la concentración de lípidos usando espectrometría de masas (lipidómica) en exosomas de leche materna.

Los miRNAs son un tipo de RNAs pequeños no codificantes, de cadena simple y de entre 18 y 23 nucleótidos de longitud, que regulan la expresión de los genes y, en consecuencia, regulan la síntesis de proteínas a nivel postranscripcional en las células eucariotas. Un mismo miRNA puede regular la expresión de numerosos genes (Alsaweed et al., 2016) y tienen una gran importancia, ya que regulan más del 60% de la expresión de los genes humanos. Los precursores de los miRNAs son exportados desde el núcleo celular al citoplasma, donde son convertidos a miRNAs maduros por la enzima DICER (Squadrito et al., 2014). Han sido identificados como reguladores clave de diversos procesos biológicos y del desarrollo en eucariotas, como en la proliferación y diferenciación celular, la apoptosis, el desarrollo de sistema inmunológico y la respuesta inmune, etc. Los miRNAs son capaces de realizar estas funciones gracias a un mecanismo de marcaje que actúa afectando a la estabilidad del mRNA durante su traducción a proteína mediante el que lo degrada o inhibe el proceso de traducción (Alsaweed et al., 2016). Al cargarse en el complejo de silenciamiento inducido por RNA (RISC), los miRNAs se unen a la zona UTR 3' de los transcritos diana para así modificar su expresión.

Más allá de los tejidos, los miRNAs pueden aislarse de diferentes fluidos corporales, tales como plasma, orina, saliva, lágrimas o leche materna (Weber et al., 2010). Además, de la misma manera que se ha mencionado anteriormente, los exosomas presentes en los fluidos corporales portan miRNAs y los protegen contra la digestión, facilitando su función reguladora en diferentes tejidos y órganos (Alsaweed et al., 2016). De esta manera, existen miRNAs específicos que están enriquecidos en exosomas con dependencia de la célula productora (Squadrito et al., 2014).

Las características que reúnen los miRNAs los convierten en un biomarcador ideal debido a que son accesibles mediante protocolos no invasivos y no excesivamente costosos de cuantificar. Además, se ha demostrado el potencial de usar concentraciones de miRNAs específicos en fluidos corporales como biomarcadores para detectar y monitorear diversas condiciones fisiopatológicas (Weber et al., 2010); e incluso valores anómalos de expresión de miRNAs han sido asociados con patologías, incluyendo diferentes tipos de cáncer, inflamación o diabetes (Lu et al., 2008).

Por otra parte, los lípidos son componentes estructurales esenciales de las membranas y poseen numerosas funciones celulares cruciales, actuando como moléculas de señalización, identificadores químicos de membranas específicas o como moléculas de almacenamiento de energía. Estos lípidos contribuyen a otorgar un carácter polar a las membranas celulares, que consisten en una cara interna hidrofóbica y una cara externa hidrofílica. La naturaleza de las moléculas hidrofóbicas las conduce a asociarse entre ellas (consecuencia de la entropía generada por la repulsión hacia agua), y la tendencia de las moléculas hidrofóbicas a interaccionar con ambientes acuosos, son la base física de la formación espontánea de membranas. Este carácter anfipático de los lípidos es una propiedad química que les permite a las células segregar su contenido interno al ambiente extracelular. Los lípidos no solo juegan este papel estructural en la membrana, sino que también pueden actuar como primeros y segundos mensajeros en procesos de transducción de señal y de reconocimiento molecular. A su vez, son moléculas que pueden adoptar los estados sólido y líquido, lo que está relacionado con diferentes configuraciones espaciales y libertad de movimiento. Todas estas características que comparten son las que les permiten realizar un repertorio tan amplio de funciones (van Meer et al., 2008).

Los lípidos también forman parte del contenido de los exosomas y se encuentran presentes en su membrana. Se sabe que hay lípidos específicos que están enriquecidos en los exosomas en comparación con sus células parentales. Entre el conjunto de lípidos enriquecidos en estas vesículas, se encuentra colesterol, esfingomielinas (SM), glicoesfingolípidos y fosfatidilserina (PS). Por otro lado,

los exosomas generalmente contienen menos concentraciones fosfatidilcolina (PC) (Skotland et al., 2017).

El estudio de los lípidos es de gran interés, ya que pueden proporcionar una imagen directa del estado metabólico celular (Han, 2016). Por ello, también son biomoléculas que nos permiten obtener información cuantitativa de las muestras y poder compararlas entre ellas.

1.5. Técnicas de secuenciación y transcriptómica para el estudio de los miRNAs

Las tecnologías de alto rendimiento *(high throughput)* permiten el estudio de la biología celular a diferentes niveles de organización molecular. Estas metodologías han evolucionado rápidamente en los últimos 15 años y dentro de ellas, destaca el uso de las nuevas tecnologías de secuenciación masiva o *Next-Generation Sequencing* (NGS), que en los últimos tiempos ha permitido mejorar la sensibilidad, especificidad y profundidad de secuenciación, unido a una notable mejora del coste del propio proceso.

Hace más de 40 años, el desarrollo de la tecnología de secuenciación Sanger (Secuenciación de Primera Generación) revolucionó el campo de la investigación biológica. Posteriormente, las implicaciones de esta técnica se volvieron aún más trascendentales con la introducción de las NGS (Secuenciación de Segunda y Tercera Generación). Dentro de las NGS de segunda generación observamos técnicas de secuenciación tales como Illumina, Solid o Roche 454. Estas técnicas lograban paralelizar el proceso de secuenciación, incrementando notablemente la cantidad de datos generada, produciendo miles o millones de secuencias al mismo tiempo. Como consecuencia, disminuyeron los costes implicados en la obtención de este tipo de datos. Estas mejoras unidas a la innovación metodológica y el desarrollo computacional han facilitado una explosión de conocimiento biológico en nuestra era (Mardis, 2017).

En la pasada década, las tecnologías de secuenciación de moléculas únicas de DNA de lectura larga han emergido en el campo de la genómica, teniendo un papel de gran importancia dentro de ella. Con la habilidad de generar lecturas de decenas de miles de kilobases de longitud con una precisión cercana a las de las tecnologías de secuenciación de lectura corta, tales como Illumina, estas plataformas han mostrado su capacidad para resolver algunas de las regiones más desafiantes del genoma humano o detectar variantes estructurales previamente inaccesibles. Aún es necesario mejorar sus prestaciones y corregir ciertos sesgos, pero las tecnologías de secuenciación de lectura larga como PacBio, se sitúan como herramientas prometedoras en el campo de la investigación biológica (Logsdon et al., 2020).

Dentro de los diferentes tipos de secuenciación, la secuenciación de RNA (en concreto, small RNA-Seq) ha demostrado que los miRNAs se encuentran en abundancia dentro de los exosomas. Este descubrimiento sugiere que la clasificación de especies de miRNAs específicos en exosomas puede estar activamente regulada, aunque los mecanismos por los que se produce aún se desconocen. Tanto el proceso de biogénesis de MVB/exosomas como los determinantes específicos de la secuencia de miRNAs pueden modular la clasificación de miRNAs en exosomas (Squadrito et al., 2014). El uso de tecnologías tales como las técnicas de secuenciación permiten obtener un conjunto de datos perteneciente a cada una de las muestras analizadas. La naturaleza de estos datos es cuantitativa, por lo que son medidas que pueden compararse entre distintos grupos o estados.

El transcriptoma es el conjunto completo de los transcritos (mRNAs, RNAs no codificantes y RNAs de pequeño tamaño) dentro de una célula y sus cantidades, para un estado de desarrollo o una condición fisiológica específica. El entendimiento del transcriptoma es esencial para interpretar los elementos funcionales del genoma y revelar los constituyentes moleculares de células y tejidos, y también para comprender los procesos de desarrollo y las enfermedades (Z. Wang et al., 2009). Asimismo, la transcriptómica se define como el estudio del transcriptoma usando métodos de alto rendimiento como análisis de microarrays o RNA-Seq. La comparación de transcriptomas permite la identificación de genes que están diferencialmente expresados en distintas poblaciones celulares o en respuesta a diferentes tratamientos (NATURE PORTFOLIO, 2021).

1.6. Lipidómica

Al conjunto total de lípidos contenidos en una célula se le conoce como lipidoma, y, por lo tanto, la lipidómica es el estudio de lipidomas mediante el uso de los principios y las técnicas de la química analítica. La lipidómica emergió en 2003 como un acercamiento para el estudio del metabolismo del lipidoma celular (Han, 2016). El poder analítico y los nuevos desarrollos de la Espectrometría de Masas (MS) han acelerado esta disciplina emergente. Estos desarrollos no solo se encuentran en esta técnica, sino que también se extienden a plataformas basadas en cromatografía líquida, nuevas estrategias de desorción o ionización en espectrometría de masas para análisis lipidómicos, y en el avance de nuevas herramientas bioinformáticas para mejorar la identificación y la cuantificación de los constituyentes moleculares individuales que componen el lipidoma de cada célula (Han et al., 2012). A su vez, la lipidómica también proporciona una herramienta poderosa para el descubrimiento de marcadores lipídicos para el estudio de estados de enfermedades (Han, 2016).

Es importante destacar que este enfoque nos permite estudiar el metabolismo celular cuantificando los cambios de clases, subclases y especies moleculares de lípidos individuales que reflejan diferencias metabólicas. Dado que las rutas y redes del metabolismo de los lípidos han sido ampliamente estudiadas, cualquier cambio en la cantidad de lípidos puede revelar variaciones a niveles enzimáticos, actividades o patrones de expresión génica (Han, 2016).

1.7. Breve descripción del estudio

El presente trabajo se realizó en colaboración con el Instituto de Investigación Sanitaria La Fe de Valencia. El objetivo del proyecto es estudiar las diferencias de composición de la leche materna entre dos grupos: madres de bebés nacidos de forma prematura y madres de bebés nacidos a término. Para ello, se llevó a cabo un estudio transcriptómico y lipidómico del contenido de exosomas extraídos de la leche materna de 10 madres, 5 de cada grupo. El estudio transcriptómico se basó en un análisis de la expresión de miRNAs obtenido mediante small RNA-Seq. Por otro lado, los datos de lipidómica se obtuvieron por Cromatografía Líquida acoplada a Espectrometría de Masas (LC-MS) con Ionización por electrospray (ESI) tanto positiva como negativa. Ambas ómicas se midieron en las mismas 10 muestras de leche y los datos fueron obtenidos mediante la extracción de exosomas de la leche materna, en cuya composición interna se encontraban almacenadas dichas moléculas. De esta manera, se pretende identificar lípidos y miRNAs cuya expresión se encuentra alterada entre ambos estados y estudiar los procesos biológicos en los que están implicados.

2. Objetivos

El objetivo principal del presente trabajo es la determinación de variaciones ómicas, específicamente transcriptómica y lipidómica, en la composición de la leche materna de madres de neonatos nacidos a término y pretérmino.

Para alcanzar el objetivo citado, nos planteamos una serie de objetivos secundarios:

- Realizar un control de calidad de los datos previo al estudio para garantizar su viabilidad.

- Aplicación y comparación de distintos métodos de pre-procesado y normalización de los datos iniciales en ambas ómicas con el fin de adaptar y optimizar el análisis a los datos con los que se trabaja.

- Aplicación de métodos estadísticos para encontrar alteraciones en la expresión de miRNAs y concentración de lípidos en la leche de los dos grupos estudiados.

- Realización de análisis de enriquecimiento para estudiar las rutas biológicas en las que las moléculas alteradas están involucradas.

3. Materiales y métodos

3.1. Datos utilizados en el estudio

3.1.1. Diseño experimental

El estudio se llevó a cabo en humanos, concretamente en madres de entre 26 y 42 años. Se estudiaron las variaciones en la composición de la leche materna en dos grupos: madres de neonatos nacidos a término y madres de neonatos nacidos de forma prematura o pretérmino, contando con 5 madres dentro de cada uno de los grupos mencionados. Los datos y la información sobre obtención de muestras, los procedimientos de extracción de exosomas, ómicas e información relacionada con el diseño experimental fue proporcionada por el Grupo de Investigación de Regeneración y Trasplante Cardíaco del Instituto de Investigación Sanitaria La Fe.

En cuanto a la información correspondiente a los propios neonatos y a sus madres se encuentra indicada en la Tabla 1. Los neonatos son de sexo tanto femenino como masculino, encontrándose ambos géneros en los dos grupos de estudio. El peso en el momento de la toma de muestras en el grupo prematuro varía de 1300 a 2620 g, valores claramente inferiores a los correspondientes al grupo a término que se encuentran entre 2860 y 3350 g, exceptuando un valor de 1930 g. Los valores de peso al nacer son similares a los del tiempo de toma de muestras en estado a término, sin embargo, en el estado pretérmino se observan pesos de nacimiento de en torno a 560 o 900 g, mucho menores a los del tiempo de la toma de la muestra. La edad gestacional varía en el grupo a término de 38 a 40 meses, mientras que en el grupo pretérmino se reduce a tiempos de entre 24 a 31 meses. Por otro lado, la cantidad de comida ingerida se mantiene con valores similares en ambos grupos.

Grupo	Peso en la toma de la muestras (g)	Peso al nacer (g)	Edad gestacional (meses)	Comida ingerida (mL/kg)	Sexo del neonato	Edad de la madre (años)
Término 1	1960	1930	38+2	170	F	42
Término 2	3315	3150	39+0	155	M	-
Término 3	3300	3350	39+4	150	F	36
Término 4	2800	2860	38+2	150	M	35
Término 5	3170	3160	40+2	150	M	42
Pretérmino 1	1820	1620	31+2	150	F	33
Pretérmino 2	2620	1400	30+1	170	F	42
Pretérmino 3	2060	900	24+6	180	F	26
Pretérmino 4	1300	1300	30+0	150	M	35
Pretérmino 5	1310	560	24+2	180	F	27

Tabla 1. Información correspondiente a los pacientes: madres y neonatos. Se muestra en cada fila las características de cada paciente, incluyendo en ella el grupo de estudio al que pertenece, término o pretérmino; el peso del neonato en gramos en el momento en el que nació y en el que se procedió a tomar la muestra de leche materna; la edad gestacional del neonato en meses, la cantidad de comida ingerida por el neonato en mL de leche por kg de peso del bebé, el sexo del neonato, correspondiendo M a masculino y F a femenino; y la edad de la madre en años.

La composición de la leche se evaluó gracias a la extracción de exosomas de la misma, a partir de los cuales se obtuvieron muestras de miRNAs y lípidos almacenados en su interior para la generación de datos ómicos, transcriptómicos y lipidómicos respectivamente, los cuales fueron analizados en el presente trabajo.

En este estudio todas las leches se han tomado en un único tiempo, siendo el mismo momento con respecto a la ingesta de leche del bebé (i.e., en el momento en el que el bebé alcanza la nutrición enteral completa en el caso de los prematuros y, equivalentemente, en el momento en el que el recién nacido a término recupera el peso al nacer).

3.1.2. Extracción de los exosomas

Los datos de transcriptómica y lipidómica se extrajeron a partir de los exosomas de la leche materna. La extracción de los exosomas fue realizada por el Instituto de Investigación Sanitaria La Fe siguiendo el protocolo de extracción detallado en la Figura 3. Este se basaba en ultracentrifugación consistente en varios ciclos de centrifugación, filtración y ultracentrifugación de muestras de 25 mL de leche materna.

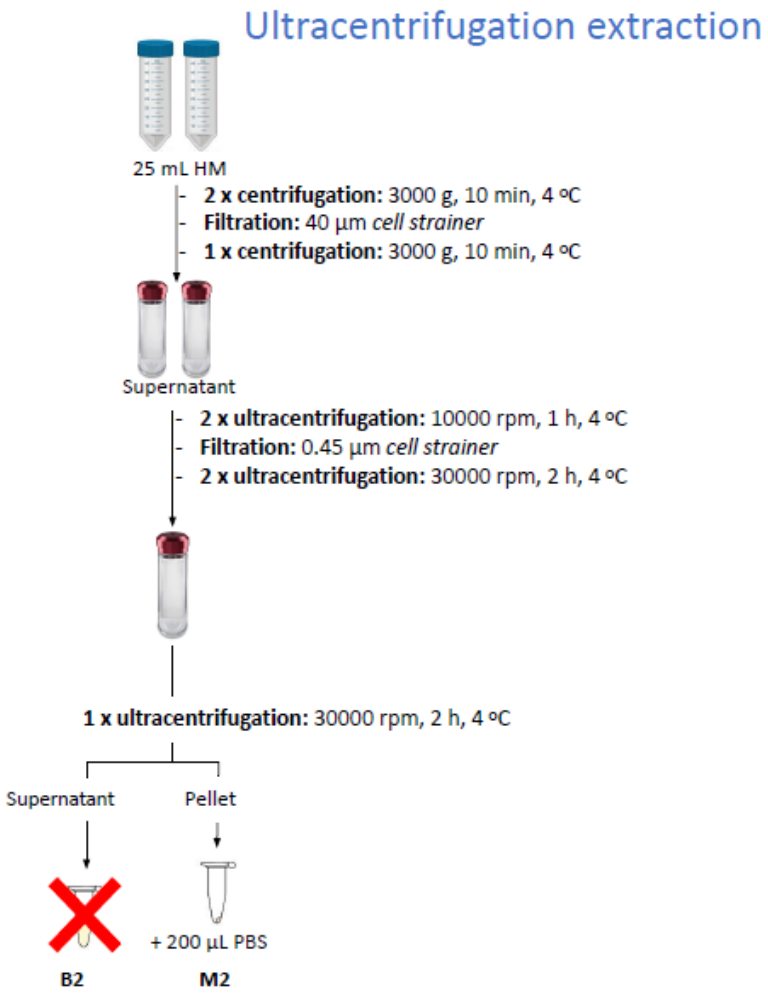


Figura 3. Condiciones de la extracción de exosomas. HM (*human milk* o leche humana), PBS (*Phosphate-buffered saline* o tampón fosfato salino).

3.1.3. Datos de transcriptómica

Los pasos iniciales del procedimiento seguido para obtener los datos de transcriptómica consistieron en realizar la extracción de RNA y posteriormente aislar la fracción de small RNA. Esta primera parte fue realizada por el Instituto de Investigación Sanitaria La Fe. Posteriormente, las muestras de small RNA fueron enviadas al CRG (Centre for Genomic Regulation) donde se prepararon las librerías y se

llevó a cabo la secuenciación RNA de pequeño tamaño (small RNA-Seq) de las 10 muestras utilizadas en el estudio. Las librerías fueron preparadas usando el kit NEBNext Small RNA Library Prep Set de Illumina (ref. E7330) de acuerdo con el protocolo del fabricante. Brevemente, 100 ng de RNA se sometieron a ligación del adaptador 3' y 5', y la síntesis de la primera hebra de cDNA. Tras ello, una PCR enriqueció selectivamente aquellos fragmentos de DNA que tenían moléculas adaptadoras en ambos extremos. La amplificación de la librería se realizó mediante PCR utilizando NEWNext Multiplex Oligos para Illumina (Index Primers Set 1, ref. E7335), (Index Primers Set 2, ref. E7500), (Index Primers Set 3, ref. E7710) e (Index Primers Set 4, ref. E7730). Todos los pasos de purificación se realizaron utilizando perlas AgenCourt AMPure XP (ref. A63882, Beckman Coulter). Las librerías finales se analizaron usando Agilent Bioanalyzer (ref. 5067-4626) para estimar la cantidad y verificar la distribución del tamaño. Se realizó una agrupación para realizar la selección de tamaño utilizando 6% Novex TBE PAGE Gels (ref. EC6265BOX) y luego la agrupación final se cuantificó mediante qPCR utilizando el KAPA Library Quantification Kit (ref. KK4835, KapaBiosystems) antes de la amplificación con cBot de Illumina. Finalmente, las librerías fueron secuenciadas en HiSeq2500 de Illumina. En cuanto a la secuenciación, la longitud de lectura se situó en 50 pares de bases (bp), obteniendo un total de lecturas cercanas a 10 millones para cada una de las diferentes muestras tal y como se muestra en la Tabla 2.

Nombre de la muestra	Número de lecturas
Término 1	9.255.941
Término 2	13.393.757
Término 3	8.148.215
Término 4	11.975.282
Término 5	10.414.733
Pretérmino 1	12.146.293
Pretérmino 2	10.746.113
Pretérmino 3	11.447.548
Pretérmino 4	10.087.235
Pretérmino 5	11.501.036

Tabla 2. Información del número de lecturas de las muestras secuenciadas.

3.1.4. Datos de lipidómica

Los datos de lipidómica fueron facilitados por el Instituto de Investigación Sanitaria La Fe. El método utilizado para analizar los lípidos contenidos en la muestra fue la técnica de cromatografía líquida acoplada a espectrómetro de masas (LC-MS). Para ello, se procedió a preparar las muestras siguiendo el protocolo mostrado en la Figura 4. Las muestras posteriormente atravesaron una columna de cromatografía líquida de alta resolución y sensibilidad denominada UPLC. La tecnología utilizada para ello fue la UPLC-QqTOF Agilent 6550, con la columna Acquity BEH C18 (100 mm x 2.1 mm, 1.7µm). El método de ionización utilizado para el análisis por MS fue la Ionización por electrospray tanto positiva como negativa (ESI +/ ESI -), obteniendo así 2 conjuntos de datos, y los datos se analizaron de forma simultánea por Full scan y por MS en tándem (MS/MS) para obtener realizar un análisis más selectivo (del Mar Gómez-Ramos et al., 2015).

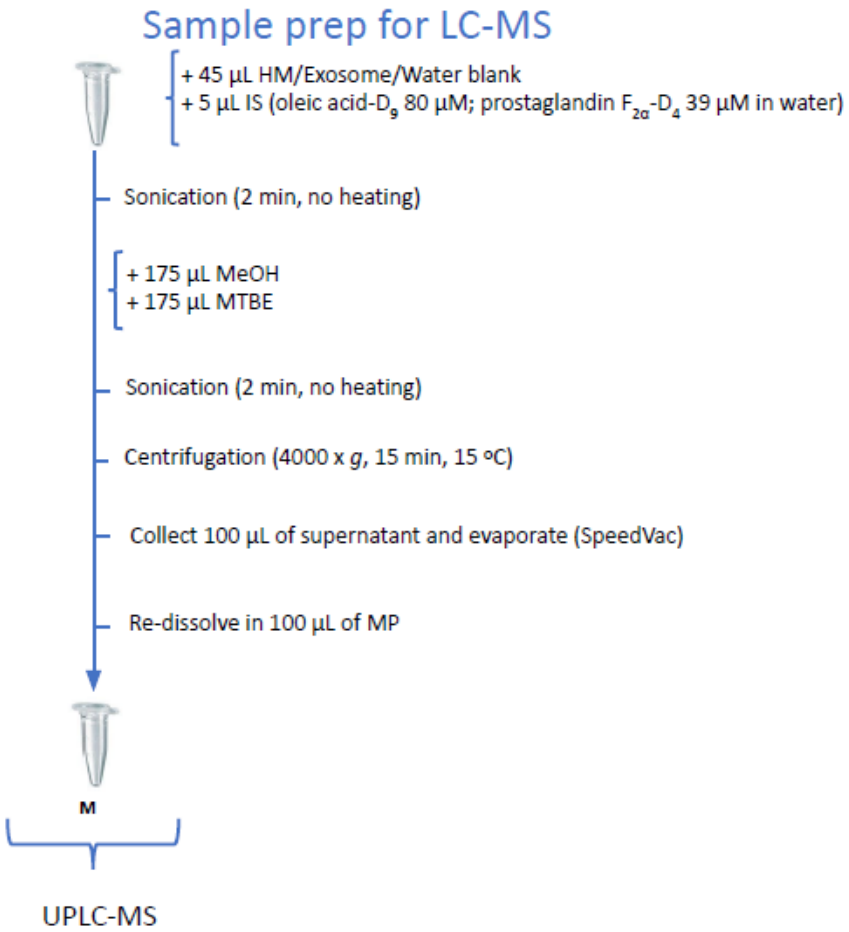


Figura 4. Protocolo de preparación de muestra para UPLC-MS. HM (*human milk* o leche humana), MeOH (metanol), MTBE (metil ter-butil éter), MP (*mobile phase* o fase móvil).

Los datos de lipidómica fueron proporcionados en dos tablas, obtenidas con ESI+ y ESI-, con valores correspondientes a las medidas de cada uno de los lípidos analizados en las 10 muestras medidas, junto con el nombre, la clase y la subclase del lípido, la relación masa carga (m/z) y el tiempo de retención (RT) obtenidos en el análisis.

3.2. Pre-procesado de los datos de transcriptómica

3.2.1. Control de calidad de las lecturas de miRNA-Seq

Antes de comenzar el procesamiento de los datos de secuenciación de las diferentes muestras, se procedió a realizar un control de calidad de las lecturas con el objetivo de hallar sesgos o problemas previos al análisis, producidos por errores asociados a las muestras o incluso a la propia tecnología de secuenciación. Para llevarlo a cabo, se utilizó la herramienta de control de calidad **FastQC** con la versión 0.11.8 (BABRAHAM BIOINFORMATICS, 2010). Entre los diferentes aspectos analizados por **FastQC** se encuentra la calidad de secuenciación, el contenido de las diferentes bases nitrogenadas a lo largo de las secuencias y el contenido de GC, entre otros.

Dicho control de calidad se volvió a realizar tras la eliminación de los adaptadores (descrita en la siguiente sección), para comprobar que el proceso se había realizado correctamente y poder observar las características finales de las muestras con las que se procedía a realizar el estudio.

3.2.2. Eliminación de los adaptadores

Una vez realizado el control de calidad inicial, se procedió a eliminar los adaptadores de las muestras que vamos a procesar. Este es un paso esencial en el pre-procesado de las lecturas, ya que los adaptadores están introducidos en la secuencia de la que se ha obtenido la lectura y si no son eliminados, interferiría en el alineamiento de las lecturas de secuenciación sobre el transcriptoma de referencia.

La eliminación de adaptadores se llevó a cabo con la herramienta **Cutadapt**, con la versión 1.9.1 (Martin, 2011). Este proceso se realizó de forma paralela con diferentes parámetros para comparar los resultados obtenidos con cada uno de ellos. La secuencia de adaptador que se incluyó para ser eliminada fue AGATCGGAAGAGCACACGTCT y se obtuvo del "Manual de Illumina de Secuencias de Adaptadores" para TruSeq.

En cuanto a los parámetros, se fijó un valor de calidad 30 (Ewing & Green, 1998) para eliminar aquellas secuencias con una calidad media inferior a dicho valor. También se utilizó otro parámetro para filtrar aquellas lecturas que no superaran cierta longitud tras la eliminación de los adaptadores. Para este parámetro se analizaron dos valores diferentes, 10 y 15, y por lo tanto, se eliminaron las lecturas con menor longitud que la determinada en cada caso. Sin embargo, la secuencia del adaptador a eliminar y el parámetro de calidad con valor de 30 se mantuvieron invariables en ambos casos.

3.2.3. Mapeo de las lecturas de miRNA-Seq

Una vez los adaptadores han sido eliminados de las lecturas, corresponde encontrar a qué secuencias nucleotídicas del material genético pertenecen dichas lecturas. Por ello, se lleva a cabo un proceso que se conoce como mapeo o alineamiento. Mapear consiste en alinear las lecturas frente a un genoma o transcriptoma de referencia, para obtener las zonas del mismo a las que pertenecen. En este caso, se mapearon las lecturas al transcriptoma humano, descargado de la base de datos miRBase con la versión 22.1 (Kozomara et al., 2019).

Para ello, se utilizó la herramienta **Bowtie2** con la versión 2.3.2 (Langmead & Salzberg, 2012). Otros mapeadores fueron también considerados, pero finalmente se decidió trabajar con **Bowtie2** por la obtención de mejores resultados en proyectos similares de nuestro grupo realizados con anterioridad. Este mapeador, a su vez, admite numerosos modos de alineación y con diferentes parámetros, por lo que se puede ajustar fácilmente a los requerimientos del estudio.

En primer lugar, es necesario indexar el transcriptoma de referencia. Este paso se realiza con la función "bowtie2-build" incorporada en **Bowtie2**. Una vez indexado el transcriptoma, se procede a mapear. Los parámetros utilizados en este caso fueron los incorporados por defecto en la propia herramienta, utilizando el índice creado en el paso previo. De esta manera, se realizó un alineamiento global, es decir, un alineamiento que involucra todos los nucleótidos de las secuencias utilizadas, y sin especificidad de hebra. El modo predeterminado de **Bowtie2** también busca múltiples alineamientos para cada secuencia, presentando el mejor de ellos. Entre los parámetros por defecto se encuentra el número de *mismatches* permitidos en 0, entendiendo por *mismatches* a las discrepancias de secuencias entre las lecturas y el transcriptoma de referencia; y la longitud de la semilla de mapeo en 20 nucleótidos, definiendo a la semilla de mapeo como la subsecuencia de una lectura que algunos mapeadores alinean en primer lugar para posteriormente extender el alineamiento en el resto de la secuencia.

3.2.4. Filtro de calidad de las lecturas mapeadas

El alineamiento de cada secuencia al transcriptoma de referencia lleva asociado un valor de calidad del mapeo, denominado MAPping Quality (MAPQ), proporcionado por Bowtie2. Este valor MAPQ está relacionado con la probabilidad de que una posición de mapeo sea errónea y varía entre 0 y 42 en Bowtie2. Un mayor valor de MAPQ corresponde a una mayor calidad de mapeo (ACGT, 2014). De esta manera, se consideró oportuno realizar varios filtros con valores de MAPQ de diferente exigencia. La herramienta utilizada para este propósito fue **SAMtools** con la versión 1.10 (Li et al., 2009), que posee diferentes utilidades para ficheros con formato de secuencia de alineamiento/mapeo (SAM). En particular, se utilizó samtools view, una funcionalidad que permite ver y convertir ficheros con este tipo específico de formato. De esta manera, se utilizó samtools view para filtrar los mapeos por 7 valores de calidad diferentes, es decir, valores de MAPQ superiores a 1, 2, 3, 4, 5, 10 ó 20.

3.2.5. Cuantificación de la expresión de miRNAs

La estimación de la expresión de los miRNAs en las tecnologías de secuenciación masiva es el número de lecturas alineadas a cada miRNA. Se utilizó un *script* en lenguaje de programación Bash específicamente diseñado para realizar dicha cuantificación tras haber aplicado los filtros de calidad. Su funcionamiento se basa en filtrar la columna del fichero de los alineamientos filtrados en la que aparece el nombre de miRNA al que pertenece ese alineamiento. Una vez obtenido el miRNA se ordenan alfabéticamente y se agrupan obteniendo una única fila para cada miRNA con un valor correspondiente a la suma de todas las veces que aparece en el fichero de alineamientos. De esta manera, obtenemos ficheros con los nombres de todos los miRNAs presentes en cada muestra y el número de veces que aparece. Esta operación se hizo tras aplicar distintos filtros con el MAPQ a los resultados del alineamiento, en concreto, MAPQ>1 (MAPQ1 en adelante) y MAPQ>5 (MAPQ5 en adelante). Por último, se crearon las matrices de conteos a partir de estos ficheros, utilizando el paquete de R **dplyr** (Mailund, 2019). A los miRNAs no detectados en alguna de las muestras, se les adjudicó un valor de expresión de 0.

Desde la obtención de la cuantificación tras ambos filtros, se procedió a seguir los análisis posteriores con ambos conjuntos de datos por vías paralelas, es decir, con los conteos tras los filtros MAPQ1 y MAPQ5. De esta manera, todos los procedimientos explicados a continuación se realizaron por duplicado para comparar resultados entre ambos conjuntos de datos y decidir posteriormente, con cuál de ellos se obtenían mejores resultados.

3.2.6. Filtro de miRNAs de baja expresión

Una vez obtenida la matriz de conteos, se procedió a aplicar un filtro de baja expresión, ya que los miRNAs poco expresados no son útiles en la clínica y además la estimación de su expresión puede ser ruidosa y poco fiable. Se filtró la matriz de conteos probando 7 valores distintos de CPM (*counts per million*, es decir, número de conteos por millón de lecturas de secuenciación), eliminando aquellos miRNAs que no contaran con unos valores de conteo que superaran el umbral de CPM aplicado en ninguno de los dos grupos estudiados (nacimientos a término y pre-término). Los 7 valores de CPM que se testaron fueron 0, 0.5, 1, 2, 3, 4 y 5. Dicho filtro se aplicó usando la función filtered.data() del paquete de R **NOISeq** (Tarazona et al., 2012). Esta función se basa en el cálculo de la media de conteos por millón en las dos condiciones estudiadas, eliminando aquellos miRNAs en los que este valor no alcance el valor fijado en ninguna de las dos condiciones.

3.3. Control de calidad de los datos ómicos

3.3.1. Análisis de componentes principales (PCA)

El PCA es un método multivariante de reducción de la dimensión, que se utilizó para explorar de manera eficiente los dos conjuntos de datos ómicos utilizados en el estudio.

El PCA es un algoritmo matemático que permite reducir el número de variables de un estudio mediante la creación de nuevas variables latentes denominadas componentes, que representan la mayor parte de la variabilidad del conjunto original de datos (Ringnér, 2008). Cada componente principal (PC) es una combinación lineal de las variables originales y explica un porcentaje de la varianza de los datos, siendo la primera componente principal (PC1) la que explica una mayor variabilidad, seguido de la segunda (PC2) y así sucesivamente. De esta manera, es posible explicar la mayor parte de la variabilidad original eligiendo un número reducido de estas componentes. El peso de las variables originales en cada una de las PCs es el *loading*. De esta manera, también se obtiene el peso de las observaciones en las PCs. Estos pesos se llaman *scores*. Se pueden representar gráficamente las proyecciones de las variables o de las observaciones en las PCs y estos gráficos pueden ayudar a comprender la relación entre ellas y sirve como control de calidad para comprobar que se las observaciones se agrupan en relación a los grupos experimentales a los que pertenecen, o si la separación de dichos grupos cumple con el comportamiento esperado.

El PCA se utilizó para examinar los datos de cuantificación de expresión en miRNA-Seq o de concentración de lípidos, realizando dicho análisis en numerosas etapas del proyecto con los valores de cuantificación crudos, o tras la aplicación de los métodos de normalización.

Para obtener los distintos PCAs se utilizó el paquete de R **NOISeq**.

3.3.2. Identificación y correlación de lípidos repetidos

De los 680 lípidos que fueron identificados por MS, muchos compartían el mismo identificador en los campos nombre, clase y subclase de lípido, difiriendo tan sólo en la relación m/z y en el RT. Este fenómeno se encontraba de diferentes maneras en los conjuntos de muestras analizadas, observando repeticiones de lípidos en variables, tanto dentro de cada tipo independiente de ionización, como entre ambos tipos. El número de repeticiones de los diferentes lípidos repetidos se situaba generalmente en el valor 2, excepto en ciertos casos, 5 en total, en los que el lípido se encontraba repetido en 3 ocasiones. Por ello, se vio conveniente examinar los datos de cuantificación correspondientes a los lípidos repetidos, analizando los valores dentro de cada grupo de lípidos repetidos. Para ello, se estudió la correlación entre los valores de cada lípido en las 10 muestras evaluadas mediante gráficos de dispersión y calculando el coeficiente de correlación de Pearson entre cada par de lípidos repetidos. Para mejorar la visualización de los valores en los gráficos de dispersión, se procedió a transformar logarítmicamente los datos de cuantificación de los lípidos una vez normalizados, dado que en ocasiones presentaban distribuciones bastante asimétricas o con valores extremos. En el caso en el que los lípidos estaban repetidos en 3 ocasiones, se realizó un gráfico para cada una de las 3 combinaciones posibles de pares de lípidos con el mismo nombre.

3.4. Normalización de los datos ómicos

La estimación de la expresión proporcionada por la tecnología miRNA-Seq puede contener sesgos que provocan que los valores de expresión no sean comparables entre las distintas muestras biológicas estudiadas (sesgos entre muestras, *between-samples*) o entre genes (sesgos intra-muestras, *within-samples*). Algunos de los sesgos debidos a la tecnología que se pueden encontrar en datos de miRNA-Seq son la profundidad de secuenciación o el contenido en GC de la secuencia del miRNA. Esto provoca que sea crítico normalizar los datos de miRNA-Seq para que los valores sean comparables (Garmire & Subramaniam, 2012).

Para evaluar la existencia del sesgo de contenido en GC se realizó un gráfico de *loadings* de PCA coloreados por su porcentaje de contenido en GC. Para ello, se calculó el contenido de GC de cada uno de los miRNAs analizados. Esto se realizó extrayendo la secuencia de cada miRNA del fichero hsa.fa del transcriptoma humano y calculando el porcentaje de bases que correspondían a G o C en el total de la secuencia. De esta manera, se procedió a representar el gráfico de loadings coloreados por contenido en GC.

Por lo tanto, los datos filtrados por CPM fueron normalizados por distintos métodos. En primer lugar, se compararon cuatro métodos de normalización: métodos UQ (*Upper-Quartile normalization*) y TMM (*Trimmed Mean of M-values*), ambos del paquete de R **NOISeq**; QN (*Quantile normalization*) del paquete de R **limma** (Ritchie et al., 2015), y CQN (*Conditional Quantile Normalization*) del paquete de R **cqn** (Hansen et al., 2012), que corrige el sesgo de contenido en GC. En el método UQ, los valores de expresión en cada muestra se dividen por el cuartil superior de todos los valores de dicha muestra. QN, iguala los percentiles de cada muestra para que tengan distribuciones idénticas. TMM se asemeja al método QN, pero iguala la distribución de los valores centrales de cada muestra, dejando libertad a los valores más extremos. CQN combina la normalización QN con modelos de regresión para estimar y corregir el sesgo de contenido en GC.

Los datos de lipidómica ya estaban pre-procesados, es decir, se disponía de la concentración de lípidos en cada muestra. No obstante, tras el control de calidad pertinente, se decidió normalizar mediante el método QN del paquete de R **limma**, tras unir los datos de ESI+ y ESI- en una sola matriz.

3.5. Identificación de variables ómicas con cambios entre grupos.

Una vez cuantificada y normalizada la expresión de miRNAs y la normalizados los datos de lípidos en las muestras analizadas, los valores ya son comparables entre sí y, por lo tanto, se continuó realizando un análisis estadístico de expresión diferencial para comparar las medias de expresión/concentración de miRNAs/lípidos entre los grupos de prematuros y a término. Para ello, se aplicó el paquete R **limma**. El análisis de expresión (concentración) diferencial que utiliza este paquete de R consiste en aplicar modelos lineales, equivalentes a una Prueba t de Student o t-test, en el caso de comparación de medias de expresión entre dos grupos, como es nuestro caso. Dado que el tamaño muestral es reducido en este tipo de experimentos, **limma** aplica un procedimiento bayesiano para estimar la varianza de las variables ómicas en cada grupo comparado de forma más robusta, teniendo en cuenta la distribución de dicha varianza en todas las variables ómicas analizadas. El paquete **limma** fue diseñado inicialmente para trabajar con datos obtenidos a partir de microarrays, que se asume que siguen una distribución de probabilidad normal, necesaria para aplicar un t-test. Dado que los datos de miRNAs o de lipidómica no tienen por qué seguir esta distribución, se transformaron con la función voom() para aproximar su distribución a la distribución normal. La estrategia voom aplica una transformación logarítmica a los

datos de conteos corregidos por la profundidad de secuenciación. Con estos datos transformados, voom estima la relación entre la media y la varianza para cada gen (o variable ómica) mediante modelos no paramétricos, y la utiliza para otorgar distintos pesos a cada observación según su precisión. Estos pesos se incorporan a los modelos lineales de limma para reducir la heterocedasticidad, tal y como se requiere en un modelo lineal (Law et al., 2014). Se consideraron diferencialmente expresados a aquellos miRNAs o lípidos con un P-valor < 0,05.

3.6. Análisis de enriquecimiento de genes diana en miRNAs diferencialmente expresados

Dada una lista de miRNAs, es común predecir qué genes están regulados por dichos miRNAs, para después estudiar en qué ruta biológica están implicados. Sin embargo, la asociación entre miRNAs y genes diana no es sencilla. Un sólo miRNA puede regular a muchos genes diferentes y un gen, a su vez, puede estar regulado por numerosos miRNAs distintos. Esto significa que la alteración de un solo miRNA puede afectar a numerosas funciones biológicas (Bleazard et al., 2015). Para disponer de la lista de genes diana predichos para cada miRNA analizado en nuestro estudio, se utilizaron las bases de datos del paquete de R **multiMiR** en las que las interacciones miRNA-diana habían sido validadas experimentalmente: miRecords, miRTarBase, and TarBase (Ru et al., 2014).

Tras el análisis de expresión diferencial obtenemos una lista de aquellos miRNAs que se encuentran alterados entre ambos grupos estudiados. Se aplicó un test de independencia de Fisher a cada gen diana, para estudiar cuáles de ellos se encontraban sobre-representados en dicho conjunto de miRNAs diferencialmente expresados. Para aplicar este análisis de enriquecimiento se elaboró un *script* de R con el siguiente funcionamiento. Se genera una tabla de contingencia de dimensiones 2x2 para cada uno de los posibles genes diana, en la que se representa el número de miRNAs diferencialmente expresados y no diferencialmente expresados, y el número de estos miRNAs que para los que el gen analizado es diana. Se consideró que un gen diana estaba significativamente sobrerrepresentado en el conjunto de miRNAs diferencialmente expresados cuando se obtenía un p-valor ajustado por el método de Benjamini y Hochberg (o FDR) inferior a 0,05. Se ajustó el p-valor para corregir el problema de tests múltiples realizados por cada gen diana, para reducir la probabilidad de error global del test.

3.7. Análisis de enriquecimiento funcional de genes diana

Una vez obtenidos los genes diana enriquecidos en los miRNAs diferencialmente expresados, es interesante estudiar en qué rutas biológicas participan de forma mayoritaria y, de esta manera, detectar cuáles de ellas se encuentran alteradas entre los grupos estudiados. Para ello, es necesario disponer de la anotación funcional, es decir, de las rutas en las que participa cada gen diana. En este trabajo, se utilizó la base de datos de la Gene Ontology (GO), que se descargó con la herramienta BioMart de Ensembl (Howe et al., 2021), y en la que se encontraban los genes humanos y los identificadores GO de las rutas en las que se encontraban involucrados. La GO es una ontología compuesta por más de 38000 definiciones precisas denominadas “términos GO” que describen las acciones moleculares de los productos de los genes, los procesos biológicos en los que tienen lugar estas acciones y la localización celular en la que están presentes (Balakrishnan et al., 2013).

El análisis de enriquecimiento funcional se llevó a cabo mediante el mismo tipo de test utilizado en el enriquecimiento de genes diana, es decir, el test exacto de independencia de Fisher. En este caso, se calculaba una tabla de contingencia para cada una de las rutas biológicas, evaluando la relación de los

genes diana y no diana con ellas. Para ello, se utilizó el mismo *script* de R que en la sección anterior, pero modificando los ficheros *input*, trabajando en este caso con los genes diana y la base de datos con la relación de estos con los términos GO. Se consideraron GOs significativamente enriquecidos en los genes diana, aquellos con p-valor ajustado por el método de Benjamini y Hochberg (o FDR) inferior a 0,05.

Para cada uno de los términos GOs más significativos, se identificaron los genes diana anotados a dicho GO. Para entender mejor la relación entre dichos genes, puesto que la GO no proporciona esta información, se utilizó la herramienta **KEGG Mapper** (Minoru Kanehisa & Sato, 2020) de la base de datos KEGG (M. Kanehisa, 2000), que permite obtener el diagrama de la ruta biológica en cuestión a partir del conjunto de genes diana proporcionados. Para ello, se elaboró un *script* de R a partir del cual se extraían los genes diana implicados en cada una de las rutas enriquecidas, y se traducía el nombre de dichos genes a su término KEGG a partir de la base de datos extraída de la KEGG API *(Application Programming Interface)*.

3.8. Análisis de enriquecimiento funcional de lípidos alterados entre grupos

Para el análisis de enriquecimiento funcional de los lípidos, se utilizó la herramienta web Lipid Ontology (**LION**). **LION** es una herramienta bioinformática específica para lipidómica que permite realizar enriquecimientos funcionales en conjuntos de datos de lípidos. **LION** también proporciona una clasificación de los lípidos por lo que se conoce como términos LION. Los términos LION son identificadores que relacionan los lípidos con descripciones químicas, características biológicas o propiedades biofísicas. De esta manera, se llevó a cabo el análisis *"Ranking mode"* de la herramienta **LION**. En este análisis, todas las especies individuales de lípidos de 2 condiciones son comparadas y clasificadas mediante un valor numérico que asocie los lípidos con dichas condiciones. En este caso, se utilizó el log2 del fold-change como valor numérico. Posteriormente, las distribuciones de todos los términos LION asociados en la lista clasificada de los lípidos, se comparan con distribuciones uniformes mediante el uso de pruebas de Kolmogorov-Smirnov, calificando como enriquecido a los términos LION cuyos lípidos asociados están mejor clasificados de lo esperado por azar (Molenaar et al., 2019). En este análisis, entre los parámetros utilizados se seleccionó la dirección de la clasificación de lípidos de mayor a menor valor numérico, comparando contra la hipótesis alternativa de dos colas y separando los lípidos *upregulated* y *downregulated* en el gráfico de barras de los resultados.

3.9. Recursos computacionales

3.9.1. Soporte informático del CIPF

A partir de los datos facilitados, todo el proyecto se realizó mediante la computadora personal (PC) del autor. En el estudio se contó con un fichero correspondiente a la secuenciación de cada una de las muestras de miRNA-Seq con un tamaño variable entre 1 y 2 GB, ocupando un total de 16,1 GB. El pre-procesado de esta cantidad de datos resulta difícil de manejar con un ordenador común, por lo que fue necesario el uso del clúster informático del Centro de Investigación Príncipe Felipe (CIPF). Un clúster es un conjunto de ordenadores que están conectados entre ellos en una red de alta velocidad actuando como si se tratara de un único ordenador de gran potencia. Esta combinación de un gran número de ordenadores produce un alto rendimiento computacional y ofrece la capacidad de almacenar mucha memoria, entre otras prestaciones. Actualmente, el clúster del CIPF está compuesto por 44 nodos, con hasta 660 CPUs y una memoria RAM acumulada de 11 TeraBytes. Además, utiliza

un sistema de almacenamiento de archivos distribuidos llamado Lustre, que cuenta con una capacidad de almacenamiento de 1 PetaByte.

Los procesos pertinentes fueron ejecutados tras el lanzamiento de *scripts* escritos en el lenguaje de programación Bash en el sistema de colas del clúster. El sistema de colas es la forma de gestionar los recursos del clúster, asignando los trabajos a las diferentes máquinas, dando diferente prioridad a los procesos de ejecución, etc. En algunos casos se prepararon un tipo especial de *scripts* para lanzar varios procesos y que fueran ejecutados de forma paralela.

3.9.2. R y R Studio

Una vez finalizado el pre-procesamiento de los datos de miRNA-Seq, la herramienta R Studio con la versión R 4.0.3 (RSTUDIO, 2020) y el lenguaje de programación R pasaron a un primer plano. Esta herramienta se utilizó en un primer momento para elaborar las matrices de cuantificación de miRNAs y lípidos y, desde ese punto, para realizar todo el conjunto de análisis del estudio, a excepción del uso de las herramientas destacadas a lo largo del apartado de Materiales y Métodos. Asimismo, se utilizaron paquetes de R mencionados en cada uno de los apartados para los que fue preciso su uso. Además, los *scripts* utilizados para los análisis del estudio están disponibles en la siguiente carpeta (https://drive.google.com/drive/folders/1cyzDfVEUAt1UzfNCkWR_2-3yi4SBfmrY?usp=sharing).

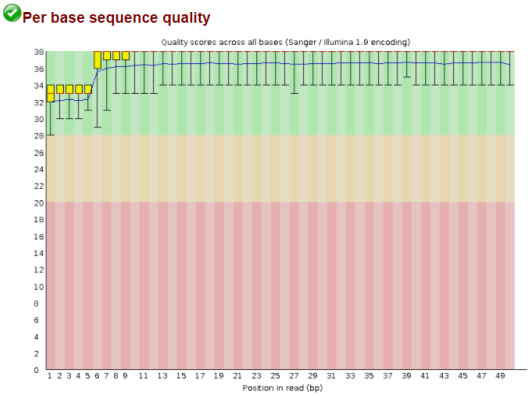
4. Resultados y discusión

4.1. Control de calidad de los datos de miRNA-Seq y eliminación de adaptadores.

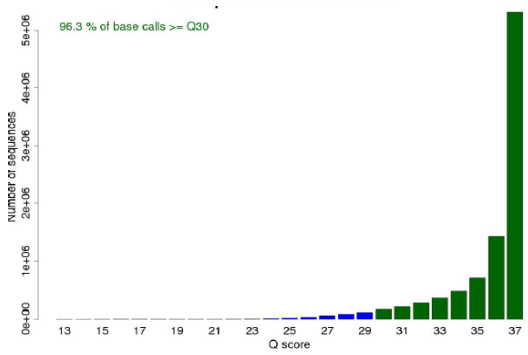
Los controles de calidad de los datos de transcriptómica realizados en este trabajo tenían el objetivo de evaluar las características de las lecturas tras la secuenciación y tras la eliminación de los adaptadores utilizados en dicha secuenciación. Dicho control de calidad se llevó a cabo mediante la herramienta **FastQC**.

En el control de calidad inicial de las lecturas de secuenciación de miRNAs se obtuvieron resultados que mostraban un buen estado de estas, en rasgos generales. Las figuras mostradas en este apartado son un ejemplo de los resultados obtenidos para la muestra "Término 1", siendo análogos para el resto de las muestras. La calidad de secuenciación se situaba en un rango óptimo en todas las muestras, con valores de *score* por encima de 28, de un rango entre 0 y 38, en todas las posiciones de la lectura. Esto también se apreciaba al observar el Q score, valor que estima la probabilidad de que una base de la lectura no sea correcta, obteniendo valores altos de Q score cuando hay menor probabilidad de error. Este valor se sitúa en cifras superiores a 30 (Figura 5-A y 5-B) en la gran mayoría de las secuencias. El contenido de GC se aproximaba a una distribución normal de la forma esperada (Figura 5-C), no se encontraban bases no determinadas (N), todas las lecturas tenían una longitud esperada, 50 bp, como resultado del tipo de secuenciación que se había llevado a cabo; se apreciaba la presencia de adaptadores y no se encontraban lecturas altamente sobrerrepresentadas a excepción de las correspondientes a dichos adaptadores. Tan solo se observó en el gráfico de calidad de secuencia *per tile*, 3 posiciones en el *flow cell* (celda de secuenciación), en las que la calidad de secuenciación disminuía en relación a otras zonas para la misma posición de base (Figura 5-D). Este gráfico era similar en todas las muestras y tan solo se obtenía menor calidad en esas 3 posiciones, por lo que no se consideró que tuviera repercusión en el estudio.

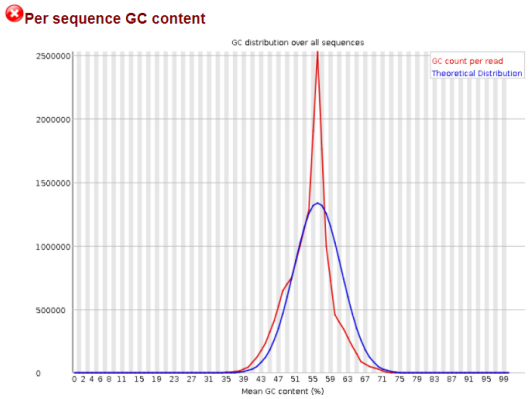
(A)



(B)



(C)



(D)

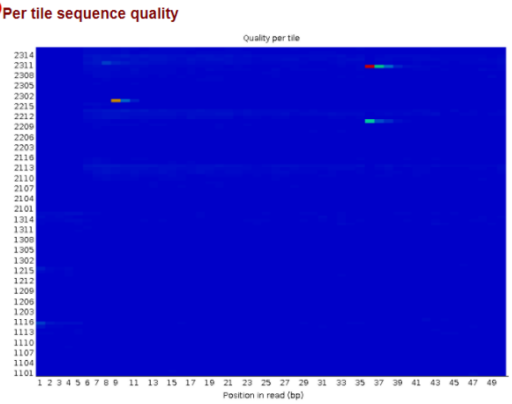


Figura 5. Gráficos obtenidos para la muestra "Término 1" en el control de calidad previo a la eliminación de adaptadores. **(A)** Gráfico de calidad de secuencia por base. En el eje X la posición de la base en la lectura; en el eje Y la calidad de la posición. **(B)** Gráfico de Q-score. En el eje X el valor de Q-score; en el eje Y el número de secuencias para los valores de Q-score. **(C)** Gráfico de contenido en GC por secuencia. En el eje X el porcentaje medio de contenido de GC; en el eje Y el número de secuencias para un determinado porcentaje de GC. La línea azul representa la distribución normal esperada de los datos; la línea roja representa la distribución de la muestra evaluada. **(D)** Gráfico de calidad de secuencia *per tile*. En el eje X la posición de la lectura en bp; en el eje Y la zona del *flow cell*. El gráfico muestra la desviación de la calidad media para cada zona del *flow cell*. Los colores fríos muestran posiciones en las que la calidad de la secuencia se sitúa en la media o sobre ella; los colores más cálidos indican que una zona tiene peor calidad de secuenciación que el resto para esa posición (BABRAHAM BIOINFORMATICS, 2019).

El control de calidad posterior a la eliminación de los adaptadores de las lecturas mostró igualmente un buen estado de estas, a pesar de algunas variaciones. Las figuras mostradas en este apartado son también un ejemplo de los resultados obtenidos para la muestra "Término 1, siendo análogos para el resto de las muestras. La calidad de secuencia por posición se mantuvo de forma general en valores por encima de 28, en el rango de máxima calidad. No obstante, dicha calidad disminuyó en las posiciones finales alcanzando un valor de 26 en las secuencias de 2 de las 10 muestras testadas. El contenido de GC de las secuencias continuó aproximándose a una distribución normal, al igual que en el control de calidad previo. El contenido de bases no determinadas en las secuencias se mantuvo en 0. En cuanto a la calidad de la secuencia *per tile*, las 3 posiciones con menor calidad comentadas anteriormente se seguían manteniendo. Adicionalmente, se aprecia una leve disminución de calidad en diferentes zonas del *flow cell* para las últimas posiciones de las lecturas (Figura 6-A). Este fenómeno se observó en todas las muestras analizadas y se explica por la propia eliminación de los adaptadores. Esta eliminación provoca que las últimas bases de las lecturas no estén presentes, variando sus valores de longitud y de calidad según la posición del adaptador y el corte realizado en dichas secuencias por el software. Por ello, es frecuente observar una disminución de la calidad de secuenciación en las bases del extremo 3' de las lecturas.

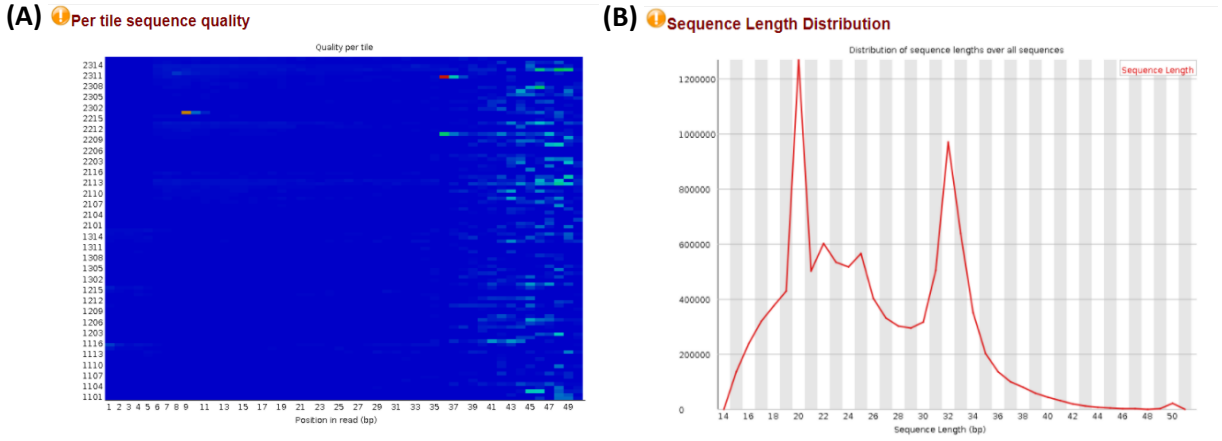


Figura 6. Gráficos obtenidos para la muestra "Término 1" en el control de calidad posterior a la eliminación de adaptadores. **(A)** Gráfico de calidad de secuencia *per tile*. Se observa una leve reducción de calidad en las posiciones finales de la secuencia con respecto al gráfico previo a la eliminación de adaptadores. **(B)** Gráfico de distribución de longitud de secuencia. En el eje X la longitud de secuencia en pares de bases (bp); en el eje Y el número de secuencias para una determinada longitud de secuencia.

La distribución de la longitud de las secuencias varió con respecto a los resultados del control de calidad previo debido a la eliminación de los adaptadores y la consecuente disminución de longitud de la lectura. Esta distribución alcanzaba un mayor número de secuencias en valores de alrededor de 20 pb, lo que equivale al tamaño normal de los miRNAs. Sin embargo, en varias muestras se observó un pico de 32 pb (Figura 6-B). De esta manera, se procedió a examinar la procedencia de dicho pico. Se buscaron las secuencias sobrerrepresentadas con la misma longitud que mostraba el pico y se utilizó la herramienta **BLAST** *(The Basic Local Alignment Search Tool)* de la web oficial del NCBI *(National Center for Biotechnology Information)* (Boratyn et al., 2013). Esta herramienta permite encontrar regiones de similaridad local entre diferentes secuencias y, en el caso de **BLASTn**, lo consigue mediante la comparación de las secuencias nucleotídicas.

En cuanto a los parámetros utilizados en la búsqueda, se introdujo en **BLASTn** la secuencia de interés y se buscaron resultados que coincidiera con esta en la base de datos "Nucleotide collection (nr/nt)" para el organismo "human (taxid:9606)". Adicionalmente, se seleccionó el programa Megablast para optimizar por secuencias altamente similares. Se obtuvieron como resultados un gran número de coincidencias con un 100% de identidad, con similar puntuación y mismo E-value. Esto se puede deber al corto tamaño de la secuencia analizada y, por tanto, la baja especificidad de esta. Entre las secuencias que producían los alineamientos de mayor significancia, cabe resaltar que muchos corresponden a clones BAC. Sin embargo, no se pudo determinar con certeza a qué secuencias corresponden estos alineamientos.

Por otro lado, la eliminación de los adaptadores fue exitosa, como muestra la gráfica de porcentaje de contenido de adaptador a lo largo de la secuencia, que tras la eliminación cambió a un valor de 0 para todas las posiciones en todas las muestras (Figura 7). Además, en la tabla de secuencias sobrerrepresentadas ya no se encontraba ninguna que perteneciera a esta fuente, hecho que sí se observaba en los resultados previos a la eliminación.

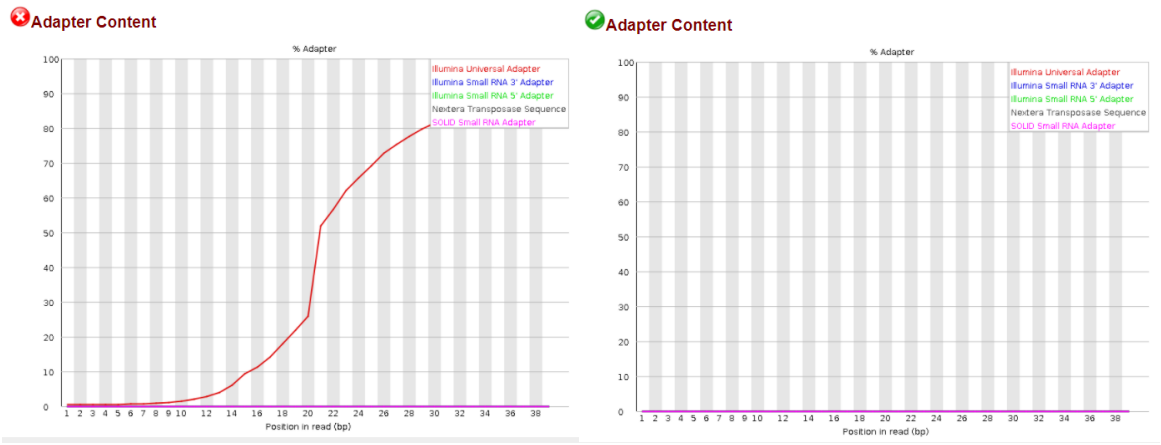


Figura 7. Gráficos de contenido de adaptador para la muestra "Término 1". En la izquierda el resultado tras el control de calidad, previo a la eliminación de adaptadores. En la derecha el resultado tras el control de calidad, posterior a la eliminación de adaptadores. En el eje X la posición de la base en la lectura en pares de bases (bp); en el eje Y el porcentaje de presencia de adaptador. La línea roja hace referencia al adaptador de Illumina utilizado para el proceso de secuenciación de las muestras. En el gráfico previo al paso de eliminación de adaptadores se observa la presencia de este, mientras que en el gráfico posterior se mantiene en 0% a lo largo de todas las posiciones de la secuencia.

La eliminación de adaptadores se llevó a cabo con dos parámetros diferentes y, por lo tanto, se procedió a comparar los resultados de ambos análisis mostrados en la Tabla 3. Cuando se fijó un valor

de longitud mínima de 15 nucleótidos, las lecturas que se eliminaban en el proceso rondaban el 10% a lo largo de las diferentes muestras, llegando a eliminar el 21,5% en una de ellas. Por otro lado, con un valor menos restrictivo de 10 para el mismo parámetro, se perdían alrededor del 2% de las muestras, alcanzando un 3,3% en la muestra con más lecturas eliminadas.

		Resultados tras aplicar cutadapt -q 30 -l 15		Resultados tras aplicar cutadapt -q 30 -l 10	
Muestra	Lecturas de partida	Lecturas eliminadas cutadapt	Lecturas tras cutadapt	Lecturas eliminadas cutadapt	Lecturas tras cutadapt
Término 1	9.255.941	979.441 (10,6%)	8.276.500 (89,4%)	220.068 (2,4%)	9.035.873 (97,6%)
Término 2	13.393.757	1.121.675 (8,4%)	12.272.082 (91,6%)	248.010 (1,9%)	13.145.747 (98,1%)
Término 3	8.148.215	1.755.641 (21,5%)	6.392.574 (78,5%)	268.547 (3,3%)	7.879.668 (96,7%)
Término 4	11.975.282	1.198.420 (10,0%)	10.776.862 (90,0%)	154.799 (1,3%)	11.820.483 (98,7%)
Término 5	10.414.733	1.113.272 (10,7%)	9.301.461 (89,3%)	178.873 (1,7%)	10.235.860 (98,3%)
Pretérmino 1	12.146.293	515.391 (4,2%)	11.630.902 (95,8%)	187.795 (1,5%)	11.958.498 (98,5%)
Pretérmino 2	10.746.113	383.878 (3,6%)	10.362.235 (96,4%)	131.423 (1,2%)	10.614.690 (98,8%)
Pretérmino 3	11.447.548	520.786 (4,5%)	10.926.762 (95,5%)	139.874 (1,2%)	11.307.674 (98,8%)
Pretérmino 4	10.087.235	750.828 (7,4%)	9.336.407 (92,6%)	95.507 (0,9%)	9.991.728 (99,1%)
Pretérmino 5	11.501.036	1.384.159 (12,0%)	10.116.877 (88,0%)	380.854 (3,3%)	11.120.182 (96,7%)
				Conjunto de datos seleccionado	

Tabla 3. Resultados del proceso de eliminación de adaptadores. Cada fila corresponde a los valores de una de las 10 muestras. Las dos últimas columnas corresponden al resultado obtenido tras aplicar el software **Cutadapt** con los parámetros seleccionados para continuar el análisis.

Tras analizar los resultados de eliminación de los adaptadores, se consideró que el porcentaje de secuencias que se perdían al marcar el parámetro de longitud en un valor de 15, alrededor del 10%, era excesivamente elevado. Por ello, se continuó el estudio con las secuencias resultantes de la eliminación de adaptadores con el parámetro de longitud marcado en un valor de 10 nucleótidos.

4.2. Mapeo de las lecturas de miRNA-Seq y filtro de calidad de mapeo

Tras realizar el mapeo con **Bowtie2**, se obtuvieron los resultados resumidos en la Tabla 4. En ella se aprecian las lecturas iniciales de partida y el ratio de lecturas que han conseguido ser alineadas (en porcentaje). Dicho ratio de alineamiento corresponde tanto a aquellas lecturas que han alineado una vez contra el transcriptoma, como aquellas que lo han hecho en varios sitios del mismo *(multi-mapping)*, contando estas últimas como un mismo alineamiento en el ratio total de alineamientos. Este ratio total varía de unas muestras a otras, pero se mantiene en cifras entre un 3 y un 11%, con una media cercana al 7% de lecturas alineadas por muestra. Aunque es un porcentaje bajo de alineamiento debido probablemente a la presencia de lecturas que corresponden a otros tipos de small RNA, se verá más adelante que fue suficiente para obtener una cuantificación adecuada de la expresión de los miRNAs.

bowtie2					
Muestra	Lecturas tras cutadapt	Ratio total de alineamiento	Lecturas alineadas 1 vez	Lecturas alineadas >1 vez	Total lecturas alineadas
Término 1	9,035,873	5.04%	339,519 (3.76%)	116,310 (1.29%)	455.829
Término 2	13,145,747	6.00%	550,650 (4.19%)	237,805 (1.81%)	788.455
Término 3	7,879,668	8.99%	493,491 (6.26%)	215,007 (2.73%)	708.498
Término 4	11,820,483	9.17%	808061 (6.84%)	276,333 (2.34%)	1.084.394
Término 5	10,235,860	6.15%	442,626 (4.32%)	186,813 (1.83%)	629.439
Pretérmino 1	11,958,498	11.17%	1,010,839 (8.45%)	324,815 (2.72%)	1.335.654
Pretérmino 2	10,614,690	3.16%	231,976 (2.19%)	103,587 (0.98%)	335.563
Pretérmino 3	11,307,674	5.41%	428,301 (3.79%)	183,503 (1.62%)	611.804
Pretérmino 4	9,991,728	7.77%	618,655 (6.19%)	157,827 (1.58%)	776.482
Pretérmino 5	11,120,182	3.95%	242,302 (2.18%)	196,911 (1.77%)	439.213

Tabla 4. Resultados del mapeo de las lecturas contra el transcriptoma humano. Cada fila corresponde a los valores de una de las 10 muestras. En la segunda columna se observan los valores correspondientes al número de lecturas con las que se realiza el mapeo, mientras que en la última columna se observa el número total de las lecturas que lograron alinear con el transcriptoma y en la tercera columna se observa la relación de ambas cifras, es decir, el porcentaje de lecturas que han mapeado. La cuarta y quinta columna corresponden al número de lecturas, dentro del total de alineadas, que han alineado con el transcriptoma una vez o más de una vez como resultado del *multi-mapping* (múltiple mapeo de una misma lectura con varias partes del transcriptoma).

Los alineamientos producidos en el mapeo fueron sometidos a distintos filtros de calidad a partir de los valores del parámetro MAPQ, y los resultados aparecen representados en la Tabla 5. Este filtro, además de eliminar los alineamientos de baja calidad, también reduce el ruido en los datos porque elimina casos de *multi-mapping* que serían incorrectos al asignar la misma lectura a varios miRNAs.

Muestra	Alineamientos tras el filtro de calidad MAPQ									
	MAPQ1		MAPQ2		MAPQ3		MAPQ4/5		MAPQ10/20	
Término 1	371.926	82%	362.751	80%	362.729	80%	357.378	78%	353.364	78%
Término 2	662.086	84%	648.654	82%	648.608	82%	640.447	81%	634.062	80%
Término 3	489.217	69%	477.509	67%	477.486	67%	465.509	66%	452.692	64%
Término 4	1.033.069	95%	1.008.825	93%	1.008.744	93%	1.005.531	93%	1.002.296	92%
Término 5	513.774	82%	497.949	79%	497.919	79%	491.064	78%	484.846	77%
Pretérmino 1	1.290.490	97%	1.264.544	95%	1.264.466	95%	1.262.396	95%	1.260.089	94%
Pretérmino 2	308.330	92%	302.122	90%	302.109	90%	299.986	89%	297.895	89%
Pretérmino 3	579.705	95%	569.271	93%	569.241	93%	567.512	93%	565.837	92%
Pretérmino 4	755.522	97%	743.779	96%	743.730	96%	742.805	96%	741.733	96%
Pretérmino 5	212.346	48%	207.253	47%	207.247	47%	195.531	45%	185.013	42%

Tabla 5. Resultados de los filtros de calidad de mapeo MAPQ. En cada fila se muestran los resultados para cada una de las diferentes muestras. En esta tabla se puede apreciar el número de lecturas que han logrado pasar el filtro para cada muestra y su porcentaje respecto al número de alineamientos totales previo al filtro. En la tabla se puede observar el agrupamiento de dos valores de MAPQ en las dos últimas columnas. Esto se debe a que los resultados obtenidos con cada uno de los filtros obtenidos fueron similares.

Al aplicar diferentes filtros de mapeos, se pretende eliminar aquellos alineamientos de baja calidad, pero sin reducir bruscamente el número de los alineamientos totales. Por ello, una vez analizados los resultados, se consideró continuar con el estudio siguiendo dos vías de forma paralela: una de ellas

con las lecturas tras aplicar el filtro de MAPQ > 1 y otra con el filtro de MAPQ > 5, con las que se procedió a cuantificar la expresión de los miRNAs. Se decidió continuar con los resultados obtenidos mediante estos filtros para contar, por un lado, con un conjunto de valores con una mayor cantidad de alineamientos totales tras superar un filtro de calidad mínimo (MAPQ1), y otro conjunto de valores con un menor número de alineamientos totales, sin alcanzar valores extremadamente bajos, aunque tras superar todos ellos un filtro de calidad más restrictivo (MAPQ5). De esta manera, se podría evaluar en análisis posteriores con cuál de los dos conjuntos de valores se obtienen mejores resultados. Además, se ha observado en trabajos previos el uso de estos dos valores del parámetro en estudios similares.

4.3. Cuantificación de la expresión de miRNAs

Mediante el uso del paquete **dplyr** de R se logró transferir la información correspondiente al número de lecturas mapeadas que superaron los filtros de calidad MAPQ hasta R Studio para poder continuar con el análisis, obteniendo finalmente una matriz de valores de expresión formada por 10 columnas, correspondientes a cada una de las muestras, y 2607 o 2454 filas, en el caso de datos obtenidos tras el filtro de MAPQ1 o MAPQ5 respectivamente, correspondientes a cada uno de los miRNAs detectados en al menos una de las muestras secuenciadas. A aquellos miRNAs no detectados en alguna muestra, se les adjudicó un valor de 0 en dicha muestra. El número de miRNAs diferentes de los que se obtuvieron conteos en cada una de las muestras se puede observar en la Tabla 6, en la que se detecta mayor cantidad en el caso de aplicar el filtro MAPQ1 debido a su carácter menos restrictivo.

Muestra	miRNAs diferentes	
	MAPQ1	MAPQ5
Término 1	1774	1431
Término 2	2119	1794
Término 3	2025	1710
Término 4	1824	1511
Término 5	2099	1774
Pretérmino 1	1879	1559
Pretérmino 2	1466	1177
Pretérmino 3	1397	1154
Pretérmino 4	1005	825
Pretérmino 5	1793	1490
Totales	2607	2454

Tabla 6. Número de miRNAs diferentes de los que se obtuvieron conteos para cada muestra. Cada fila muestra los resultados para una de las 10 muestras, mientras que la última fila muestra el número de miRNAs diferentes en el total de las muestras.

Para comprobar el efecto de la aplicación de los filtros de MAPQ1 y MAPQ5 sobre las matrices de expresión y la relación entre ambos conjuntos de valores, se procedió a evaluar la correlación entre las 10 muestras tras cada filtro utilizando los miRNAs comunes a ambos filtros MAPQ. Para ello, se normalizaron previamente los datos para evitar el sesgo de la profundidad de secuenciación mediante la función rpkm() del paquete de R **NOISeq**. Los valores de correlación obtenidos fueron, en todos los casos, superiores a 0.999, por lo que podemos concluir que los valores procedentes de la aplicación

de ambos filtros estaban muy correlacionados, indicando que el filtro de calidad no modifica sustancialmente la distribución de los valores de expresión, más allá de que dejen de detectarse algunos miRNAs con baja expresión al ser más exigentes con el filtro de calidad.

Por otro lado, se exploraron ambos conjuntos de datos mediante la aplicación de PCAs. Para ello, se realizaron los gráficos de PCAs de *scores* que se muestran en la Figura 8. Estos gráficos no solo se utilizaron para observar la distribución de las observaciones, sino también para compararlos con otros gráficos análogos obtenidos en pasos posteriores. En la Figura 8 se observa una agrupación similar de las muestras para ambos filtros. La separación de las muestras en los grupos "Término" y "Pretérmino" se establece principalmente en relación a la segunda componente principal (PC2), que representa tan solo un 10 y un 11% de la variabilidad total de los datos, mientras que en la primera componente principal (PC1) recae un 50 y un 47% de la variabilidad.

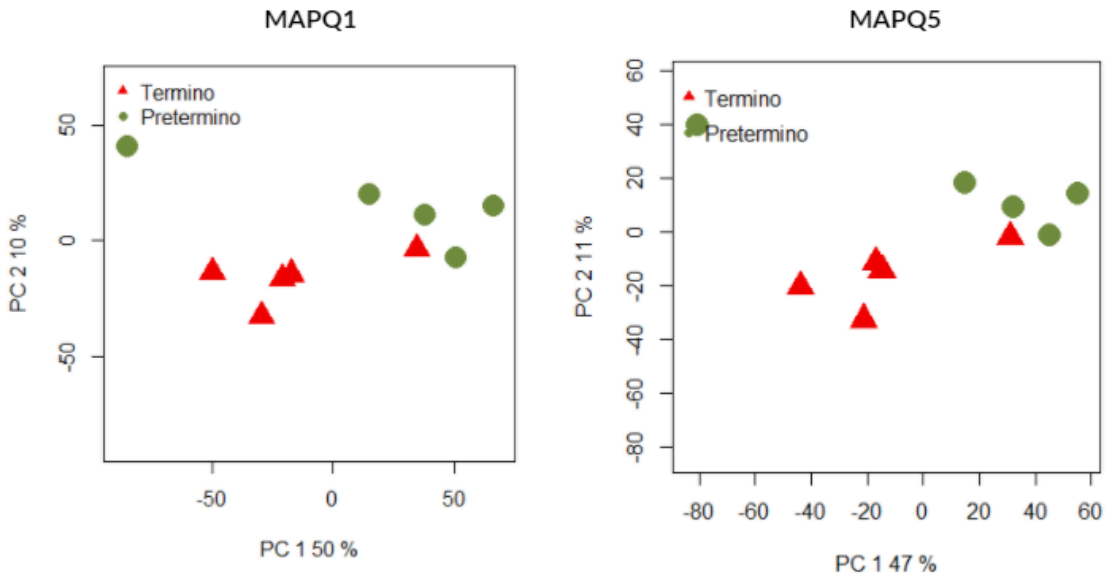


Figura 8. Gráfico de *scores* de PCA tras los filtros MAPQ1 y MAPQ5. Se observa una mayor separación entre los grupos estudiados mediante la segunda componente principal (PC2), siendo similares ambas distribuciones.

4.4. Filtro de baja expresión CPM

A las matrices de expresión obtenidas tras los filtros de MAPQ1 y MAPQ5 se les aplicó el filtro de baja expresión (CPM) con 7 valores diferentes obteniendo los resultados mostrados en la Tabla 7. Con este filtro, se pretendía eliminar aquellos miRNAs que estuvieran bajamente expresados en ambos grupos y que, por lo tanto, no serían útiles en el estudio e incrementarían el ruido de fondo de los datos. Se decidió seleccionar los datos de MAPQ1 tras aplicar el filtro de CPM > 3 y los datos de MAPQ5 tras aplicar el filtro de CPM > 2. En ambos casos, el objetivo era reducir el número de miRNAs "ruidosos" o irrelevantes para el análisis por su baja expresión, pero siempre teniendo en cuenta que el número de miRNAs restantes fuera suficiente para que este paso no conllevara una pérdida de información relevante en el estudio.

Filtro (c.p.m)	miRNAs de partida: 2607 (MAPQ1)		miRNAs de partida: 2454 (MAPQ5)		Variación
cpm	miRNAs restantes	% miRNAs restantes	miRNAs restantes	% miRNAs restantes	
0	1600	**61%**	1212	**49%**	**76%**
0,5	1600	**61%**	1212	**49%**	**76%**
1	1591	**61%**	1204	**49%**	**76%**
2	1437	**55%**	**1042**	**42%**	**73%**
3	**1238**	**47%**	876	**36%**	**71%**
4	1060	**41%**	761	**31%**	**72%**
5	934	**36%**	662	**27%**	**71%**

Tabla 7. Resultados de la aplicación de los filtros de baja expresión. Se muestran los resultados de la aplicación de 7 valores diferentes de filtro CPM en los datos del filtro MAPQ1 y del filtro MAPQ5. Las casillas resaltadas con el fondo morado corresponden al número de miRNAs restantes obtenidos con el filtro CPM seleccionado para continuar el estudio con cada conjunto de datos.

4.5. Evaluación de sesgos y normalización de los datos ómicos

Con los dos conjuntos de datos seleccionados tras los filtros de CPM se continuó el estudio evaluando la posible existencia de sesgos en los datos introducidos por la tecnología de secuenciación, además del sesgo debido a la distinta profundidad de secuenciación en cada una de las muestras analizadas.

En miRNA-Seq, cabe valorar si hay un sesgo debido al contenido en GC en las secuencias de los miRNAs. Para ello, se aplicó un PCA a los datos para ver si se detectaba un efecto importante del contenido en GC en la expresión de los miRNAs la Figura 9 muestra los gráficos de *loadings* del PCA antes y después de corregir el sesgo de contenido en GC, con los miRNAs representados en las dos primeras componentes principales, que recogen las principales fuentes de variabilidad de los datos. Los miRNAs se han coloreado según su contenido en GC y, en el gráfico de la izquierda, se observa una ligera agrupación de los miRNAs con alto contenido en GC (alrededor del 75%). No obstante, este efecto es leve y, además, no es el esperado, ya que, si existe sesgo por GC, la expresión de los miRNAs con alto o bajo contenido en GC tiende a ser más baja que la de los miRNAs con contenido en GC alrededor del 50%, cuya expresión tendería a ser más alta. Aun así, se decidió aplicar el método de normalización CQN (Figura 9, derecha), que permite corregir este sesgo. Como se puede observar, los miRNAs con alto contenido en GC siguen estando agrupados en una zona del gráfico por lo que no parece que el método CQN haya conseguido corregir bien este leve sesgo. No obstante, resultados de aplicar este método de normalización se compararán a continuación con los de los otros métodos de normalización que no tienen en cuenta este sesgo.

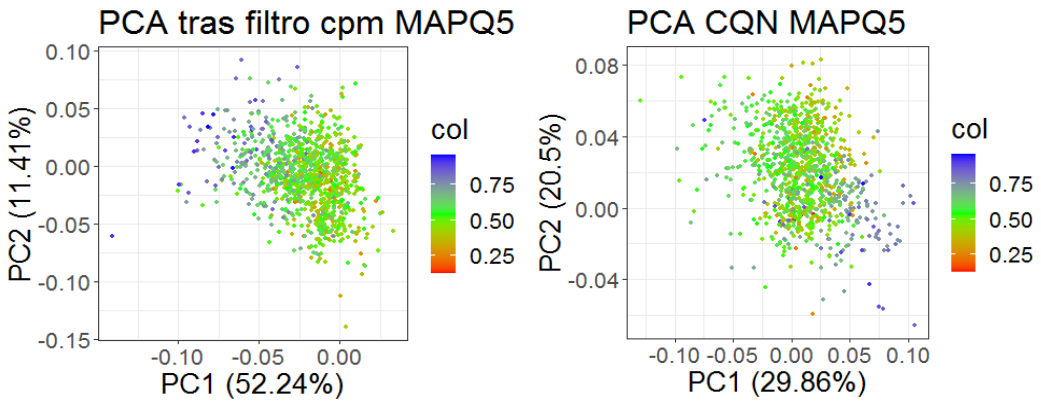


Figura 9. Gráficos de *loadings* de PCA coloreados por porcentaje de contenido en GC. El gráfico de la izquierda pertenece a los datos de MAPQ5 tras el filtro CPM > 2, mientras que el gráfico de la derecha pertenece a los datos de MAPQ5 tras el filtro CPM > 2 normalizados por el método CQN. *Loadings* coloreados de tonalidades

azules muestran un porcentaje elevado de contenido en GC, mientras que tonalidades rojas muestran un porcentaje bajo y tonalidades verdes en torno al 50%. Se observan agrupaciones independientes entre los puntos correspondientes a variables con bajo y elevado porcentaje de contenido en CG.

Los demás métodos de normalización aplicados fueron: UQ, TMM y QN. Estos métodos de normalización unidos a CQN, fueron aplicados a los datos MAPQ1 y MAPQ5 con los correspondientes filtros CPM. Tras comparar la agrupación de las muestras de cada grupo estudiado en los gráficos de *scores* de los correspondientes PCAs para las dos primeras componentes principales (ver Figura 10 y Anexo 1), se decidió continuar el estudio con los valores pertenecientes al filtro MAPQ5 normalizado mediante el método QN. Aunque no había grandes diferencias entre los gráficos de *scores* de las distintas normalizaciones y filtros, las muestras se agrupaban en mejor en prematuros y a término para la opción elegida. Además, el método CQN apenas corrigió el sesgo del contenido en GC como se observa en la Figura 10, ya que tras su aplicación seguían observándose agrupamientos de *loadings* en relación con su contenido en GC que se encontraban separados por la PC1 y la PC2.

En el caso de los datos de lipidómica, se decidió aplicar también método de normalización por cuantiles, ya que las distribuciones eran muy desiguales en las distintas muestras.

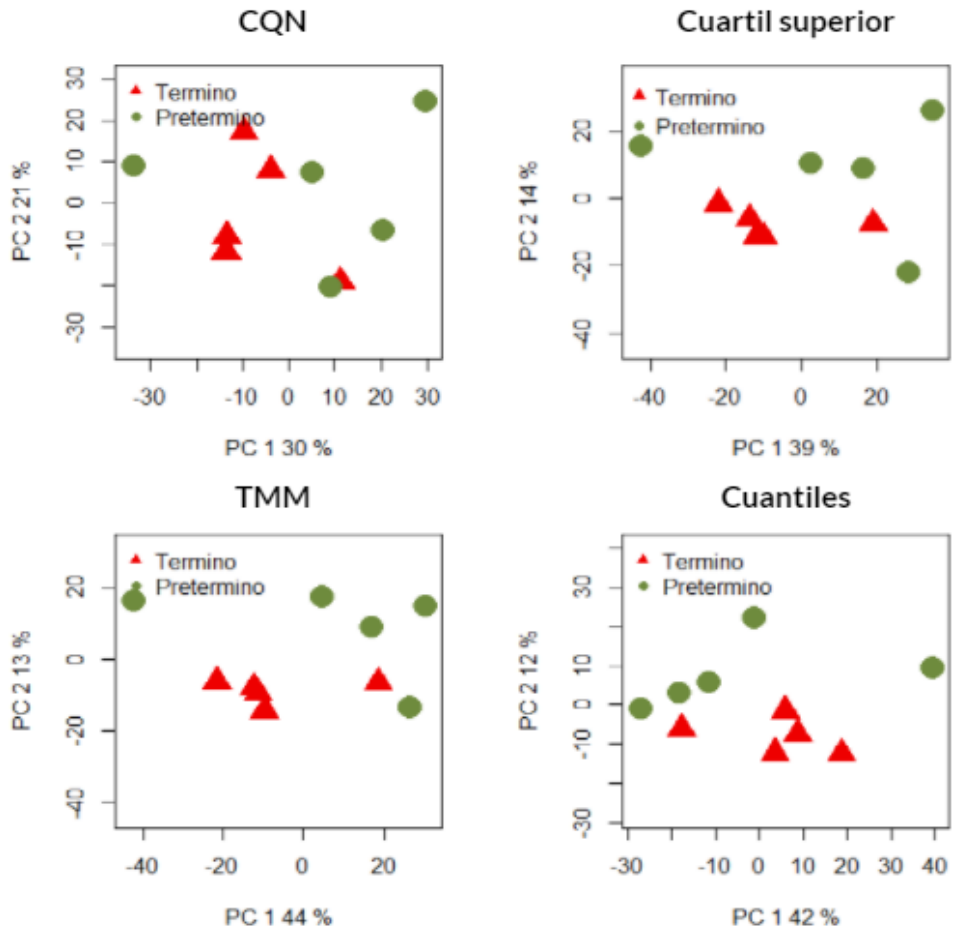


Figura 10. Gráfico de *scores* de PCA de los cuatro métodos de normalización aplicados a los datos de transcriptómica de MAPQ5 tras el filtro CPM > 2. Se observa una mayor separación entre los grupos estudiados mediante la segunda componente principal (PC2) al aplicar la normalización por cuantiles.

Cabe resaltar la presencia de una muestra del grupo "Pretérmino" que se aleja del resto y se sitúa en la zona central derecha en el caso del gráfico de scores tras la normalización por cuantiles de la Figura 10. Esta muestra es la "Pretérmino 5" que, tal y como se observa en la Tabla 1, corresponde a una madre de las más jóvenes del estudio (27 años) cuyo bebé nació de forma prematura y alcanzó un peso de nacimiento de 560 g, valor mucho menor que el resto de los pesos dentro del grupo "Pretérmino". Esta muestra se encuentra alejada del resto incluso trans la aplicación de los diferentes métodos de normalización, por lo que este alejamiento se debe a las características de la misma.

4.6. Análisis diferencial de variables ómicas entre grupos de pacientes

Posteriormente, se realizó un análisis estadístico de las distintas variables ómicas, para identificar aquellas con diferencias significativas en expresión o concentración entre el grupo a término y prematuros. Este análisis se llevó a cabo con el paquete **limma** de R. Se obtuvieron 56 miRNAs expresados diferencialmente entre ambos grupos (p-valor < 0,05) del total de 1042 analizados. En la Tabla 8 se pueden observar los 10 miRNAs más significativos, mientras que el resto se encuentran listados en el Anexo 2. Por otro lado, también se obtuvieron 30 lípidos con concentraciones alteradas entre ambos grupos de un total de 680 analizados (p-valor < 0,05). Los 10 lípidos más relevantes se encuentran en la Tabla 8, mientras que el resto se encuentran listados en el Anexo 3. Cabe destacar que no se utilizó el p-valor ajustado por tests múltiples para determinar las diferencias estadísticamente significativas porque no se detectaba ninguna variable ómica con cambios entre grupos al utilizar dicho p-valor ajustado, por lo que se tuvo en cuenta solo el p-valor sin ajustar.

Top 10 miRNAs y lípidos más relevantes						
Nº	miRNA	p-valor	logFC	Lípidos	p-valor	logFC
1	hsa-miR-21-5p	4,31E-04	-0,945	Stearoylcarnitine	3,22E-03	-1,280
2	hsa-miR-335-5p	3,90E-03	-1,076	SM d19:0_24:1 [M+H]+	9,21E-03	0,792
3	hsa-miR-30d-5p	6,52E-03	1,644	SP d16:0 [M+H]+	1,20E-02	-2,472
4	hsa-miR-4658	2,05E-03	1,473	SM d18:0_14:1 [M+H]+	1,95E-02	0,944
5	hsa-miR-1287-3p	1,93E-03	1,664	TG(12:0/i-14:0/14:0)	2,17E-02	1,250
6	hsa-miR-106b-5p	3,83E-03	-1,309	DG(14:0/16:1(9Z)/0:0)	1,64E-02	0,955
7	hsa-miR-375-3p	7,37E-03	0,914	DG(12:0/14:0/0:0)	1,92E-02	1,541
8	hsa-miR-19b-3p	4,95E-03	-1,387	TG(14:0/14:0/14:1(9Z))	3,31E-02	1,196
9	hsa-miR-29a-3p	8,89E-03	-1,002	SM d18:3_16:1 [M+H]+	3,07E-02	0,961
10	hsa-miR-99b-5p	6,48E-03	1,085	SM d19:0_24:4 [M+H]+	3,71E-02	0,983

Tabla 8. Top 10 miRNAs y lípidos más relevantes obtenidos mediante el análisis diferencial de variables ómicas entre grupos de pacientes. En la tabla se observa el nombre de los miRNAs y lípidos más relevantes, junto con su p-valor y el logaritmo del Fold-change (logFC). Aquellos con un valor de logFC positivo son miRNAs sobre-expresados o lípidos con mayor concentración en el grupo "Pretérmino", mientras que aquellos con un valor negativo lo están en el grupo "Término". Los caracteres junto al nombre del lípido hacer referencia a los tipos de enlaces que contiene y sus características. SM (esfingomielina), SP (esfingolípido), TG (triacilglicérido), DG (diacilglicérido).

4.7. Análisis de enriquecimiento de genes diana en miRNAs diferencialmente expresados y análisis de enriquecimiento de rutas biológicas en genes diana

El análisis de enriquecimiento de genes diana en miRNAs diferencialmente expresados dio como resultado un total de 2358 genes diana sobrerrepresentados en dicho conjunto de miRNAs. A su vez,

el análisis de enriquecimiento de rutas biológicas en estos genes diana, reportó un total de 26 términos GO significativamente sobrerrepresentados (p-valor ajustado < 0,05) (Anexo 4). En la Tabla 9 se pueden observar los 10 términos GO alterados más relevantes en el estudio. En esta tabla, los 6 primeros términos GO coloreados de morado (filas de 1 a 6 de la Tabla 9), se encuentran en el conjunto de los 26 significativamente sobrerrepresentados en relación al p-valor ajustado. Sin embargo, se han añadido otros 4 términos GO que a pesar de no estar alterados de forma significativa si se tiene en cuenta el p-valor ajustado, se observó en análisis posteriores que tenían importancia en el contexto del estudio. Cabe resaltar que estos 4 términos GO, a pesar de no ser significativos por el p-valor ajustado, sí tienen un p-valor < 0,05. Además, los 2 términos GO coloreados de rojo (filas 7 y 8 de la Tabla 9), tienen un p-valor ajustado < 0,1, situándose muy cerca del valor umbral.

Top 10 Términos GO más relevantes			
Nº	**Término GO**	**p-valor**	**p-valor aj.**
1	Apoptotic process	1,24E-08	2,92E-05
2	Defense response to virus	3,31E-07	6,93E-04
3	Type I interferon signaling pathway	1,67E-06	2,86E-03
4	Viral process	3,26E-06	4,39E-03
5	Inmune system process	4,20E-06	5,28E-03
6	Cytokine-mediated signaling pathway	6,27E-05	4,72E-02
7	Positive regulation of apoptotic process	8,26E-05	5,37E-02
8	Innate inmune response	1,80E-04	9,41E-02
9	Neutrophil chemotaxis	1,07E-03	2,74E-01
10	Regulation of ext. apoptotic signaling pathway via DD receptors	1,14E-03	2,76E-01

Tabla 9. Top 10 términos GO más relevantes del estudio. La tabla contiene los términos GO y los valores de p-valor y p-valor ajustado de cada uno de ellos.

Los resultados obtenidos en el enriquecimiento funcional de términos GO muestran los procesos biológicos supuestamente alterados entre la leche de madres de bebés nacidos a término y madres de bebés prematuros, entre los que cabe destacar "apoptotic process", "immune system process", "defense response to virus", "type I interferon signaling pathway" y "cytokine-mediated signaling pathway". Se procedió a utilizar la herramienta "**KEGG mapper**" de la base de datos KEGG para representar los genes diana asociados a dichos términos GO sobre las rutas biológicas de KEGG e interpretar así mejor su relevancia biológica y la conexión entre ellos. En los siguientes apartados se comentan las principales rutas biológicas obtenidas en este análisis.

4.7.1. Ruta de señalización de la prolactina

Entre los términos GO alterados se encuentra la ruta de señalización de la prolactina ("cytokine-mediated signaling pathway" en Anexo 4). La prolactina es una hormona que promueve la lactancia (Freeman et al., 2000) y su ruta de señalización depende de otras rutas, como las rutas PI3K y JAK/STAT.

La ruta de señalización JAK-STAT es el mecanismo principal de señalización para un gran conjunto de citoquinas y factores de crecimiento (Byfield et al., 2009). JAK (*Janus-activated kinase*) es una quinasa que activa STAT (*signal transducer and activator of transcription*), que se encarga de activar la transcripción de diversos genes. Esta ruta también se encontró alterada ("cytokine-mediated signaling pathway" en Anexo 4) y está implicada en la respuesta a infecciones virales y regula la apoptosis y el

ciclo celular. Dentro de esta ruta de señalización también encontramos otras moléculas cuyos genes son diana de los miRNAs alterados entre los dos grupos, como es el caso del supresor de señalización de citoquinas (SOCS). Dicha alteración podría deberse al estrés oxigénico experimentado por neonatos nacidos de forma prematura, desencadenando la señalización a través de JAK/STAT (Byfield et al., 2009).

Entre los argumentos que respaldan la alteración de esta ruta, también se observa que la ruta de señalización PI3K/AKT se encuentra alterada. PI3K participa en la progresión celular y es un regulador importante de numerosos tipos de cáncer, por lo que muchas terapias están dirigidas a esta ruta.

Su función recae en fosforilar PIP2 pasando a PIP3, siendo de esta manera reconocido por proteínas con dominios PH como AKT que continúa con la ruta de señalización. Dentro de esta ruta, se encuentra también alterado FOXO, factor de transcripción inhibido por Akt (Carter & Brunet, 2007). La alteración de los componentes de la ruta de señalización PI3K se muestran en Anexo 5.

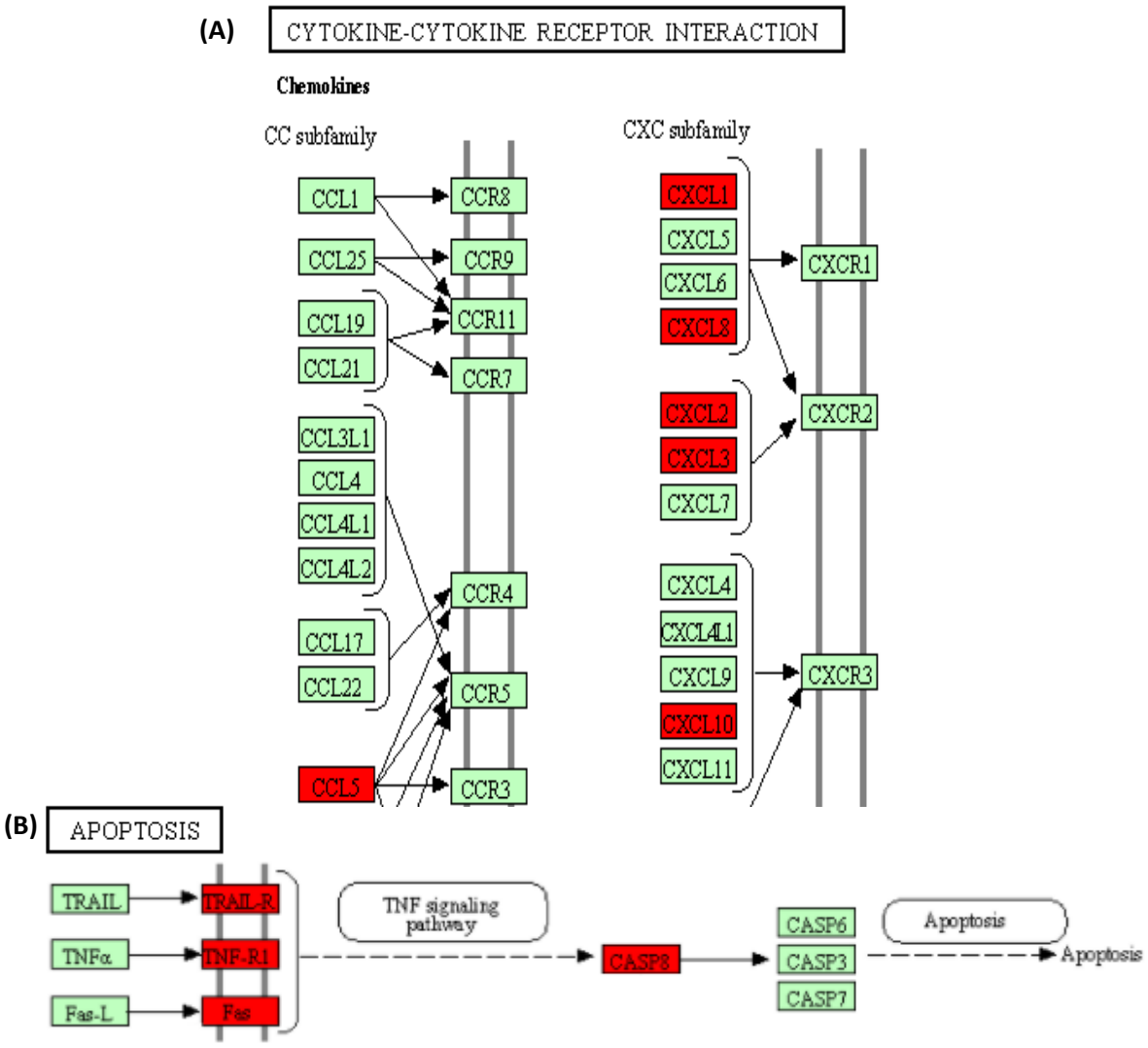


Figura 11. Figuras extraídas de la herramienta "**KEGG mapper**" de la base de datos KEGG. Los cuadros coloreados de rojo hacen referencia a los componentes de la ruta que se han obtenido en el análisis de enriquecimiento de genes diana a partir de los miRNAs diferencialmente expresados. **(A)** Componentes alterados en el sistema inmune. **(B)** Componentes alterados en la ruta de señalización de la apoptosis.

Además, encontramos otros componentes que se encuentran alterados en la ruta de señalización de la prolactina prolactina ("cytokine-mediated signaling pathway" en Anexo 4). Entre ellos, observamos moléculas como FOXO3a, factor de transcripción regulado por la ruta de señalización de PI3K. Además, encontramos la ciclina D1, común en cánceres de mama cuya función se basa en controlar la progresión de la célula a la fase de síntesis (S) del ciclo celular (Roy & Thompson, 2006); y también al receptor de la prolactina (PRLR).

En este contexto, varios artículos coinciden en la alteración tanto de los niveles de prolactina en plasma, como de producción de leche materna entre madres de bebés nacidos a término y madres de bebés nacidos a pretérmino (Chatterton et al., 2000; Mazor et al., 1996). Sin embargo, a pesar de que la variación de los niveles de prolactina parece clara, se observa cierta controversia en cuanto a la relación de la concentración de la misma entre ambos grupos, ya que algunos artículos defienden que los niveles de prolactina son significativamente menores en madres de neonatos prematuros (Chatterton et al., 2000), mientras que en otros postulan que la concentración de prolactina es significativamente mayor en madres de neonatos a pretérmino que en madres a término de forma general (Mazor et al., 1996).

4.7.2. Ruta de señalización de la apoptosis

Entre los términos GO alterados también se encuentra la ruta de señalización de la apoptosis ("apoptotic process" en Anexo 4). La apoptosis es un tipo de muerte celular programada que tiene un papel fundamental en el desarrollo y la homeostasis de los individuos. Tres genes diana encontrados en esta ruta (TRAILR2, Fas y TNF-R1) pertenecen a la superfamilia de receptores TNF-R (Sheikh & Fornace, 2000). A estos receptores, se les denomina receptores de muerte y desencadenan la apoptosis extrínseca, es decir, la generada por señales externas a la célula como moléculas solubles u otras células, a partir de la interacción con sus ligandos específicos. La interacción de los receptores con los ligandos desencadena la activación de las caspasas y sus consiguientes rutas de señalización, como es el caso de la caspasa 8, una proteasa que participa en la iniciación de este mecanismo de apoptosis extrínseca y que también resultó ser gen diana de los miRNAs alterados entre los dos grupos de muestras estudiadas. La alteración de los componentes de la ruta de apoptosis mencionados se muestra en la Figura 11-A.

Otras moléculas cuyos genes diana han sido encontrados en este estudio y relacionados con la actividad apoptótica, son el receptor TFGBR2, participando como mediador de la apoptosis (Schuster & Krieglstein, 2002); ASK1, una MAPKKK (*Mitogen Activated Protein (MAP) kinase kinase kinase*) cuya sobreexpresión induce la muerte celular apoptótica (Ichijo, 1997); y el IGF-BP3 que actúa como inhibidor del proceso de apoptosis mediante la inhibición de IGF (Ryan & Vousden, 1998). La idea de que la ruta de señalización de la apoptosis se encuentre alterada entre ambos grupos se encuentra respaldada por los resultados del estudio Hargitai et al., 2001, en el que se sugiere que la apoptosis desempeña un papel destacado en la patogénesis de numerosas enfermedades en bebés prematuros.

4.7.3. Sistema inmune

El sistema inmune es uno de los términos GO que está alterado entre los dos grupos de muestras y que aparece con el nombre de "immune system process" (Anexo 4). En el funcionamiento del sistema inmune están implicados diferentes tipos celulares y moléculas.

4.7.3.1. Ruta de señalización del interferón de tipo I

Entre los procesos implicados en el sistema inmune, la ruta de señalización del interferón de tipo I o IFN ("type I interferon signaling pathway") se encontró alterada (Anexo 4). Entre ambos grupos de muestras se encuentran diferentes moléculas relacionadas con el IFN, que realiza una acción antiviral en células infectadas por virus y en tejidos cercanos a estas, mediando la expresión de múltiples genes (Fitzgerald, 2011). Dentro de las moléculas relacionadas con el IFN cuya expresión se encuentra afectada, encontramos IPS-1 y IRF3/7, que juegan un papel esencial como inductores de la producción de IFN en respuesta a la infección viral (Potter et al., 2008; R.-P. Wang et al., 2008). Otro inductor alterado es la proteína denominada MAVS, que media la activación de IRF3 y, por lo tanto, participan en la inducción del interferón (Seth et al., 2005). También se observan otras moléculas que, en lugar de inducir, se encuentran inducidas por el IFN, como la quimiocina CXCL10/IP-10 (M. Liu et al., 2011), el IRF7 o la viperina, molécula efectora clave de la respuesta inmune contra virus (Fitzgerald, 2011). Asimismo, también la Ribonucleasa L (*RNase L*), es un importante efector de la respuesta antiviral innata (Urisman et al., 2006).

4.7.3.2. Sistema del complemento

Otro proceso relacionado con el sistema inmune que se vio alterado es el sistema del complemento. El sistema del complemento participa en la respuesta inmune innata y se encontraron alteraciones en las moléculas FH (factor H), MCP (proteína cofactor de membrana o CD46) y clusterina, que tienen como función la inhibición de la cascada del complemento (Ekdahl et al., 2016).

4.7.3.3. Citoquinas

Dentro del sistema inmune participan numerosas moléculas, entre las que destacan las citoquinas y quimiocinas cuyo GO ("cytokine-mediated signaling pathway") también se ha identificado como significativo en el análisis. En concreto, encontramos como genes diana de los miRNA estudiados algunos ligandos de la subfamilia de quimiocinas CXC (como CXCL1, CXCL2, CXCL3, CXCL8 y CXCL10) y también el CCL5 de la subfamilia CC (Figura 11-B). Un estudio reciente demuestra que los niveles de CCL5 están elevados en madres de bebés nacidos a pretérmino en relación con los de madres de bebés nacidos a término (Fattahpour et al., 2021). Estas moléculas tienen una función proinflamatoria, ya que inducen mediadores como el TNF; y está involucrada en procesos de dolor (Zychowska et al., 2015). Además, tanto las citoquinas como las quimiocinas están implicadas en procesos de polarización de macrófagos (REF). Los macrófagos son células del sistema inmune que pueden presentar un fenotipo pro-inflamatorio o M1 y pro-resultivo o M2, siendo ambos subtipos necesarios en los procesos inflamatorios. La polarización hacia un subtipo u otro de macrófagos depende de las citoquinas y quimiocinas que en ese momento se estén expresando. En este trabajo hemos identificado CXCL10, CCL-5 y TNFα como marcadores de M1, y CCL20, IL-10 y VEGF como marcadores de M2 (REF).

En este sentido, el laboratorio de la Dra. Pilar Sepúlveda del Instituto de Investigación Sanitaria La Fe ha realizado ensayos de polarización de macrófagos *in vitro*. Han observado que los macrófagos M1 tratados con vesículas extracelulares (EVs) de leche materna disminuyen la expresión de CD80 y CD86 (marcadores de M1) y no aumentan los niveles de CD163 (marcador de M2). Estos resultados sugieren que los EVs de leche materna podrían tener un papel anti-inflamatorio. Además, con los marcadores estudiados, no hemos observado diferencias en la polarización entre las EVs de leche materna a término o pretérmino.

De la misma manera, esta alteración de la ruta de señalización mediada por citoquinas está relacionada a su vez con la alteración de la ruta de señalización de la apoptosis. Esta relación se debe a que las quimiocinas activan las rutas de señalización de Jak2/3 o PI3K, que conducen a la activación del factor de transcripción FOXO3, factor de transcripción de la familia FOXO que juega un papel en la regulación de la inflamación y la apoptosis. Todas estas moléculas y rutas se encuentran alteradas en los datos analizados en el estudio. A su vez, esta relación viene reforzada por el hecho de que el parto prematuro está asociado a un aumento de moléculas de señalización proinflamatoria (Lim et al., 2013).

4.8. Análisis de enriquecimiento funcional de lípidos alterados entre grupos

Tras el análisis de enriquecimiento funcional de lípidos alterados entre grupos se obtuvo el gráfico mostrado en la Figura 12. En ella, se encuentra el conjunto de lípidos (términos LION) cuya concentración está alterada significativamente entre los dos grupos estudiados. En la mitad superior del gráfico se observan aquellas especies de lípidos con niveles más elevados en las muestras del grupo "Pretérmino", mientras que en la mitad inferior se encuentran aquellos con mayores niveles en el grupo "Término". Estos resultados mostraron que los términos LION relacionados con los niveles de triacilglicéridos (TGs) y diacilglicéridos (DGs) son más altos en las muestras correspondientes a madres de bebés prematuros. En cambio, los términos LION asociados a esfingomielinas (SM) y ceramidas (Cer) son más altos en las muestras de las madres con bebés a término.

Por un lado, DGs y TGs pertenecen al grupo de los acilgliceroles, que son ésteres del glicerol con uno o varios ácidos grasos, siendo 2 y 3 en el caso de DG y TG. Estas moléculas tienen funciones distintas a pesar de pertenecer al mismo grupo de lípidos. Los DGs tienen funciones relacionadas con la señalización celular, ya que actúan como segundos mensajeros y son un producto de la hidrólisis del PIP2 (bifosfato de fosfatidilinositol) catalizada por la enzima fosfolipasa C (PLC); mientras que los TGs tienen una función relacionada con la reserva energética, siendo almacenadas en el citosol en forma de gotas de grasa.

Por otro lado, los lípidos que se encuentran con mayores niveles en las muestras del grupo "Término", Cer y SM, pertenecen al grupo de los esfingolípidos y tienen distinta naturaleza y funciones. Los esfingolípidos son moléculas bioactivas que participan en diferentes procesos como la regulación del crecimiento celular, la muerte o la inflamación (Huang et al., 2011). Dentro de los esfingolípidos, las SM son los más abundantes en las células y componen un elemento esencial en la membrana plasmática de células y exosomas. El contenido de SM está estrictamente regulado por enzimas, cuya actividad crea una balanza entre la síntesis y la degradación. Como consecuencia de la hidrólisis de SM por la enzima esfingomielinasa (SMasa), se incrementa la concentración de la otra especie de lípido en cuestión, las ceramidas (Bienias et al., 2016). Las ceramidas son moléculas que constituyen el esqueleto hidrofóbico de todos los esfingolípidos complejos, entre ellos la SM, y estructuralmente consisten en un ácido graso de longitud de cada variable unido a un grupo amino, generalmente de la esfingosina. Se encuentran en la membrana plasmática como elementos de soporte del exosoma a niveles muy bajos, aunque estos pueden incrementarse significativamente de forma rápida y estable bajo condiciones de estrés celular o en respuesta a diferentes estímulos como citoquinas o ligandos de receptores de muerte, provocando diferentes respuestas biológicas (Castro et al., 2014). Entre estas respuestas, destaca la regulación de múltiples funciones biológicas, especialmente de la apoptosis, en la que los efectos citopáticos producidos por las ceramidas son de carácter pro-apoptótico (Huang et al., 2011).

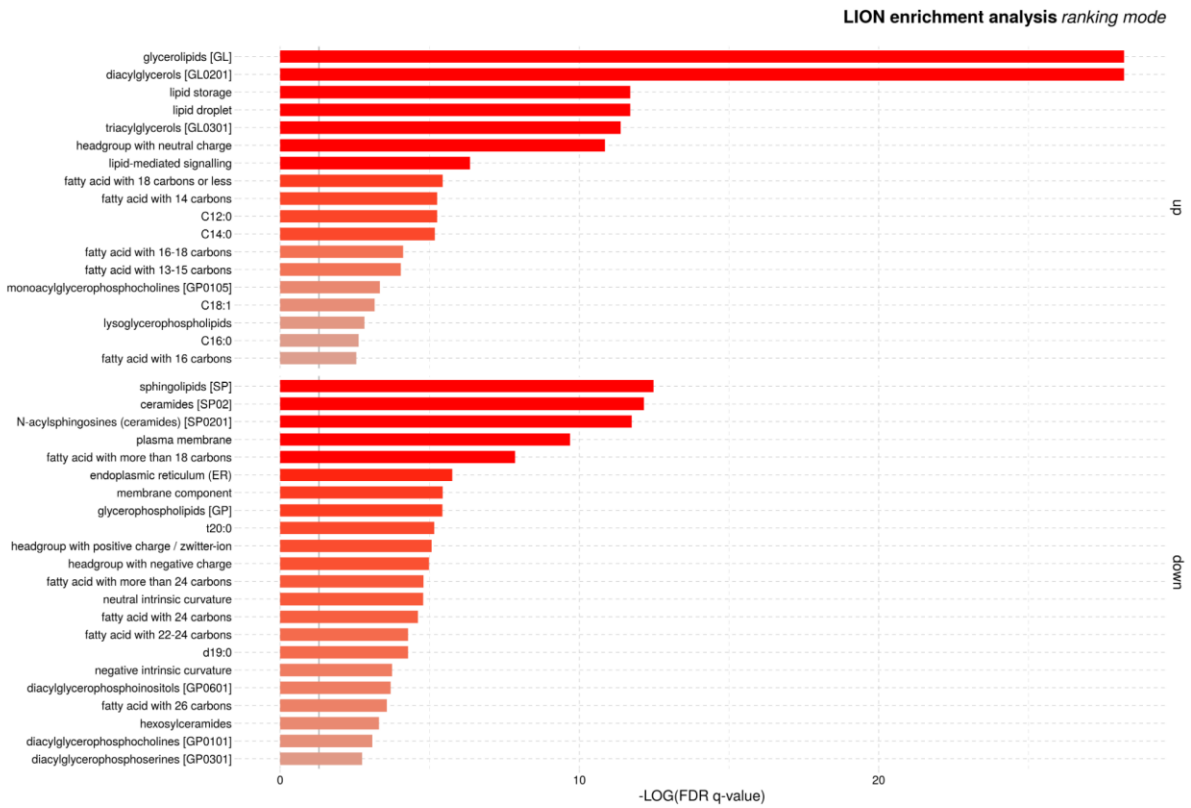


Figura 12. Resultados del análisis de enriquecimiento funcional de lípidos realizado con el *"Ranking mode"* de **LION**. Se observan las especies de lípidos con concentraciones significativamente alteradas entre los grupos "Término" y "Pretérmino".

Cabe esperar que, al igual que ocurre con los miRNAs, los lípidos atraviesen el tracto digestivo de los neonatos protegidos por la membrana de los exosomas y, por lo tanto, que participen en las rutas mencionadas en este apartado. Sin embargo, aún no hay estudios concluyentes que lo demuestren.

Tras finalizar los análisis del estudio, se observa que los resultados obtenidos del análisis lipidómico están relacionados en parte con los pertenecientes a los datos de transcriptómica. Esta relación se debe a que en ambos casos se han observado alteraciones en la ruta de la señalización de la apoptosis entre los grupos estudiados. Estos resultados, a su vez, abren campo a futuras investigaciones que amplíen los hallazgos del presente estudio.

5. Conclusiones

Las aportaciones de este TFG se enmarcan en dos ámbitos diferentes: el metodológico y el biológico.

Desde el punto de vista metodológico, se ha optimizado un procedimiento bioinformático para el análisis transcriptómico y lipidómico, que ha demostrado ser efectivo para extraer información biológica relevante en experimentos de alto rendimiento como son los datos ómicos.

Por una parte, se ha observado la importancia que conlleva la aplicación de distintos procedimientos de pre-procesado de los datos de transcriptómica, tales como la eliminación de adaptadores y el mapeo de lecturas de miRNA-Seq, la aplicación de filtros como MAPQ y CPM, la evaluación de sesgos propios de la secuenciación o la normalización de los datos mediante distintos métodos que también fueron aplicados a los datos de lipidómica. En este aspecto, cabe destacar que pese a obtener un bajo porcentaje de mapeo de lecturas contra el transcriptoma humano, este hecho no comprometió el resultado final del proyecto.

Por otra parte, se ha propuesto el uso de diferentes métodos de análisis para la extracción de información biológica. Entre ellos, se ha aplicado el análisis diferencial de variables ómicas entre grupos de pacientes mediante el que se han obtenido 56 miRNAs diferencialmente expresados y 30 lípidos con concentraciones alteradas entre ambos grupos estudiados. Además, se han aplicado distintos análisis de enriquecimiento que han permitido obtener información sobre los genes diana enriquecidos en los miRNAs diferencialmente expresados, relacionar estos genes diana con las rutas biológicas en las que intervienen y realizar un enriquecimiento funcional de los lípidos alterados. Mediante estos análisis de enriquecimiento se pudieron obtener los datos necesarios para realizar una correcta interpretación biológica.

Desde el punto de vista biológico, este estudio ha permitido un mejor entendimiento de los efectos de los nacimientos prematuros en la composición de la leche materna. Se ha constatado que existen diferencias en la composición de la leche materna de madres de bebés nacidos a término y a pretérmino. Estas diferencias han sido observadas mediante alteraciones en moléculas relacionadas con procesos como la respuesta inmune, la ruta de señalización de la prostaglandina o la ruta de la señalización de la apoptosis, y han sido respaldadas por estudios publicados sobre la materia. Incluso dentro de estos procesos, se han encontrado otros más específicos también alterados, como las rutas de señalización Jak/STAT y PI3K, la subfamilia de las quimiocinas CXC, la ruta de señalización del IFN I o la ruta de activación del complemento.

Además, este trabajo deja abierta la puerta a futuros análisis que continúen con el estudio de las alteraciones en la leche materna de madre de bebés a término y a pretérmino. Entre las diferentes posibilidades se encuentra llevar a cabo un análisis integrativo de la información obtenida mediante ambas ómicas o realizar un estudio exhaustivo de las deficiencias en la respuesta inmune y el mayor riesgo a sufrir enfermedades de neonatos a pretérmino y su relación con los componentes alterados en la leche de sus madres.

6. Bibliografía

Abels, E. R., & Breakefield, X. O. (2016). Introduction to Extracellular Vesicles: Biogenesis, RNA Cargo Selection, Content, Release, and Uptake. *Cellular and Molecular Neurobiology*, *36*(3).https://doi.org/10.1007/s10571-016-0366-z

Alsaweed, M., Lai, C. T., Hartmann, P. E., Geddes, D. T., & Kakulas, F. (2016). Human milk miRNAs primarily originate from the mammary gland resulting in unique miRNA profiles of fractionated milk. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep20680

Balakrishnan, R., Harris, M. A., Huntley, R., Van Auken, K., & Cherry, J. M. (2013). A guide to best practices for Gene Ontology (GO) manual annotation. *Database*, *2013*(0). https://doi.org/10.1093/database/bat054

Ballard, O., & Morrow, A. L. (2013). Human Milk Composition. *Pediatric Clinics of North America*, *60*(1). https://doi.org/10.1016/j.pcl.2012.10.002

Bienias, K., Fiedorowicz, A., Sadowska, A., Prokopiuk, S., & Car, H. (2016). Regulation of sphingomyelin metabolism. *Pharmacological Reports*, *68*(3). https://doi.org/10.1016/j.pharep.2015.12.008

Bleazard, T., Lamb, J. A., & Griffiths-Jones, S. (2015). Bias in microRNA functional enrichment analysis. *Bioinformatics*, *31*(10). https://doi.org/10.1093/bioinformatics/btv023

Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., Adler, A., Vera Garcia, C., Rohde, S., Say, L., & Lawn, J. E. (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*, *379*(9832). https://doi.org/10.1016/S0140-6736(12)60820-4

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(W1). https://doi.org/10.1093/nar/gkt282

Byfield, G., Budd, S., & Hartnett, M. E. (2009). The Role of Supplemental Oxygen and JAK/STAT Signaling in Intravitreous Neovascularization in a ROP Rat Model. *Investigative Opthalmology & Visual Science*, *50*(7). https://doi.org/10.1167/iovs.08-3256

Carter, M. E., & Brunet, A. (2007). FOXO transcription factors. *Current Biology*, *17*(4). https://doi.org/10.1016/j.cub.2007.01.008

Castro, B. M., Prieto, M., & Silva, L. C. (2014). Ceramide: A simple sphingolipid with unique biophysical properties. *Progress in Lipid Research*, *54*. https://doi.org/10.1016/j.plipres.2014.01.004

Chatterton, R. T., Hill, P. D., Aldag, J. C., Hodges, K. R., Belknap, S. M., & Zinaman, M. J. (2000). Relation of Plasma Oxytocin and Prolactin Concentrations to Milk Production in Mothers of Preterm Infants: Influence of Stress [1]. *The Journal of Clinical Endocrinology & Metabolism*, *85*(10). https://doi.org/10.1210/jcem.85.10.6912

del Mar Gómez-Ramos, M., Rajski, Ł., Heinzen, H., & Fernández-Alba, A. R. (2015). Liquid chromatography Orbitrap mass spectrometry with simultaneous full scan and tandem MS/MS for highly selective pesticide residue analysis. *Analytical and Bioanalytical Chemistry*, *407*(21). https://doi.org/10.1007/s00216-015-8709-z

Ekdahl, K. N., Huang, S., Nilsson, B., & Teramura, Y. (2016). Complement inhibition in biomaterial- and biosurface-induced thromboinflammation. *Seminars in Immunology*, *28*(3). https://doi.org/10.1016/j.smim.2016.04.006

Ewing, B., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. *Genome Research*, *8*(3). https://doi.org/10.1101/gr.8.3.186

Farooqi, A. A., Desai, N. N., Qureshi, M. Z., Librelotto, D. R. N., Gasparri, M. L., Bishayee, A., Nabavi, S. M., Curti, V., & Daglia, M. (2018). Exosome biogenesis, bioactivities and functions as new delivery systems of natural compounds. *Biotechnology Advances*, *36*(1). https://doi.org/10.1016/j.biotechadv.2017.12.010

Fattahpour, S., Moogooei, M., Aminzadeh, F., Ghorashi, Z., Khorramdelazad, H., & Hassanshahi, G. (2021). *Various pattern of CC chemokine expression in term and pre-term neonates along with their respected mothers.* (EBSCOhost, Vol. 2). Iranian Journal of Reproductive Medicine.

Fitzgerald, K. A. (2011). The Interferon Inducible Gene: Viperin. *Journal of Interferon & Cytokine Research*, *31*(1). https://doi.org/10.1089/jir.2010.0127

Freeman, M. E., Kanyicska, B., Lerant, A., & Nagy, G. (2000). Prolactin: Structure, Function, and Regulation of Secretion. *Physiological Reviews*, *80*(4). https://doi.org/10.1152/physrev.2000.80.4.1523

Garcia, N. A., Moncayo-Arlandi, J., Sepulveda, P., & Diez-Juan, A. (2016). Cardiomyocyte exosomes regulate glycolytic flux in endothelium by direct transfer of GLUT transporters and glycolytic enzymes. *Cardiovascular Research*, *109*(3).

https://doi.org/10.1093/cvr/cvv260

Garmire, L. X., & Subramaniam, S. (2012). Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*, *18*(6). https://doi.org/10.1261/rna.030916.111

Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The Lancet*, *371*(9606). https://doi.org/10.1016/S0140-6736(08)60074-4

Han, X. (2016). Lipidomics for studying metabolism. *Nature Reviews Endocrinology*, *12*(11). https://doi.org/10.1038/nrendo.2016.98

Han, X., Yang, K., & Gross, R. W. (2012). Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. *Mass Spectrometry Reviews*, *31*(1). https://doi.org/10.1002/mas.20342

Hansen, K. D., Irizarry, R. A., & WU, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, *13*(2). https://doi.org/10.1093/biostatistics/kxr054

Hargitai, B., Szabó, V., Hajdú, J., Harmath, Á., Pataki, M., Farid, P., Papp, Z., & Szende, B. (2001). Apoptosis in Various Organs of Preterm Infants: Histopathologic Study of Lung, Kidney, Liver, and Brain of Ventilated Infants. *Pediatric Research*, *50*(1). https://doi.org/10.1203/00006450-200107000-00020

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa942

Huang, W.-C., Chen, C.-L., Lin, Y.-S., & Lin, C.-F. (2011). Apoptotic Sphingolipid Ceramide in Cancer Therapy. *Journal of Lipids*, *2011*. https://doi.org/10.1155/2011/565316

Hurley, J. H., & Odorizzi, G. (2012). Get on the exosome bus with ALIX. *Nature Cell Biology*, *14*(7). https://doi.org/10.1038/ncb2530

Ichijo, H. (1997). Induction of Apoptosis by ASK1, a Mammalian MAPKKK That Activates SAPK/JNK and p38 Signaling Pathways. *Science*, *275*(5296). https://doi.org/10.1126/science.275.5296.90

Kahn, S., Liao, Y., Du, X., Xu, W., Li, J., & Lönnerdal, B. (2018). Exosomal MicroRNAs in Milk from Mothers Delivering Preterm Infants Survive in Vitro Digestion and Are Taken Up by Human Intestinal Cells. *Molecular Nutrition & Food Research*, *62*(11). https://doi.org/10.1002/mnfr.201701050

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1). https://doi.org/10.1093/nar/28.1.27

Kanehisa, Minoru, & Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science*, *29*(1). https://doi.org/10.1002/pro.3711

Keith Bradnam. (2014, December 16). *Understanding MAPQ scores in SAM files: does 37 = 42?* ACGT. http://www.acgt.me/blog/2014/12/16/understanding-mapq-scores-in-sam-files-does-37-42

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1141

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4). https://doi.org/10.1038/nmeth.1923

Laulagnier, K., Grand, D., Dujardin, A., Hamdi, S., Vincent-Schneider, H., Lankar, D., Salles, J.-P., Bonnerot, C., Perret, B., & Record, M. (2004). PLD2 is enriched on exosomes and its activity is correlated to the release of exosomes. *FEBS Letters*, *572*(1–3). https://doi.org/10.1016/j.febslet.2004.06.082

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2). https://doi.org/10.1186/gb-2014-15-2-r29

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16). https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Du, X., Li, J., & Lönnerdal, B. (2017). Human milk exosomes and their microRNAs survive digestion in vitro and are taken up by human intestinal cells. *Molecular Nutrition & Food Research*, *61*(11). https://doi.org/10.1002/mnfr.201700082

Lim, R., Barker, G., & Lappas, M. (2013). A Novel Role for FOXO3 in Human Labor: Increased Expression in Laboring Myometrium, and Regulation of Proinflammatory and Prolabor Mediators in Pregnant Human Myometrial Cells. *Biology of Reproduction*, *88*(6). https://doi.org/10.1095/biolreprod.113.108126

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., & Black, R. E. (2016). Global,

regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, *388*(10063). https://doi.org/10.1016/S0140-6736(16)31593-8

Liu, M., Guo, S., Hibbert, J. M., Jain, V., Singh, N., Wilson, N. O., & Stiles, J. K. (2011). CXCL10/IP-10 in infectious diseases pathogenesis and potential therapeutic implications. *Cytokine & Growth Factor Reviews*. https://doi.org/10.1016/j.cytogfr.2011.06.001

Llorente, A., Skotland, T., Sylvänne, T., Kauhanen, D., Róg, T., Orłowski, A., Vattulainen, I., Ekroos, K., & Sandvig, K. (2013). Molecular lipidomics of exosomes released by PC-3 prostate cancer cells. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, *1831*(7). https://doi.org/10.1016/j.bbalip.2013.04.011

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10). https://doi.org/10.1038/s41576-020-0236-x

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., & Cui, Q. (2008). An Analysis of Human MicroRNA and Disease Associations. *PLoS ONE*, *3*(10). https://doi.org/10.1371/journal.pone.0003420

Mailund, T. (2019). Manipulating Data Frames: dplyr. In *R Data Science Quick Reference*. Apress. https://doi.org/10.1007/978-1-4842-4894-2_7

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, *12*(2). https://doi.org/10.1038/nprot.2016.182

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1). https://doi.org/10.14806/ej.17.1.200

Mazor, M., Hershkowitz, R., Ghezzi, F., Cohen, J., Chaim, W., Wiznitzer, A., Levy, J., Leiberman, J. R., & Glezerman, M. (1996). Prolactin concentrations in preterm and term pregnancy and labour. *Archives of Gynecology and Obstetrics*, *258*(2). https://doi.org/10.1007/BF00626026

Molenaar, M. R., Jeucken, A., Wassenaar, T. A., van de Lest, C. H. A., Brouwers, J. F., & Helms, J. B. (2019). LION/web: a web-based ontology enrichment tool for lipidomic data analysis. *GigaScience*, *8*(6). https://doi.org/10.1093/gigascience/giz061

NATURE PORTFOLIO. (2021, June 25). *Transcriptomics*. NATURE PORTFOLIO. https://www.nature.com/subjects/transcriptomics

Oftedal, O. T. (2012). The evolution of milk secretion and its ancient origins. *Animal*, *6*(3). https://doi.org/10.1017/S1751731111001935

Ontoria-Oviedo, I., Dorronsoro, A., Sánchez, R., Ciria, M., Gómez-Ferrer, M., Buigues, M., Grueso, E., Tejedor, S., García-García, F., González-King, H., Garcia, N. A., Peiró-Molina, E., & Sepúlveda, P. (2018). Extracellular Vesicles Secreted by Hypoxic AC10 Cardiomyocytes Modulate Fibroblast Cell Motility. *Frontiers in Cardiovascular Medicine*, *5*. https://doi.org/10.3389/fcvm.2018.00152

Potter, J. A., Randall, R. E., & Taylor, G. L. (2008). Crystal structure of human IPS-1/MAVS/VISA/Cardif caspase activation recruitment domain. *BMC Structural Biology*, *8*(1). https://doi.org/10.1186/1472-6807-8-11

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3). https://doi.org/10.1038/nbt0308-303

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7). https://doi.org/10.1093/nar/gkv007

Roy, P. G., & Thompson, A. M. (2006). Cyclin D1 and breast cancer. *The Breast*, *15*(6). https://doi.org/10.1016/j.breast.2006.02.005

RStudio Team. (2020). *RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA*. http://www.rstudio.com/

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., Mahaffey, S., Rossi, S., Calin, G. A., Bemis, L., & Theodorescu, D. (2014). The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Research*, *42*(17). https://doi.org/10.1093/nar/gku631

Ryan, K. M., & Vousden, K. H. (1998). Characterization of Structural p53 Mutants Which Show Selective Defects in Apoptosis but Not Cell Cycle Arrest. *Molecular and Cellular Biology*, *18*(7). https://doi.org/10.1128/MCB.18.7.3692

Schuster, N., & Krieglstein, K. (2002). Mechanisms of TGF-β-mediated apoptosis. *Cell and Tissue Research*, *307*(1). https://doi.org/10.1007/s00441-001-0479-6

Seth, R. B., Sun, L., Ea, C.-K., & Chen, Z. J. (2005). Identification and Characterization of MAVS, a Mitochondrial Antiviral Signaling Protein that Activates NF-κB and IRF3. *Cell*, *122*(5). https://doi.org/10.1016/j.cell.2005.08.012

Sheikh, M., & Fornace, A. (2000). Death and decoy receptors and p53-mediated apoptosis. *Leukemia*, *14*(8). https://doi.org/10.1038/sj.leu.2401865

Simon Andrews. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Simon Andrews. (2019, January 8). *Per Tile Sequence Quality*. Babraham Bioinformatics.

Skotland, T., Sandvig, K., & Llorente, A. (2017). Lipids in exosomes: Current knowledge and the way forward. *Progress in Lipid Research*, *66*. https://doi.org/10.1016/j.plipres.2017.03.001

Squadrito, M. L., Baer, C., Burdet, F., Maderna, C., Gilfillan, G. D., Lyle, R., Ibberson, M., & De Palma, M. (2014). Endogenous RNAs Modulate MicroRNA Sorting to Exosomes and Transfer to Acceptor Cells. *Cell Reports*, *8*(5). https://doi.org/10.1016/j.celrep.2014.07.035

Tarazona, S., García, F., Ferrer, A., Dopazo, J., & Conesa, A. (2012). NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.Journal*, *17*(B).https://doi.org/10.14806/ej.17.B.265

Urisman, A., Molinaro, R. J., Fischer, N., Plummer, S. J., Casey, G., Klein, E. A., Malathi, K., Magi-Galluzzi, C., Tubbs, R. R., Ganem, D., Silverman, R. H., & DeRisi, J. L. (2006). Identification of a Novel Gammaretrovirus in Prostate Tumors of Patients Homozygous for R462Q RNASEL Variant. *PLoS Pathogens*, *2*(3). https://doi.org/10.1371/journal.ppat.0020025

van Meer, G., Voelker, D. R., & Feigenson, G. W. (2008). Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology*, *9*(2). https://doi.org/10.1038/nrm2330

Wang, R.-P., Zhang, M., Li, Y., Diao, F.-C., Chen, D., Zhai, Z., & Shu, H.-B. (2008). Differential regulation of IKKα-mediated activation of IRF3/7 by NIK. *Molecular Immunology*, *45*(7). https://doi.org/10.1016/j.molimm.2007.10.034

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1). https://doi.org/10.1038/nrg2484

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., How Huang, K., Jen Lee, M., Galas, D. J., & Wang, K. (2010). The MicroRNA Spectrum in 12 Body Fluids. *Clinical Chemistry*, *56*(11). https://doi.org/10.1373/clinchem.2010.147405

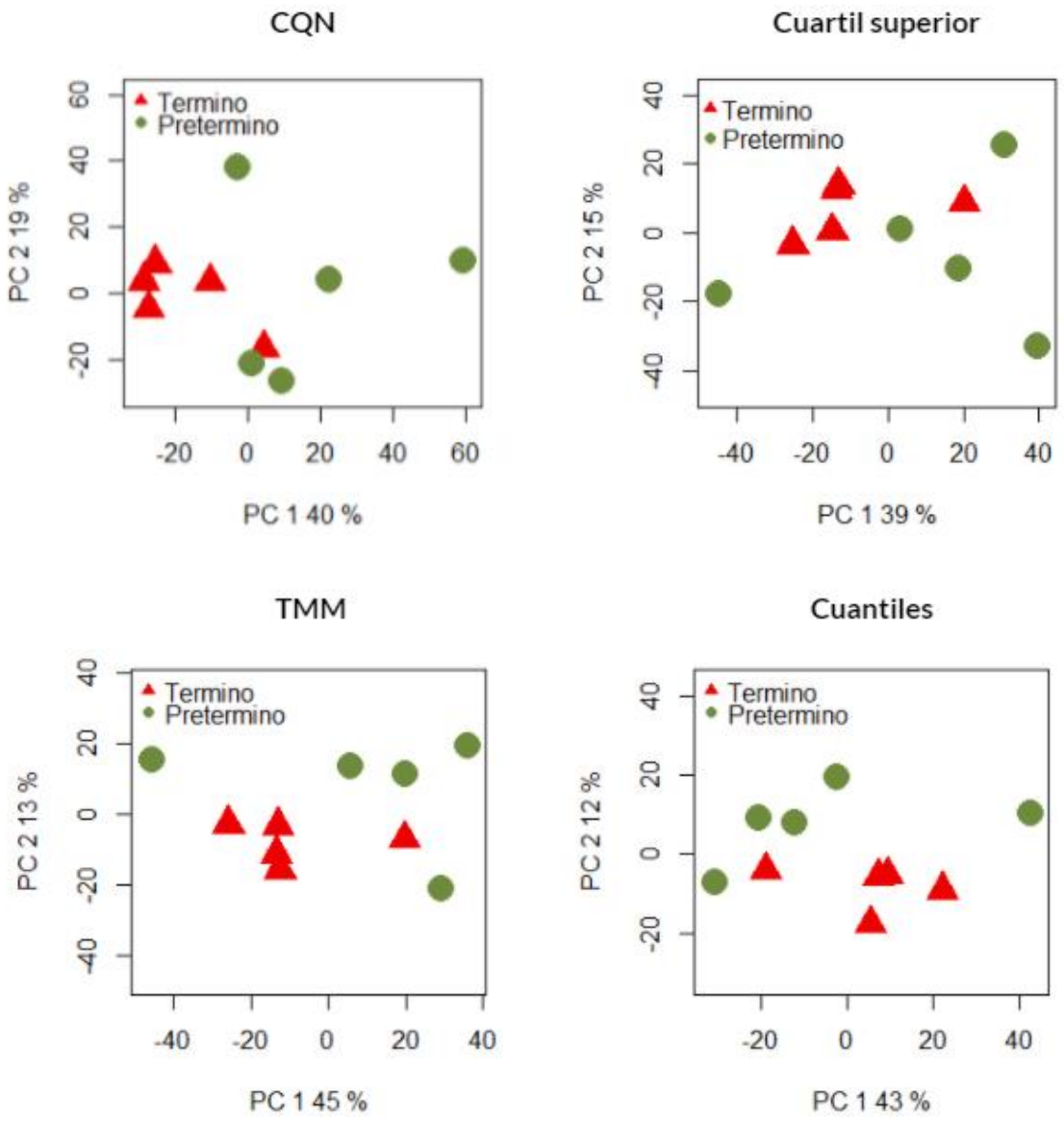
World Health Organization. (2018, February 19). *Preterm birth*. World Health Organization.

Zempleni, J., Aguilar-Lozano, A., Sadri, M., Sukreet, S., Manca, S., Wu, D., Zhou, F., & Mutai, E. (2017). Biological Activities of Extracellular Vesicles and Their Cargos from Bovine and Human Milk in Humans and Implications for Infants. *The Journal of Nutrition*, *147*(1). https://doi.org/10.3945/jn.116.238949

Zychowska, M., Rojewska, E., Pilat, D., & Mika, J. (2015). The Role of Some Chemokines from the CXC Subfamily in a Mouse Model of Diabetic Neuropathy. *Journal of Diabetes Research*, *2015*. https://doi.org/10.1155/2015/750182

7. Anexos

7.1. **Anexo 1:** gráficos de *scores* de PCA de los cuatro métodos de normalización aplicados a los datos de transcriptómica de MAPQ1 tras el filtro CPM > 2



7.2. **Anexo 2:** listado de los miRNAs diferencialmente expresados con p-valor y logFC

miRNAs diferencialemente expresados							
Nº	miRNA	p-valor	logFC	Nº	miRNA	p-valor	logFC
1	hsa-miR-21-5p	4,3E-04	-0,945	29	hsa-miR-4763-5p	2,3E-02	-1,250
2	hsa-miR-335-5p	3,9E-03	-1,076	30	hsa-miR-195-5p	3,1E-02	-1,312
3	hsa-miR-30d-5p	6,5E-03	1,644	31	hsa-miR-19a-3p	3,5E-02	-1,224
4	hsa-miR-4658	2,1E-03	1,473	32	hsa-miR-27a-5p	1,8E-02	2,287
5	hsa-miR-1287-3p	1,9E-03	1,664	33	hsa-miR-4637	3,0E-02	1,332
6	hsa-miR-106b-5p	3,8E-03	-1,309	34	hsa-miR-762	3,3E-02	-4,123
7	hsa-miR-375-3p	7,4E-03	0,914	35	hsa-miR-600	3,8E-02	1,359
8	hsa-miR-19b-3p	4,9E-03	-1,387	36	hsa-miR-370-5p	4,1E-02	1,211
9	hsa-miR-29a-3p	8,9E-03	-1,002	37	hsa-miR-26b-3p	2,3E-02	1,418
10	hsa-miR-99b-5p	6,5E-03	1,085	38	hsa-miR-15a-5p	3,8E-02	-1,060
11	hsa-miR-4668-3p	3,4E-03	1,488	39	hsa-miR-4729	3,7E-02	1,238
12	hsa-miR-20a-5p	8,0E-03	-1,060	40	hsa-miR-6842-3p	3,6E-02	1,480
13	hsa-miR-17-5p	9,1E-03	-1,128	41	hsa-miR-4428	4,7E-02	1,117
14	hsa-let-7c-5p	1,2E-02	1,049	42	hsa-miR-3921	4,7E-02	-0,951
15	hsa-miR-4778-3p	6,1E-03	1,533	43	hsa-miR-660-5p	3,7E-02	-1,239
16	hsa-miR-199a-5p	4,8E-03	2,284	44	hsa-miR-5703	2,7E-02	-1,417
17	hsa-miR-499b-3p	5,7E-03	-1,491	45	hsa-miR-34a-5p	3,3E-02	-1,602
18	hsa-miR-497-5p	8,0E-03	-1,668	46	hsa-miR-210-5p	3,8E-02	1,103
19	hsa-miR-101-3p	1,9E-02	-0,899	47	hsa-miR-1236-5p	4,5E-02	-1,416
20	hsa-miR-548an	3,8E-03	2,421	48	hsa-miR-4277	3,4E-02	-1,354
21	hsa-miR-146a-5p	2,0E-02	-1,849	49	hsa-miR-130a-3p	3,5E-02	-1,975
22	hsa-miR-765	5,7E-03	-1,924	50	hsa-miR-4474-3p	4,9E-02	1,038
23	hsa-miR-6792-3p	8,4E-03	1,913	51	hsa-miR-10399-3p	3,9E-02	1,426
24	hsa-miR-3065-3p	1,2E-02	1,580	52	hsa-miR-105-5p	4,8E-02	1,269
25	hsa-miR-511-5p	2,0E-02	1,755	53	hsa-miR-4259	4,9E-02	-1,387
26	hsa-miR-3170	2,0E-02	1,122	54	hsa-miR-5087	4,4E-02	-2,037
27	hsa-miR-27b-3p	3,0E-02	-1,125	55	hsa-miR-3197	4,6E-02	-1,360
28	hsa-miR-3615	1,5E-02	1,578	56	hsa-miR-584-5p	4,8E-02	1,249

7.3. **Anexo 3:** listado de los lípidos con concentraciones alteradas entre ambos grupos con p-valor y logFC

Lípidos con concentraciones alteradas entre grupos			
Nº	Lípidos	p-valor	logFC
1	Stearoylcarnitine	3,2E-03	-1,280
2	SM d19:0_24:1 [M+H]+;	9,2E-03	0,792
3	DG 18:1_20:1 [M+NH4]+;DG 18:0_20:2 [M+NH4]+;DG 14:1_24:1 [M+NH4]+;	4,0E-03	-1,340
4	SP d16:0 [M+H]+;	1,2E-02	-2,472
5	PS 18:2_18:1 [M+H]+;PS 18:3_18:0 [M+H]+;	1,4E-02	0,896
6	SM d18:0_14:1 [M+H]+;	2,0E-02	0,944
7	TG(12:0/i-14:0/14:0)	2,2E-02	1,250
8	DG(14:0/16:1(9Z)/0:0)	1,6E-02	0,955
9	PE(18:1(11Z)/18:2(9Z,12Z))	2,6E-02	0,641
10	DG(12:0/14:0/0:0)	1,9E-02	1,541
11	TG(14:0/14:0/14:1(9Z))	3,3E-02	1,196
12	Glycerol 1,2-didodecanoate 3-tetradecanoate	3,3E-02	1,044
13	SM d18:3_16:1 [M+H]+;	3,1E-02	0,961
14	4-(4-Hydroxyphenyl)-2-butanone	3,6E-02	-0,986
15	SM d19:0_24:4 [M+H]+;	3,7E-02	0,983
16	TG(14:0/16:1(9Z)/14:1(9Z))	4,1E-02	0,889
17	TG 12:0_12:0_18:2 [M+NH4]+;	3,6E-02	0,996
18	TG 12:0_14:0_14:0 [M+Na]+;TG 12:0_12:0_16:0 [M+Na]+;TG 10:0_14:0_16:0 [M+Na]+;	4,1E-02	1,025
19	Plasmenyl-PC P-20:0_13:0 [M+H]+;Plasmenyl-PC P-18:0_15:0 [M+H]+;	3,2E-02	-0,979
20	TG(14:0/14:0/14:1(9Z))*TG(12:0/i-16:0/16:0)*Glycerol 1,3-ditetradecanoate 2-(9Z-octadecenoate)	3,9E-02	1,572
21	DG 16:0_22:1 [M+NH4]+;DG 12:0_26:1 [M+NH4]+;DG 18:1_20:0 [M+NH4]+;DG 18:0_20:1 [M+NH4]+;	4,3E-02	-1,096
22	TG 14:1_16:0_20:6 [M+NH4]+;	4,2E-02	0,651
23	LysoPC 18:1 [M+H]+;	8,2E-03	1,520
24	TG 17:2_17:1_17:1 [M+NH4]+;	4,9E-02	-0,662
25	DG 18:1_18:0 [M+NH4]+;	5,0E-02	-0,675
26	LysoPC 16:0 [M+H]+;	4,2E-02	1,198
27	PE 18:2_18:1 [M+H]+; & Alkenyl-DG P-14:0_22:4 [M+H]+;	4,9E-02	0,578
28	SM d19:0_24:1 [M+Ac-H]-;SM d17:1_26:0 [M+Ac-H]-;	4,5E-02	0,709
29	DG 12:0_18:2 [M+NH4]+;	3,7E-02	1,102
30	TG 14:0_20:1_22:1 [M+NH4]+;TG 16:0_20:1_20:1 [M+NH4]+;	1,1E-02	-1,306

7.4. **Anexo 4:** listado de términos GO significativamente sobrerrepresentados por p-valor ajustado

Términos GO significativamente sobrerrepresentados por p-valor ajustado			
Nº	Término GO	p-valor	p-valor aj.
1	protein binding	1,49E-21	2,81E-17
2	cytoplasm	1,13E-12	6,10E-09
3	cytosol	9,04E-13	6,10E-09
4	nucleoplasm	3,13E-10	1,18E-06
5	nucleus	5,49E-10	1,72E-06
6	cell cycle	5,57E-09	1,50E-05
7	apoptotic process	1,24E-08	2,92E-05
8	defense response to virus	3,31E-07	6,93E-04
9	identical protein binding	4,68E-07	8,80E-04
10	type I interferon signaling pathway	1,67E-06	2,86E-03
11	negative regulation of viral genome replication	2,18E-06	3,42E-03
12	cytoskeleton	2,99E-06	4,33E-03
13	viral process	3,26E-06	4,39E-03
14	immune system process	4,20E-06	5,28E-03
15	cell division	6,62E-06	7,79E-03
16	positive regulation of transcription, DNA-templated	7,70E-06	8,52E-03
17	perinuclear region of cytoplasm	1,48E-05	1,53E-02
18	cell cycle arrest	1,54E-05	1,53E-02
19	microtubule	1,71E-05	1,61E-02
20	protein phosphorylation	1,99E-05	1,78E-02
21	response to virus	2,59E-05	2,22E-02
22	ubiquitin protein ligase binding	3,80E-05	3,11E-02
23	protein ubiquitination	4,16E-05	3,27E-02
24	cytokine-mediated signaling pathway	6,27E-05	4,72E-02
25	lipid droplet	6,53E-05	4,73E-02
26	peptidyl-serine phosphorylation	7,07E-05	4,93E-02

7.5. **Anexo 5:** componentes alterados de la uta de señalización PI3K obtenida con la herramienta "KEGG mapper"