TESIS DOCTORAL POR COMPENDIO DE PUBLICACIONES

# Aplicación de redes de expresión y regulación para el análisis de enfermedades raras

José Córdoba Caballero


UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica
Programa de Doctorado en Biotecnología Avanzada

Dirigida por Juan Antonio García Ranea y Pedro Seoane Zonjic

Málaga, 8 de enero de 2025



AUTOR: José Córdoba Caballero

https://orcid.org/0000-0002-1821-5742

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

# DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña JOSÉ CÓRDOBA CABALLERO
Estudiante del programa de doctorado DOCTORADO EN BIOTECNOLOGÍA AVANZADA
de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: APLICACIÓN DE REDES DE EXPRESIÓN Y REGULACIÓN PARA EL ANÁLISIS DE ENFERMEDADES RARAS

Realizada bajo la tutorización de ANA GRANDE PÉREZ y dirección de JUAN ANTONIO GARCÍA RANEA Y PEDRO SEOANE ZONJIC (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 20 de DICIEMBRE de 2024

| | |
|---|---|
| Fdo.:<br>Doctorando/a | Fdo.:<br>Tutor/a |
| Fdo.: | |


Edificio Pabellón de Gobierno. Campus El Ejido. 29071
**Tel.:** 952 13 10 28 / 952 13 14 61 / 952 13 71 10
**E-mail:** doctorado@uma.es

| Director/es de tesis |
| --- |





Edificio Pabellón de Gobierno. Campus El Ejido. 29071
**Tel.:** 952 13 10 28 / 952 13 14 61 / 952 13 71 10
**E-mail:** doctorado@uma.es

D. JUAN ANTONIO GARCÍA RANEA, Catedrático y Director del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga, D. PEDRO SEOANE ZONJIC, Investigador Posdoctoral del Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER) y contratado en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga y Dña. ANA GRANDE PÉREZ, Catedrática del Departamento de Biología Celular, Genética y Fisiología de la Universidad de Málaga:

## CERTIFICAN

Que D. José Córdoba Caballero, Graduado en Biología por la Universidad de Málaga, ha realizado bajo su dirección conjunta en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga el trabajo de investigación correspondiente a su Tesis Doctoral que lleva por título "Aplicación de redes de expresión y regulación para el análisis de enfermedades raras". Que las publicaciones en coautoría en revistas científicas relevantes con revision por pares que avalan de forma idonea la presente Tesis Doctoral por compendio son: "*Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer*", "*Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis*", "*Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite*" y "*Exploring miRNA-target gene pair detection in disease with coRmiT*", las cuales no han sido utilizadas en tesis anteriores. Tras revisar y supervisar este trabajo, lo consideran apropiado para que se presente y defienda ante el tribunal de tesis correspondiente, por lo cual

## AUTORIZAN

Su presentación y defensa para optar al grado de Doctor.

Málaga, 20 de diciembre de 2024

Los directores y tutora,

D. Juan A. García Ranea D. Pedro Seoane Zonjic D. Ana Grande Pérez

# Índice general

# Agradecimientos

Como preámbulo a esta memoria me gustaría dedicar unas palabras de agradecimiento a todas las personas que me han acompañado y apoyado antes y durante la realización de esta tesis doctoral.

En primer lugar, quiero agradecer a mi padre Pepe y a mi madre Carmen el esfuerzo por apoyar incondicionalmente mis estudios y por luchar contra viento y marea para que sus hijos tuvieran una vida mejor. A mi hermano Jorge, pretendo agradecerle todo el apoyo y co-sufrimiento en cervezas, teniendo claro que no hay cebada suficiente en el mundo para cumplir dicho propósito. A mi segunda familia, mis tíos Nieves y Jorge (el Gordo) y a mis primos Merche y Benito quiero agradecerles su apoyo y cariño, dejando claro que sin ellos esto no habría sido posible. También me gustaría mencionar a mi tío Paco, que se encargó de regar la semilla de la curiosidad que hoy florece. Además me gustaría agradecer al resto de mi familia, a los Caballero y a los Córdoba y a los Moriñigo-Muñoz-Leiva que me han acogido como uno más, en especial a Rocío por ser el pilar que más peso soporta.

En segundo lugar, quiero agradecer especialmente el apoyo de mis directores. A Juan Antonio García Ranea por darme la oportunidad de realizar este trabajo en su grupo de investigación y apoyarme en los momentos más complicados, y a Pedro Seoane Zonjic por guiarme, por fortalecer las técnicas que han hecho posible este trabajo y por haber sido un amigo más que un director de tesis. En este apartado me gustaría mencionar a James Richard Perkins que, aunque no conste en los papeles como director, ha cumplido con creces la labor de guiarme hasta conseguir mi mejor trabajo hasta la fecha y ha llegado a convertirse en un amigo.

En tercer lugar (y no por ello menos importante), quiero agradecer el apoyo a mis amigos. A Pedro y Héctor que siguen aguantándome después de tantos años. A Juanan que me ayudo a lanzarme a la piscina de la investigación. A Clara y Fran que han estado ahí desde el principio de esta tesis. A Clau por provocar resurgir poderoso del guerrero y ayudar a llegar al Olimpo y robar el fuego. A Lidia por no dejar que se me lleve el aire más que a dos palmos del suelo. A Elena que fuimos uña y carne antes de que la naturaleza pegara en su puerta. A Rafa por escuchar las mayores divagaciones que han salido de mi cabeza. A Pablo, que

aunque tardío, ha sido indudablemente el mayor descubrimiento que he hecho. Y un agradecimiento especial a todos los hermanos restantes de la pseudosecta de casas rurales que tenemos montada.

Por último, quiero agradecer a los compañeros que han participado de forma directa e indirecta en esta tesis doctoral, en especial a Fernando Jabato con quien trabajé mano a mano como co-minion durante los primeros años. Me gustaría agradecer además al Servicio de Supercomputación y Bioinformática (SCBI) por disponer los resursos y asistencia y al Departamento de Biología Molecular y Bioquímica de la Facultad de Ciencias de la Universidad de Málaga por darme un lugar para realizar este trabajo.

**Gracias**

# Capítulo 1

# Artículos avales

A continuación se listan los artículos publicados en revistas científicas que avalan esta tesis doctoral.

- Rojano E, **Córdoba-Caballero J**, Jabato FM, Gallego D, Serrano M, Pérez B, Parés-Aguilar Á, Perkins JR*, Ranea JAG†, Seoane-Zonjic P†. Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer. *J Pers Med.* 2021 Jul 27;11(8):730. doi: 10.3390/jpm11080730.[1][2]

- Pagano-Márquez R, **Córdoba-Caballero J**, Martínez-Poveda B, Quesada AR, Rojano E*, Seoane P*, Ranea JAG†, Ángel Medina M†. Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis. *Brief Bioinform.* 2022 Jul 18;23(4):bbac220. doi: 10.1093/bib/bbac220.[1] [2]

- Jabato FM, **Córdoba-Caballero J**, Rojano E, Romá-Mateo C, Sanz P, Pérez B, Gallego D, Seoane P*, Ranea JAG†, Perkins JR†. Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite. *Sci Rep.* 2021 Jul 23;11(1):15062. doi: 10.1038/s41598-021-94343-w.[1] [2]

- **Córdoba-Caballero J**, Perkins JR*, García-Criado F, Gallego D, Navarro-Sánchez A, Moreno-Estellés M, Garcés C, Bonet F, Romá-Mateo C, Toro R, Perez B, Sanz P, Kohl M, Rojano E, Seoane P, Ranea JAG. Exploring miRNA-target gene pair detection in disease with coRmiT. *Brief Bioinform.* 2024 Jan 22;25(2):bbae060. doi: 10.1093/bib/bbae060.[1]

[1] * autor de correspondencia
[2] † misma contribución.

# Acrónimos y abreviaciones

***HPO***:*Human Phenotype Ontology.*
***DECIPHER***: *DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources.*
***OMIM***: *Online Mendelian Inheritance in Man.*
**ADN**: Ácido desoxirribonucleico.
**ARN**: Ácido ribonucleico.
**miARN**: Micro ARN.
**ARNcirc**: ARN circular.
***PacBio***: *Pacific Biosciences.*
**GEDs**: Genes con expresión diferencial.
**RNA-seq**: Secuenciación de ARN.
**TPM**: Transcritos por millón.
***RTq-PCR***: Reacción en cadena de la polimerasa con transcripción inversa cuantitativa en tiempo real.
***WGCNA***: *Weighted Gene Correlation Network Analysis.*
***GO***: *Gene Ontology.*
***KEGG***: *Kyoto Encyclopedia of Genes and Genomes.*
***GSEA***: Gene Set Enrichment Analysis.
**PMM2-CDG**: *PMM2 Congenital Disorder of Glycosylation.*
**LMNA-DCM**: *LMNA Dilated Cardiomyopathy.*
**ER-A**: Enfermedades raras asociadas a procesos de angiogénesis.
**ACP**: Análisis de Componentes Principales.
**CP**: Componente Principal/ Componentes Principales.
**AJCP**: Agrupación Jerárquica de los Componentes Principales.
**ICARS**: Escala internacional cooperativa de la ataxia.
**MVRD**: El diámetro relativo medio-sagital del vermis.
**NCPRS**: Escala de puntuación pediátrica de desordenes congénitos de la glucosilación de Nijmegen.
**SSY**: Síndrome de Schaaf-Yang.
**ACM**: Análisis de Correspondencias Múltiples.
**AFM**: Análisis Factorial Múltiple.
***FDR***: *False Discovery Rate.*
**ANOVA**: Análisis de la varianza.

# Parte I

# Introducción

# Capítulo 2

# Introducción

## 2.1. Enfermedades raras

Las enfermedades raras, también denominadas minoritarias o huérfanas, se definen según el reglamento nº 141/2000 del Parlamento Europeo y del Consejo del 16 de diciembre de 1999 sobre medicamentos huérfanos (www.boe.es) como aquellas que afectan a menos de una persona por cada 2000 habitantes [1]. Se estima que en la actualidad existen entre 6.000 y 8.000 enfermedades raras conocidas, afectando a aproximadamente 36 millones de personas en la Unión Europea, lo cual significa que el impacto acumulado en la salud pública es mucho mayor de lo que implica el término "enfermedad rara"[2]. La mayoría de enfermedades genéticas son raras, pero no todas las enfermedades raras son genéticas, por ejemplo, hay enfermedades infecciosas, autoinmunes o cánceres de muy baja prevalencia [1] [3]. La mayoría de los pacientes con enfermedades raras manifiestan los primeros signos al nacer o durante la infancia [4]. Sin embargo, el diagnóstico temprano continúa siendo un desafío significativo [5]. Más del 80 % de estos pacientes desarrolla una discapacidad certificada, lo que genera un impacto profundo a nivel psicológico, laboral y económico, tanto en el ámbito personal como en el familiar [2] [6, 7]. Además, las enfermedades raras suponen un impacto económico relevante a las administraciones públicas y sanitarias, en gran parte porque la mayoría de ellas carecen de tratamientos adecuados [8]. Se ha demostrado que se necesitan más de 5 años y 16 pruebas clínicas de media para su diagnóstico completo [4], lo cual supone un alto coste de personal y material sanitario que se suma a las atenciones sanitarias específicas de cada paciente. Por lo tanto, se requieren de más estudios clínicos que generen conocimiento sobre estas patologías para redu-

[1] https://www.orpha.net/es/other-information/about-rare-diseases

[2] https://www.enfermedades-raras.org/enfermedades-raras/conoce-mas-sobre-er/enfermedades-raras-en-cifras

cir costes económicos, acortar el tiempo de diagnóstico y proponer nuevas dianas terapéuticas que consigan paliar total o parcialmente sus síntomas.

Los estudios clínicos con poblaciones pequeñas limitan la generación de conocimiento y están sujetos a la elección apropiada de los individuos y la metodología, cuyo equilibrio es fundamental [9]. Por un lado, la cohorte de estudio o el organismo modelo experimental, en la medida de lo posible, deben ser escogidos con rigurosidad. En el caso de las cohortes de pacientes se debe considerar la heterogeneidad de las poblaciones, el estado de progresión de la enfermedad y las particularidades moleculares que puedan presentar individuos concretos [10]. Por otro lado, la metodología debe ser escogida acorde a la cohorte de estudio y debe presentar un equilibrio entre los métodos estadísticos establecidos en la comunidad científica y nuevos métodos que puedan generar conocimiento adicional [9].

## 2.2. Fenotipado de pacientes y recursos

El fenotipado de calidad de pacientes con enfermedades raras es un paso fundamental para el diagnóstico y, además, es una fuente de información muy útil a la hora de investigar el origen genético, molecular o ambiental de los fenotipos patológicos. En el contexto de la investigación es muy importante un fenotipado detallado que recoja los fenotipos característicos para una determinada enfermedad, conocido como fenotipado profundo [11]. Este fenotipado debe estar codificado y estandarizado para asegurar la rigurosidad e integración de los estudios clínicos.

La *Human Phenotype Ontology* (*HPO*) es un recurso reconocido globalmente como el estándar de codificación formal de fenotipos que estructura la información fenotípica de manera jerárquica en una ontología de términos (llamados términos *HPO*) donde los fenotipos se organizan desde los más generales a los más específicos en una estructura de grafo dirigido y acíclico. Esta codificación permite la aplicación de algoritmos avanzados que combinen los análisis fenotípicos y genómicos [12]. Por ejemplo, algunos métodos de similitud semántica permiten cuantificar el parecido entre dos grupos de términos teniendo en cuenta la posición de dichos términos en una ontología, como por ejemplo los métodos de Robinson [13], Lin [14] y Resnik [15] [16]. Existen iniciativas biomédicas que usan los términos *HPO* para fenotipar pacientes como la *DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources* (*DECIPHER*) [17] y para caracterizar enfermedades como la *Online Mendelian Inheritance in Man* (*OMIM*) [18].

Además, existen otras iniciativas como la *Monarch Initiative* que integran bases de datos con información de genes como la propia *HPO*, *OMIM*, *ClinVar* [19], *Orphanet* [20], *Comparative Toxicogenomics Database* [21] y el catálogo *Genome-wide association study* [22], entre otras, y unifican la información para que sea más accesible. Esta información integrada es usada por programas como *Exomiser* [23]

para identificar variantes causantes de la enfermedad de pacientes.

A pesar de los esfuerzos de integración, la información disponible debe curarse mediante el esfuerzo de personal experto, y en muchos casos la información fenotípica disponible, traducida a *HPO*, suele ser incompleta por el escaso número de términos en los perfiles fenotípicos y la falta de profundidad en el fenotipado. Para ayudar en esta labor, han surgido recientemente programas que usan modelos de lenguaje para traducir la información fenotípica de las historias clínicas en términos *HPO* [11] que tienen mejor rendimiento que los métodos de minería de texto con el mismo propósito [24]. Sin embargo, se necesitan de herramientas para analizar si la calidad del fenotipado de los pacientes de una cohorte es suficiente para extraer información de la misma con análisis computacionales posteriores.

## 2.3. Tecnologías de secuenciación

Las tecnologías de secuenciación son un conjunto de métodos capaces de leer cadenas de nucleótidos presentes en los organismos. Las tecnologías de secuenciación han sufrido un proceso de mejora y refinamiento desde que Sanger y colaboradores publicaran el primer método capaz de leer la secuencia de ácido desoxiribonucléico (ADN) del fago $\phi$X174 en 1977 [25, 26]. Este método se basa en la incorporación de nucleótidos modificados llamados didesoxinucleótidos que detienen la síntesis de ADN generando fragmentos de distinto tamaño. Estos fragmentos son separados en un gel de electroforesis y se detectan mediante radioactividad o fluorescencia. Aunque esta tecnología fue crucial para los avances iniciales en biología molecular y se sigue usando para la secuenciación de genes individuales y en estudios de variantes genéticas en clínica, esta limitada en cuanto a velocidad y escalabilidad ya que solo permite secuenciar una cadena de ADN a la vez [26, 27].

Las tecnologías de secuenciación masiva o de alto rendimiento de segunda generación, llamadas comúnmente cómo tecnologías de secuenciación de próxima generación (*Next Generation Sequencing*) revolucionaron la genómica y transcriptómica por su capacidad de generar grandes volúmenes de datos de secuenciación en paralelo. Las plataformas más conocidas son las de *Illumina* y *Ion Torrent* que utilizan la técnica de secuenciación por síntesis. Este tipo de tecnología se basa en la síntesis de cadenas complementarias a fragmentos de ADN molde que han sido previamente fijados a una superficie sólida mediante un adaptador. Debido a que los adaptadores son universales, este tipo de secuenciación no requiere conocimiento previo de la secuencia de ADN molde para poder secuenciarla, lo que permite identificar secuencias nuevas, incluidas las secuencias de ácidos ribonucléicos (ARN) como los micro ARN (miARN) y ARN no codificantes. El proceso de síntesis en estas plataformas incorpora de manera secuencial nucleótidos modificados que liberan fluorescencia (en el caso de *Illumina*) o iones de hidrógeno (en el caso de

*Ion Torrent*) que son detectadas por un sensor [27]. Pese a que las secuencias de estas plataformas son más cortas (entre 100-300 pb), la enorme cantidad de lecturas generada permite la reconstrucción de genomas completos, la secuenciación de transcriptomas, exomas y análisis de variantes genéticas, de modificaciones epigenéticas y de metagenomas [27, 28].

Las tecnologías de tercera generación introdujeron avances relevantes en la longitud de las lecturas y la capacidad de secuenciar moléculas completas de ADN sin necesidad de amplificación. Además fueron las primeras tecnologías capaces de secuenciar directamente ARN sin una retrotranscripción previa a ADN. Las principales tecnologías de esta generación son *Pacific Biosciences* (*PacBio*) y *Oxford Nanopore*. Por un lado, *PacBio* utiliza la secuenciación en tiempo real de moléculas únicas. En esta tecnología, la molécula de ADN (de hasta 30.000 nt) pasa por un pozo donde una ADN polimerasa fijada incorpora nucleótidos marcados con fluorescencia. El hecho de conocer el sitio exacto donde ocurre la reacción de polimerización permite al sistema focalizarse en una sola molécula [27]. Por otro lado, *Oxford Nanopore* emplea una técnica basada en nanoporos. Un nanoporo, en este contexto, es un complejo proteico en una membrana sintética semipermeable que fuerza el paso controlado de una molécula de ADN o ARN. Debido a que hay un gradiente electroquímico entre ambos lados de la membrana, el paso de la molécula de ADN o ARN produce un cambio de voltaje que se puede medir por un sensor [27]. Ambas tecnologías de tercera generación presentan ventajas sobre los métodos de segunda generación y la longitud de las lecturas que producen facilitan procesos como los ensamblajes *de novo* de genomas y transcriptomas [28], pero producen lecturas con una tasa de error muy alta y, aunque han mejorado con ciertos algoritmos de corrección de errores, a día de hoy no son las técnicas más usadas.

## 2.4. ARN no codificantes como elementos reguladores

La hipótesis central de la biología molecular[3] propuesta por Francis Crick en 1958 [30] es ampliamente conocida en el campo y postula que la información genética fluye en una sola dirección, del ADN al ARN y de este a la proteína, o del ARN directamente a la proteína. Sin embargo, esta hipótesis, aunque es parcialmente cierta, quedó obsoleta debido a que la información genética puede fluir desde ARN a ADN [31], y hay casos en los que no es necesario que la información genética

[3]Aunque dicha hipótesis se conoce como "dogma central de la biología molecular" el uso del término "dogma" no es correcto en un entorno científico. El propio autor, Francis Crick, reconoció en su biografía que la elección de dicho término no fue acertada [29].

cambie de molécula para ejercer su función, como es el caso de los ARN no codificantes.

En los últimos 40 años se han descubierto numerosas moléculas de ARN que no codifican proteínas, llamados ARN no codificantes. Hay muchos tipos de ARN no codificantes que realizan funciones esenciales en una célula como pueden ser el ARN transferente, el ARN ribosómico y los ARN pequeños nucleares, entre muchos otros [32]. Sin embargo hay un conjunto de ARN no codificantes cuyo papel es la regulación post-transcripcional de otros ARN mediante hibridación, este es el caso de los micro ARN (miARN).

Los miARN son moléculas pequeñas de ARN de, aproximadamente, 22 nt conservadas entre especies cuya función principal es la inactivación de transcritos a distintos niveles. Los miARN se transcriben como pre-miARN que son moléculas de unos 70 nt con una estructura secundaria característica, en la que los extremos 3' y 5' hibridan entre si ya que son cuasi-complementarios generándose un bucle monocatenario en la región central del pre-miARN [33]. Este pre-miARN se exporta al citoplasma y es procesado por la proteína DICER. Esta proteína corta el blucle central y captura una de las dos hebras restantes que recibe el nombre de miARN. DICER forma parte del complejo miRISC, que captura una serie de proteínas de las cuales AGO (proteína Argonauta) es la más relevante. Cuando AGO se une a un miARN es capaz de cortar ARN complementarios al miARN, promoviendo su degradación y, además es capaz de inhibir la iniciación de la traducción [33]. Este mecanismo de regulación por miARN es un importante método para regular genes diana específicos con un potencial enorme para la biotecnología. Además, se ha demostrado que tienen un papel relevante en distintos tipos de cáncer [34, 35, 36], enfermedades cardíacas [37, 38, 39] y enfermedades raras [40, 41, 42]. Hasta tal punto que su descubrimiento ha sido premiado con el Premio Nobel de Fisiología o Medicina en el año 2024[4].

Sin embargo, la regulación por miARN es algo más compleja de lo que parece. El pequeño tamaño de estas moléculas aumenta la probabilidad de unirse a más dianas, de hecho, se ha comprobado que hay miARN que afectan a numerosas dianas y hay transcritos a los que se pueden unir diferentes miARN [33]. Esto genera una competición endógena de los ARN por unirse a los miARN que puede modular la potencia de represión del miARN [43]. Además, no solo compiten entre sí ARN codificantes, sino que una de las funciones descritas para los ARN largos no codificantes, y específicamente para el subrupo de ARN circulares, es precisamente la de capturar miARN para modular el mecanismo de represión [44].

La complejidad de estas relaciones de regulación obliga a los investigadores a analizarlas con datos de secuenciación masiva y aplicando análisis de redes de coexpresión, como se desarrollará a lo largo de esta tesis doctoral.

[4]https://www.nobelprize.org/prizes/medicine/2024/summary

## 2.5. Análisis de la expresión génica

Los análisis de la expresión de genes o génica son métodos para estudiar la actividad de los genes de una o varias células y se basan en la cuantificación de las copias de ARN que se generan mediante la transcripción de la secuencia de los genes. Por un lado, hay técnicas capaces de medir la expresión de genes concretos como es el caso de la reacción en cadena de la polimerasa a tiempo real o cuantitativa (RTq-PCR por sus siglas en inglés). Esta técnica permite la cuantificación de la expresión transcripcional de genes concretos en muestras biológicas de distintas condiciones experimentales para comprobar, posteriormente, los cambios en la expresión de los mismos de forma estadística [27]. Los genes que cambian su expresión entre distintas condiciones suelen llamarse genes expresados diferencialmente o con expresión diferencial (GEDs). Los GEDs suelen tener asociadas dos valores: el $log_2FC$ que es el logaritmo en base 2 del ratio de cambio de la expresión de un gen entre dos grupos de muestras y es una medida cuantitativa del cambio de expresión y el valor P relativo a la prueba estadística, generalmente la prueba $t$ de Student, que es una probabilidad que indica la significancia del $log_2FC$. Por otro lado, existen métodos que permiten la cuantificación de la expresión de un gran número de genes a la vez, como es el caso de las micromatrices (del inglés *microarrays*) de expresión y la secuenciación de ARN (RNA-seq por sus siglas en inglés)[5].

Las micromatrices permiten cuantificar de manera relativa un gran número de genes. Esta técnica se basa en la hibridación de moléculas plantilla del ADN complementario (ADNc) de moléculas de ARN generados por retrotranscripción y marcadas con fluorescencia a unas sondas de 60 nucleótidos que están unidas de forma covalente a una placa de cristal. La fluorescencia de las plantillas que hibriden a las sondas puede ser escaneada y procesada por programas informáticos específicos para cuantificar la expresión indirecta de los genes [45]. Los GEDs se pueden calcular mediante la cuantificación por micromatrices con algoritmos específicos como puede ser *limma* [46], que utiliza transformaciones logarítmicas y aplica modelos lineales. Esta técnica está limitada por la cantidad de sondas que hay fijadas a la placa en doble sentido ya que la cantidad de sondas distintas limita la cantidad de genes que pueden cuantificarse y la cantidad de sondas del mismo gen afecta a la precisión de la cuantificación del mismo [45, 47]. A causa de las limitaciones de esta técnica y la existencia de las técnicas de RNA-seq, las micromatrices están en desuso.

Por otra parte, las técnicas de RNA-seq usan las tecnologías de secuenciación, siendo el uso de las de segunda generación el más extendido. En este caso, los ARN

[5]RNA-seq es el acrónimo comercial y globalmente extendido en la comunidad científica para referirse a las tecnologías de secuenciación de ARN y viene del inglés *RiboNucleic Acid sequencing*

se retrotranscriben a ADN y se dividen en fragmentos del mismo tamaño para generar la librería de secuenciación [27]. Estos fragmentos, una vez digitalizados en lecturas por las plataformas de secuenciación, se pueden mapear con programas como *STAR* [48] o *Bowtie* [49] para poder posicionarlos sobre el genoma o transcriptoma de referencia. Posteriormente, estos alineamientos se pueden cuantificar para generar una tabla de conteos. Esta tabla, que contiene la cantidad de lecturas que se ha obtenido para cada gen en cada una de las muestras, es susceptible de la aplicación de algoritmos para calcular los GEDs pero necesita de transformaciones previas para corregir posibles sesgos técnicos. Las técnicas de RNA-seq suelen presentar una serie de sesgos de cuantificación. Por ejemplo, está ampliamente estudiado que los genes cuyos transcritos son de mayor tamaño se separan en más fragmentos de ADNc y, por tanto, las plataformas de secuenciación tienden a generar mayor cantidad de lecturas para estos genes [50]. Además, estas plataformas suelen generar distinto número de lecturas para cada muestra [51]. Para corregir esto la tabla de conteos debe ser sometida a procedimientos de normalización como pueden ser el cálculo de los transcritos por millón (TPM) o la normalización por muestras virtuales de referencia [52]. Una vez se obtiene la matriz normalizada se pueden aplicar algoritmos específicos para calcular la expresión diferencial como *limma-voom*, *DESeq2* [52], *edgeR* [53] y *NOISeq* [54]. Aunque la cuantificación de la expresión con RNA-seq no es perfecta, supera las desventajas que causan el uso de sondas de las micromatrices, permitiendo así el estudio de todos los ARN expresados en una o varias células. Aún así, los resultados obtenidos por los análisis de expresión diferencial suelen validarse como precaución por RTq-PCR para confirmar el cambio observado.

Además de los análisis de expresión diferencial, la cuantificación realizada por las micromatrices o por RNA-seq son susceptibles de estudios de la expresión complementarios al calculo de la expresión diferencial como puede ser el estudio de la coexpresión de genes mediante redes de correlación, siendo el algoritmo de *WGCNA* [55] el de uso más extendido. Dado que el algoritmo de *WGCNA* aplica métodos de análisis de redes, será comentado en detalle en el próximo apartado.

## 2.6. Análisis de redes biológicas

La gran cantidad de información que se puede obtener de bases de datos biológicas o de técnicas experimentales que generan grandes cantidades de datos se puede modelar y analizar usando modelos de redes. Éste puede ser aplicado a cualquier conjunto de datos que se relacionen entre sí de forma compleja. De esta manera, los datos se organizan como nodos que se relacionan entre si mediante enlaces, los cuales pueden tener pesos o distintos atributos [56]. Los nodos de la red se distribuyen según su grado, que es la cantidad de conexiones directas que un nodo tiene

con el resto de la red. Las redes biológicas tienen una estructura libre de escala donde hay pocos nodos con un grado muy alto (llamados *hubs* en inglés) mientras que la mayoría tienen un grado bajo. Este tipo redes son robustas porque mantienen su estructura cuando faltan nodos aleatorios pero, a su vez, son vulnerables cuando faltan nodos *hubs* [56].

Además existen varios tipos de redes según los datos que se modelen. Cuando se estudian datos con las mismas propiedades, como puede ser genes y sus relaciones con otros genes, la red tiene una sola capa o tipo de información y recibe el nombre de red monopartita. Cuando los datos provienen de distintas fuentes o tienen distintas características, como pueden ser genes, fenotipos y enfermedades, se pueden modelar en redes n-partitas o multipartitas [56, 57, 58, 59, 60].

El modelado de redes en biología es una herramienta muy útil ya que permite un análisis holístico, en el que el foco de estudio es el conjunto de nodos y la estructura de la propia red [61, 56, 62, 63, 64]. Uno de los análisis comunes de estas redes es calcular agrupaciones de genes con cualquier algoritmo que tenga en cuenta la estructura de la red, aunque en algunos casos se requiere calcular las distancias que existen entre cada pareja de nodos como por ejemplo, el camino más corto. La medida del camino más corto entre dos nodos es la mínima cantidad de nodos o mínima distancia de enlaces que los separa. Esta medida puede ser útil, por ejemplo, para evaluar la cohesión de un grupo de genes en una red de interacción [65, 66]. Otro ejemplo de análisis de redes puede ser la predicción de interacciones entre nodos de dos capas distintas gracias a las interacciones con una tercera capa. Por ejemplo, en el estudio de Rojano y colaboradores [60] se predijeron anotaciones funcionales de dominios de proteínas en una red tripartita de dominios-proteínas-funciones mediante el cálculo de asociación basado en una distribución hipergeométrica [67].

El análisis de redes también se puede aplicar a datos de expresión. Un ejemplo es el cálculo de la coexpresión con el análisis de redes ponderadas de correlación (*WGCNA*, del inglés *Weighted Correlation Network Analysis*) [55]. Este análisis parte de que un grupo de genes regulados por los mismos mecanismos deben cambiar su expresión de manera similar y se basa en el cálculo de correlación. A partir de la tabla de conteos, este algoritmo calcula la correlación de cada pareja de genes para establecer una red de correlación. Dado que la correlación presenta valores de R desde -1 a 1, por conveniencia, se usa el valor absoluto de R como valor de similitud para los cálculos posteriores y el signo como el sentido de dicha similitud. Posteriormente, los valores de similitud se elevan a una potencia para reducir los valores más bajos a practicamente cero y transformar la red para que presente una estructura libre de escala. A esta red se le aplica un algoritmo de agrupamiento jerárquico para definir grupos de genes coexpresados, comúnmente llamados módulos de coexpresión. De cada módulo se calcula un perfil de expresión representativo

que puede ser los valores de expresión del gen más conectado del módulo (gen *hub*) o el autovector del primer componente principal del módulo [55].

El análisis de datos desde el punto de vista de redes es una potente herramienta para procesar grandes conjuntos de datos y predecir interacciones entre los mismos, campo en el que el grupo de investigación tiene una larga experiencia en la aplicación de este tipo de algoritmos [57, 58, 59, 60].

## 2.7. Enriquecimiento funcional de genes

Los análisis, que se han descrito hasta ahora están centrados en producir conjuntos de genes que deben ser estudiados posteriormente. Por ejemplo, el análisis de expresión génica diferencial resulta en un conjunto de GEDs, el análisis de redes de interacción de proteínas pueden generar agrupaciones de genes y el análisis de redes de coexpresión produce módulos de genes coexpresados. Estos conjuntos de genes pueden estudiarse de manera exhaustiva mediante una búsqueda bibliográfica exponiendo las funciones moleculares que realiza cada gen y argumentando la interacción entre los mismos. Sin embargo, este tipo de estudio es laborioso y está sujeto a la interpretación del investigador.

Los análisis de enriquecimiento funcional son aproximaciones estadísticas que relacionan conjuntos de genes con nomenclaturas que permiten interpretar los resultados mediante la aplicación de algún tipo de test estadístico asociado. Las nomenclaturas son anotaciones de genes con otro tipo de información relevante. Estas nomenclaturas pueden ser: ontologías como la *HPO*, que contiene información de genes que afectan a ciertos fenotipos y la *Gene Ontology* (*GO*) [68], cuyos términos son funciones moleculares, biológicas o localizaciones celulares de los genes; bases de datos como *Reactome Pathway Database* [69], *Wikipathways* [70] o *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) [71] que contienen información sobre las rutas metabólicas y de señalización en la que participan los genes, entre otras. En resumen, las nomenclaturas contienen parejas gen-anotación y las anotaciones, ya sean fenotipos, funciones o rutas metabólicas pueden ser vistas como grupos de genes. Por tanto, las asociaciones estadísticas entre los conjuntos de genes de estudio y las anotaciones pueden calcularse mediante métodos de asociación de conjuntos.

La elección del método de asociación depende en gran medida de si los genes distribuidos en conjuntos tienen asociado un peso cuantitativo, como puede ser el $log_2FC$ de los genes con expresión diferencial o el grado del gen en una red de interacción. El método más extendido es el análisis de sobrerrepresentación que aplica una prueba estadística de una cola basado en una distribución hipergeométrica (también llamado prueba exacta de Fisher), pero este método no tienen en cuenta los pesos de los genes [72]. Por otro lado existen métodos capa-

ces de asociar conjuntos de genes teniendo en cuenta los pesos como puede ser la $t$ de Student que compara los valores de dos conjuntos de genes o el análisis de enriquecimiento de conjuntos de genes (*GSEA* del inglés *Gene Set Enrichment Analysis*) [73, 74] que aplica una prueba de Kolmogorov-Smirnov al orden de los genes según sus pesos. Además hay estrategias para aplicar estos algoritmos teniendo en cuenta la estructura ontológica de nomenclaturas como *GO* [72, 75, 76]. Todos estos análisis estadísticos devuelven una probabilidad que permite filtrar las asociaciones según su significancia estadística, sin embargo esta probabilidad debe ser corregida previamente, debido a las pruebas múltiples, con métodos como el de Benjamini-Hochberg [77] o Bonferroni [78].

En definitiva, la aplicación de los distintos métodos de enriquecimiento funcional dota de significancia estadística al estudio de las funciones de los conjuntos de genes y facilita la interpretación biológica de dichas agrupaciones.

# Capítulo 3

# Hipótesis y objetivos

Las enfermedades raras tienen un impacto significativo en la sociedad. La investigación en este ámbito resulta esencial para mitigar dicho impacto, promoviendo además avances en el conocimiento biológico y en la práctica clínica. En la actualidad, es posible acceder a datos fenotípicos de pacientes y enfermedades a partir de diversas fuentes, complementados por tecnologías como el RNA-seq, que generan grandes volúmenes de datos. No obstante, el análisis efectivo de estos datos requiere el desarrollo y aplicación de metodologías computacionales capaces de extraer información biológica relevante. En este contexto, el presente trabajo se estructura en función de las siguientes hipótesis:

**Hipótesis I** La calidad del fenotipado de los pacientes de enfermedades raras es esencial para poder relacionarlos correctamente entre sí o con el conocimiento ya existente de otras enfermedades.

**Hipótesis II** El estudio de redes de relaciones fenotípicas entre enfermedades permite identificar subgrupos que son similares a nivel fenotípico y su relación con genes causales, lo que ayuda al estudio de mecanismos moleculares implicados en estos subconjuntos de manera integrada.

**Hipótesis III** La aplicación del análisis de redes de correlación y de diferentes algoritmos de expresión diferencial sobre datos de transcriptómica permite encontrar módulos de genes coexpresados y genes diferencialmente expresados en experimentos con muestras relacionadas con el estudio de enfermedades raras.

**Hipótesis IV** Los ARN no codificantes tienen un papel fundamental dentro de las redes de regulación génicas y consecuentemente son un punto clave en el estudio de las enfermedades raras.

**Hipótesis V** La aplicación de métodos de reducción de la dimensionalidad sobre datos de expresión de genes permite detectar patrones de expresión ligados a factores experimentales que pueden contribuir a explicar los mecanismos moleculares de la enfermedad.

**Hipótesis VI** La integración de datos de expresión génica y datos fenotípicos mediante técnicas multivariante es eficaz para inferir conocimiento sobre los mecanismos de la enfermedad, permitiendo identificar patrones que contribuyen a una mejor comprensión de su patogénesis.

Conforme a estas hipótesis, el objetivo principal de esta tesis doctoral es el estudio integrado de datos fenotípicos y de expresión génica de enfermedades raras mediante técnicas de análisis de expresión y modelos de red para ayudar a su caracterización y tratamiento. Concretamente, este objetivo principal se desglosa de la siguiente manera:

**Objetivo I** Llevar a cabo análisis a nivel fenotípico de pacientes y enfermedades para estudiar la calidad del fenotipado, establecer agrupaciones y usarlas para caracterizar mecanismos moleculares comunes implicados en el desarrollo de la enfermedad.

**Objetivo II** Integrar herramientas de análisis de expresión diferencial y redes de coexpresión en un flujo de trabajo para el estudio de datos de expresión de enfermedades raras.

**Objetivo III** Relacionar ARN no codificantes, tales como micro ARN y circulares, con sus genes diana mediante estrategias de correlación, usando los datos de secuenciación de ARN de distintas enfermedades raras para encontrar posibles biomarcadores o dianas terapéuticas.

**Objetivo IV** Estudiar la expresión génica de enfermedades raras mediante Análisis de Componentes Principales para localizar patrones de expresión relacionados con factores experimentales que faciliten la comprensión de su desarrollo.

**Objetivo V** Estudiar e integrar los datos fenotípicos y de expresión génica de PMM2-CDG mediante la aplicación de análisis de reducción de la dimensionalidad para relacionar estos datos con las escalas de gravedad tales como ICARS, NCPRS o MVRD usadas en esta enfermedad e identificar genes o fenotipos asociados a dicha gravedad.

# Parte II

# Estudio de cohortes de pacientes y grupos de enfermedades raras mediante análisis de red

# Capítulo 4

# Estudio de cohortes pacientes con enfermedades genéticas mediante el uso de la *Human Phenotype Ontology*

Cuando se trabaja con datos de expresión o genómica de pacientes con enfermedades genéticas es necesario un fenotipado completo y profundo de los mismos para relacionar correctamente las variantes genéticas asociadas o los cambios de expresión a los fenotipos patológicos tal y como se describe en la Sección 2.2. Por tanto, precediendo al uso de cohortes de pacientes con enfermedades genéticas, se requiere de una evaluación tanto de la calidad del fenotipado de dichos pacientes cómo de la homogeneidad de los datos genómicos.

## 4.1. Evaluación, filtrado y agrupamiento de cohortes de enfermedades genéticas basado en datos de la *Human Phenotype Ontology* con *Cohort Analyzer*

En el marco de esta tesis doctoral he participado en el desarrollo de *Cohort Analyzer*, una herramienta que permite i) evaluar la profundidad del fenotipado de una cohorte en base a la *Human Phenotype Ontology* (*HPO*) [12], ii) inspeccionar la cobertura de los datos genómicos de los pacientes y, además, iii) agrupar los pacientes en función de la similitud semántica de sus fenotipos patológicos.

La aplicación de esta herramienta a tres cohortes de pacientes de enfermedades

genéticas dió lugar a la publicación "*Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer*". Mi contribución a este trabajo consistió en desarrollar y aplicar un flujo de trabajo que usa *Cohort Analyzer* para analizar las cohortes de 1) DECIPHER, una base de datos de pacientes de distintos centros y países con enfermedades genéticas heterogéneas [79], 2) ID/MCA, una cohorte con pacientes con diversas anomalías congénitas fenotipados con los términos de la *HPO Intelectual disability* y *Developmental delay*) [80], y 3) una cohorte de 27 pacientes diagnosticados con un desorden de la N-glicosilación derivado de mutaciones en la fosfomanomutasa 2 (PMM2-CDG). Este flujo de trabajo i) aplica *Cohort Analyzer* para evaluar el fenotipado de las cohortes completas, ii) elimina los pacientes con menos de 3 fenotipos, iii) evalúa el fenotipado de las cohortes filtradas y realiza un agrupamiento basado en la similitud semántica de los fenotipos de los pacientes al aplicar de nuevo *Cohort Analyzer*, y por último, iv) realiza un análisis de enriquecimiento funcional en las regiones genómicas afectadas en los pacientes de cada grupo para estudiar los mecanismos moleculares asociados. Para realizar los enriquecimientos en varios grupos de genes desarrollé el programa *clusters_to_enrichments.R* dentro de *ExpHunterSuite* [81].

**Artículo:** Rojano E, Córdoba-Caballero J, Jabato FM, Gallego D, Serrano M, Pérez B, Parés-Aguilar Á, Perkins JR*, Ranea JAG†, Seoane-Zonjic P†. Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer. *J Pers Med*. 2021 Jul 27;11(8):730. doi: 10.3390/jpm11080730.[1]

[1] * corresponding author; † equal contribution.

# Capítulo 5

# Estudio de las enfermedades raras usando redes de similitud semántica basada en fenotipos y técnicas de agrupamiento aplicado a enfermedades relacionadas con angiogénesis

Como se ha desarrollado en el Capítulo 4 los perfiles fenotípicos en términos de la *HPO* permite medir asociaciones entre los pacientes de una cohorte mediante el cálculo de la similitud semántica y estas asociaciones se pueden usar para localizar grupos de pacientes. Este mismo procecimiento se puede aplicar a los perfiles fenotípicos de un conjunto de enfermedades raras. El agrupamiento de enfermedades por su parecido fenotípico, la integración de distintas fuentes de datos y su análisis desde el punto de vista de redes permiten el estudio de los mecanismos moleculares y fisiológicos subyacentes.

## 5.1. Aplicación de la similitud semántica de fenotipos y análisis de redes para la generación de conocimiento en enfermedades raras asociadas a angiogénesis

La necesidad de un estudio riguroso sobre el conjunto de enfermedades raras asociadas a la angiogénesis (ER-A) dentro del grupo de investigación, el cual cuenta con una larga trayectoria en estudios de este tipo [57, 58, 59, 60], dio lugar a la publicación del trabajo *"Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis"*. En dicho trabajo, se actualizó la lista de ER-A siguiendo el protocolo de un estudio previo del grupo [82], y se realizó un análisis sistemático de las ER-A basado en la similitud semántica de los fenotipos. Este estudio ha permitido encontrar subgrupos de enfermedades ligadas a angiogénesis cuyos genes asociados comparten funciones biológicas y ha permitido proponer genes asociados candidatos.

Mi contribución en este trabajo consistió en la participación en el desarrollo y aplicación de un flujo de análisis de las ER-A encontradas en la bibliografía. Dicho flujo de trabajo incluye i) la extracción de los fenotipos de las ER-A mediante sus códigos de *Orphanet* [82] y los genes asociados a fenotipos desde *Monarch Initiative* [83], ii) la aplicación de *Cohort Analyzer* (descrito en el Capítulo 4) para generar grupos de ER-A, y iii) el análisis de los grupos de genes asociados a ER-A, incluyendo sus genes en la red de *STRINGDB* [84].

**Artículo:** Pagano-Márquez R, Córdoba-Caballero J, Martínez-Poveda B, Quesada AR, Rojano E*, Seoane P*, Ranea JAG†, Ángel Medina M†. Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis. *Brief Bioinform.* 2022 Jul 18;23(4):bbac220. doi: 10.1093/bib/bbac220.[1]

[1] * corresponding author; † equal contribution.

# Parte III

# Análisis de datos de transcriptómica en enfermedades raras

# Capítulo 6

# Análisis integrado de la expresión génica en *ExpHunterSuite* y su aplicación a enfermedades raras.

Actualmente, el análisis de expresión génica a partir de datos de RNA-seq es muy útil para generar nuevas hipótesis en el estudio de enfermedades. Debido a la heterogeneidad de las muestras de pacientes y a la poca cantidad de muestras que derivan de estudios clínicos de enfermedades raras se requiere de un análisis riguroso, combinando varias aproximaciones, para determinar genes relevantes para el proceso de enfermedad.

## 6.1. Integración de la expresión de genes y detección de módulos de coexpresión aplicado a desórdenes del metabolismo de glúcidos usando ExpHunterSuite

Dada la necesidad del análisis de datos de secuenciación de ARN mensajeros en varias enfermedades raras, participé en el desarrollo conjunto de *ExpHunterSuite*. Esta herramienta consiste en un paquete del lenguaje de programación *R*, alojado en *Bioconductor*, que se puede usar tanto desde la línea de comandos como de la propia consola de *R*. Este paquete incluye varias funcionalidades para el análisis de datos de expresión como lo son los análisis de expresión diferencial, coexpresión y enriquecimiento funcional de genes.

La aplicación de esta herramienta a dos enfermedades raras como PMM2-CDG y la enfermedad de Lafora, nos ha permitido demostrar que la integración de

los resultados de varios algoritmos de expresión diferencial tiene ventaja sobre la aplicación independiente de estos algoritmos.

Este estudio dió lugar a la publicación "*Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite*" en el cual he contribuido en el desarrollo y mantenimiento del paquete *ExpHunterSuite* y su aplicación a los conjuntos de datos de PMM2-CDG y de la enfermedad de Lafora.

**Artículo:** Jabato FM, Córdoba-Caballero J, Rojano E, Romá-Mateo C, Sanz P, Pérez B, Gallego D, Seoane P*, Ranea JAG, Perkins JR. Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite. *Sci Rep.* 2021 Jul 23;11(1):15062. doi: 10.1038/s41598-021-94343-w.[1]

[1] * corresponding author.

# Capítulo 7

# Integración de estrategias de cálculo de correlación para buscar interacciones miARN-gen diana en datos de secuenciación y su aplicación a enfermedades raras

Los miARN, como se ha descrito en la Sección 2.4, son moléculas que añaden una capa más de complejidad a la regulación post-transcripcional. Se ha observado que las desregulaciones de miARN pueden tener un impacto esencial en ciertas enfermedades raras [41, 39], lo que provoca interés en la comunidad científica por desvelar el papel regulatorio de los miARN a través del efecto que éstos tienen en sus genes diana. Hay numerosas formas de buscar las relaciones miARN-gen diana a partir de datos de secuenciación de ARN, pero los métodos de correlación miden los efectos en la expresión de las dianas causados por la regulación de un miARN. Sin embargo, hay múltiples formas de abordar este problema mediante correlación [85, 86, 87, 88] y no se ha encontrado ningún criterio sistemático para evaluarlas.

## 7.1. Exploración de la detección de parejas miARN-gen diana en enfermedades con coRmiT

Con el fin de abordar el análisis de miARN, se desarrolló un flujo de trabajo para identificar y cuantificar miARN a partir de datos de secuenciación de ARN pequeños. Este mismo flujo es capaz de analizar la expresión, por separado, de miARN y genes gracias a la herramienta *ExpHunterSuite* (Capítulo 6).

En cuanto a la búsqueda de los genes diana de miARN (miARN-diana), se encontraron diversas estrategias de asociación en la bibliografía. Ante la falta de criterio a la hora de escoger el método de asociación adecuado, se implementaron todos los métodos de correlación en *coRmiT*, un programa incluido en *ExpHunter Suite* [89]. Este programa evalúa las estrategias para cada miARN calculando la razón de probabilidades o *odds ratio* entre las parejas correlacionadas y un conjunto de validación, que por defecto es la base de datos de *multiMiR* [90]. Esta medida nos permitió seleccionar las parejas óptimas miARN-diana para tres experimentos de enfermedades raras como PMM2-CDG, la enfermedad de Lafora y la cardiomiopatía dilatada asociada a lamina. Tanto la descripción y evaluación de *coRmiT* como el estudio de las parejas encontradas en las distintas enfermedades raras se publicaron en el trabajo "*Exploring miARN–target gene pair detection in disease with coRmiT*". Mi participación en dicho estudio consitió en el co-desarrollo del flujo de trabajo para analizar miARN y genes, el diseño y desarrollo de *coRmiT* dentro del paquete de *Bioconductor ExpHunterSuite*[81], el análisis de los distintos conjuntos de datos de enfermedades raras y la evaluación computacional de los resultados de *coRmiT*. Gracias a este estudio se ha podido proponer biomarcadores y dianas terapéuticas para las distintas enfermedades raras analizadas y se ha descrito la presencia de una vía de regulación postranscripcional no canónica del miR-155 en la enfermedad de Lafora.

**Artículo:** Cordoba-Caballero J, Perkins JR*, García-Criado F, Gallego D, Navarro-Sánchez A, Moreno-Estellés M, Garcés C, Bonet F, Romá-Mateo C, Toro R, Perez B, Sanz P, Kohl M, Rojano E, Seoane P, Ranea JAG. Exploring miRNA-target gene pair detection in disease with coRmiT. *Brief Bioinform.* 2024 Jan 22;25(2):bbae060. doi: 10.1093/bib/bbae060.[1]

[1] * corresponding author.

# Capítulo 8

# Aplicación de coRmiT a la miocardiopatía dilatada por mutación de Lamina

El análisis de las relaciones entre los miARN y sus genes diana puede ser de utilidad para determinar mecanismos moleculares relevantes para el desarrollo de la enfermedad como se detalló en el Capítulo 7. Además, el análisis en detalle de estas interacciones y las funciones afectadas pueden usarse para estudiar la validez de un modelo *in vivo* para distintas enfermedades como es el caso de la miocardiopatía dilatada por mutación del gen de la Lamina (*Lmna*).

## 8.1. Una aproximación integrada para la identificación de nuevas redes de interacción miRNA-ARNm en cardiomiopatias-LMNA

Este artículo es una exploración exhaustiva de la transcriptómica de la cepa de ratones $Lmna^{R249W}$ demostrando su idoneidad cómo modelo para el estudio de la miocardiopatía dilatada asociada a Lamina (LMNA-DCM).

Para este estudio, a los datos de expresión de genes y miARN descritos en el Capítulo 7 se aplicó *coRmiT* con un umbral de correlación fijo (R de Pearson < -0,8) para seleccionar las parejas obtenidas por la estrategia con la mayor razón de probabilidades (*Odds ratio*) general. Se seleccionaron 2197 parejas miARN-diana validadas y mediante el enriquecimiento funcional de los genes, se estudiaron las funciones afectadas por cada miARN. En base a los resultados obtenidos en dicho enriquecimiento, se seleccionaron siete miARN para medir su expresión mediante reacción en cadena de la polimerasa a tiempo real (*RT-qPCR*), de los cuales se

consiguió validar la infraexpresión de los miR-133a-5p, miR-139-5p, miR-149-5p, miR-155-5p y miR-196-5p en los ratones defectivos para *Lmna*. Las dianas validadas relacionadas con estos miARN cumplen funciones en el desarrollo del corazón, la contracción del músculo cardíaco, la $\beta$-oxidación de los ácidos grasos, la adhesión celular y la unión a calcio, entre otras. Estas funciones suelen estar desreguladas en pacientes con LMNA-DCM y en otras cepas de ratones usadas como modelo de dicha enfermedad. Esto prueba, además, que los miARN identificados con *coRmiT* son susceptibles de ser biomarcadores o dianas terapéuticas.

Mi contribución a este trabajo consistió en el análisis de la expresión de miARN y genes, la aplicación de *coRmiT* para encontrar las parejas miARN-diana y, finalmente, el enriquecimiento funcional tanto de los genes con cambios de expresión como de las parejas miARN-diana.

**Artículo:** José Córdoba-Caballero, Fernando Bonet, Oscar Campuzano, Georgia Brugada-Sarquella, Ignacio Pérez de Castro Insua, Borja Vilaplana-Martí, Pedro Seoane-Zonjic, Alipio Mangas, Juan A. G. Ranea, Rocio Toro. An integrative approach to identify novel miRNA-mRNA interaction networks in LMNA-cardiomyopathy.
- En revisión

## Title

# An integrative approach to identify novel miRNA-mRNA interaction networks in *LMNA*-cardiomyopathy

## Short title

RNA interactome in lamin cardiomyopathy


## Authors

*José Córdoba-Caballero[1,2], *Fernando Bonet[1], Oscar Campuzano[3,4,5], Georgia Brugada-Sarquella[3,6,7], Ignacio Pérez de Castro Insua[8], Borja Vilaplana-Martí[9], Pedro Seoane-Zonjic[2], Alipio Mangas[1,10,11], &Juan A. G. Ranea[2,12,13,14], &Rocio Toro[1]

*These authors are the co-first authors.

&These authors are the co-senior authors.

## Affiliations

[1].- Biomedical Research and Innovation Institute of Cadiz (INiBICA), Research Unit, Puerta del Mar University Hospital, Cádiz, Spain.

[2].- Department of Molecular Biology and Biochemistry, University of Málaga, Málaga, Spain.

[3].- Medical Science Department, School of Medicine, University of Girona, Girona, Spain.

[4].- Institut d'Investigació Biomèdica de Girona (IDIBGI-CERCA), Salt, Spain.

[5].- Centro Investigación Biomédica en Red, Enfermedades Cardiovasculares (CIBERCV), Madrid, Spain.

[6].- Pediatric Arrhythmias, Inherited Cardiac Diseases and Sudden Death Unit, Cardiology Department, Sant Joan de Déu Hospital, Barcelona, Spain.

[7].- Arrítmies Pediàtriques, Cardiologia Genètica i Mort Sobtada, Malalties Cardiovasculars en el Desenvolupament, Institut de Recerca Sant Joan de Déu, Barcelona, Spain.

[8].- Center for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, Madrid, Spain.

[9].- Servicio Cirugía Experimental, Instituto Aragonés Ciencias de la Salud (IACS), Spain.

[10].- Medicine Department, Medical School of Cádiz, Cádiz University, Cádiz, Spain.

[11].- Internal Medicine Department, Puerta del Mar University Hospital, Cádiz Spain.

[12].- Institute of Biomedical Research in Málaga (IBIMA Plataforma BIONAND), Málaga, Spain.

[13.-] Center for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, U741, Málaga, Spain.

[14].- Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Instituto de Salud Carlos III (ISCIII), Madrid, Spain.

## Correspondence

Dra. Rocío Toro, MD, PhD
ORCID Dra Rocío Toro: 0000-0003-3136-1776
Dr Oscar Campuzano, BSc, MSc, PhD
ORCID Dr Campuzano: 0000-0001-5298-5276
Medicine Department, School of Medicine, University of Cadiz, 11003 Cádiz, Spain.
rocio.toro@uca.es / oscar@brugada.org



## Abstract

**Aims.** Dilated cardiomyopathy caused by variants in the *LMNA* gene leads to malignant arrhythmogenic events, faster phenotype progression and high risk of sudden cardiac death. The pathophysiological mechanisms triggering disease progression remains poorly understood.

**Methods and results.** We investigated the mRNA and miRNA transcriptome in the myocardial tissue of 50-week-old *LMNA*$^{R249W}$ mice developing dilated cardiomyopathy. We found 2148 genes and 53 miRNAs that were differentially expressed in *LMNA*$^{R249W}$ hearts. Gene ontology and pathway enrichments showed that differentially expressed genes were enriched mainly for fatty acid metabolism, muscle contraction, cell adhesion and dilated cardiomyopathy pathways. The miRNA-mRNA interactions analysis identified 2197 miRNA-target pairs with an anti-correlation between differentially expressed genes and miRNAs. Gene ontology and pathway enrichments revealed that the most significant functions of miRNA targets are mainly related to heart development, cardiac muscle contraction, fatty acid β-oxidation, cell adhesion and calcium binding pathways, among others.

**Conclusion.** Our study provides new insights into the molecular mechanisms that determine dilated cardiomyopathy due to pathogenic variants in the *LMNA* gene, and identified several target pairs that are of potential interest for further studies.

**Keywords**: *LMNA*-related dilated cardiomyopathy*;* microRNA; mRNA; RNA-sequencing; miRNA-gene interaction, biological pathways.

## Translational Perspective

Dilated cardiomyopathy (DCM) is a malignant arrhythmogenic entity leading to heart transplant. Aggressive arrythmias leading to sudden cardiac death are usual in DCM patients due to pathogenic *LMNA* variants (*LMNA*-DCM). Molecular and cellular pathophysiological pathways remain to clarify. We performed a comprehensive analysis of the miRNA-mRNA interactome to identify molecular mechanisms regulating the pathogenesis of *LMNA*-DCM. We identified more than 2000 miRNA-target pairs directly associated with aggressive arrythmias and which should be comprehensively analyzed as potential therapeutic targets in prevention of lethal arrythmias in DCM patients.



## 1. Introduction

Dilated cardiomyopathy (DCM) is a cardiac disorder that represents the major cause of heart failure and cardiac transplantation worldwide [1, 2]. The disease is characterized by left ventricle enlargement and impair systolic function [3]. The natural history of DCM is determined by the heterogeneous etiology, due to genetic and non-genetic causes [4]. Nowadays, about one-third to one-half of patients with idiopathic DCM show a genetic origin [5]. Pathogenic alterations in more than 30 genes have been linked to the genetic form of DCM. Among them, deleterious variants in the *LMNA* gene are responsible for about 10% of cases mainly following a dominant pattern of inheritance [6-8]. The *LMNA* gene encodes nuclear lamin A and C, intermediate filament proteins that are critically important for the structural properties of the nucleus, as they provide nuclear shape and mechanical stability. Several studies have also supported a role in DNA replication, gene expression, chromatin organization, and cell cycle progression [9]. A significant number of DCM patients due to pathogenic *LMNA* variants (*LMNA*-DCM) display complications only in the cardiovascular system [10]. Furthermore, most of reported *LMNA*-DCM patients display a more aggressive arrhythmogenic behavior, thromboembolisms with fatal outcomes, and faster phenotype progression as compared to other forms of DCM [11]. In addition, the cardiac dysfunction frequently preceded by the conduction system disease and/or arrhythmia leading to great concern to cardiologist [12]. Both incomplete penetrance and variable expressivity, influenced by external factors, difficult the risk stratification of *LMNA*-DCM patients. The current medical treatment for *LMNA*-DCM follows the standard heart failure management recommendations involving pharmaceutical and non-pharmaceutical therapies with a suboptimal result. Hence, there is an urgency in understanding the pathophysiological mechanism that underpin this malignant entity to shed light on novel treatment approaches [13, 14].

MicroRNA (miRNA) are single-stranded non-coding RNA (ncRNA) of ~22 nucleotides that repress gene expression by binding to complementary sequences in the 3' untranslated region (3' UTR) of mRNAs to target them for degradation or transcription suppression [15]. A single miRNA can target multiple genes and a single gene can be targeted by many miRNAs [16]. Numerous studies have demonstrated critical functions of miRNAs in the progression of various DCM etiologies [17]. However, the role of miRNAs in the onset, progression and outcome of *LMNA*-DCM remains to be deeply analyzed so far. Although prior RNA-sequencing (RNA-seq) studies have evidenced different gene expression profile in distinct *Lmna* mutant mouse lines as compared to wild-type, to date no miRNA sequencing studies on *Lmna* mutants have been conducted, with only a brief analysis being reported [18]. The goal of this study was to use screen critical miRNA-gene target pairs to elucidate the mechanisms of miRNA regulation in the progression of *LMNA*-DCM. For this purpose, we used high-throughput mRNA and miRNA-seq of myocardial tissue from 50-week-old wild-type and a *Lmna* mutant mouse line, namely *LMNA*$^{R249W}$, characterized by a unique DCM phenotype. Then, we performed a comprehensive analysis of the miRNA-mRNA interactome to identify important molecular mechanisms regulating the pathogenesis of *LMNA*-DCM.



## 2. Methods

### 2.1 Mice

The murine model used in our study carries the widely reported pathogenic variant p.Arg249Trp (p.R249W, c.745C>T, rs121912496) in the *LMNA* gene, the most prevalent in *LMNA*-related congenital muscular dystrophy (L-CMD) [19]. Our mouse model (*Lmna*$^{R249W}$) recapitulates many of the features of L-CMD and its detailed characterization will be published elsewhere. We analyzed both mRNA and miRNA expression profiles by RNA-seq in myocardial tissue of 50-week-old *Lmna*$^{R249W}$ (n = 6) and age-matched wild-typed type mice (n = 6). Mice were housed at the specific pathogen-free barrier area of the Instituto de Salud Carlos III (ISCIII) (Madrid, Spain). Mice were observed daily and sacrificed when they showed overt signs of morbidity in accordance with the Guidelines for Humane Endpoints for Animals Used in Biomedical Research from the Council for International Organizations of Medical Sciences (CIOMS). All animal procedures were performed according to the procedures (PROEX 164-18) approved by the ISCIII Ethics Committee for Research and Animal Welfare (CEIyBA) and are conformed to the guidelines from Directive 2010/63/EU of the European Parliament on the protection of animals used for scientific purposes. No anaesthetic was used in this work. The euthanasia was performed by inhalation of CO2 in a specially designed chamber available in our animal facility. This study was approved by the Andalusian Biomedical Research Ethics committee (CEICA, reference number PI17/0023). The ethical research principles were fulfilled following the Helsinki Declaration and the Belmont report. The study also adhered to two legal provisions governing human research and the Spanish Organic Law 15/1999 for the Regulation of Automated Processing of Personal Data.

### 2.2 RNA isolation and qRT-PCR

Total RNA was extracted and purified by using TRI Reagent (Sigma-Aldrich, St Louis, MO, USA) according to the manufacturer's instructions followed by DNase treatment and purification using RNA clean and concentrator-5 kit (Zymo Research, Irvine, CA, USA). RNA was quantified using a Qubit RNA High-Sensitivity Assay kit in the Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). For miRNA expression analysis, 5 ng of RNA was reversed transcribed using miRCURY LNA RT Kit (Qiagen, Hilden, Germany) according to manufacturer´s instruction. qRT-PCR was performed on a CFX96 Real-Time PCR system (Bio-Rad) using the miRCURY LNA SYBR Green PCR Kit (Qiagen, Hilden, Germany), expression was normalized against U6 and 5s, and data was analyzed using the $2^{-\Delta\Delta Ct}$ algorithm.

### 2.3 RNA sequencing

The quality and integrity of total RNA were controlled on the Agilent Technologies 2100 Bioanalyzer. Standard-specific mRNA-seq libraries were generated using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA), and single-end libraries were sequenced on an Illumina PE75 Platform with an output of 70.7 M reads per sample. Standard miRNA libraries were generated using the NEXTFLEX small RNA-seq kit v3 (Perkin Elmer, Waltham, MA, USA), and single-end libraries were sequenced on an Illumina SE75 Platform with an output of 22.1 M reads per sample.

### 2.4 Expression and miRNA-gene target network analysis

RNA-seq and small RNA-seq samples were separately quantified using DEG_workflow [18]. The analysis of the differential expression and the co-expression were analyzed in RNA and miRNA counts tables using degenes_hunter. to get the DEMs using the 4 implemented differential expression algorithms and WGCNA, the co-expressed miRNA modules, the DEGs and the co-expressed genes modules. The overrepresentation analysis of Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome pathways in DEGs was analyzed using functional_hunter.R. The results of degenes_hunter.R for miRNA and mRNA were used as input of coRmiT.R [18] to detect correlated miRNA target gene pairs. A threshold of Pearson's R < -0.8 was used. Only previously validated miRNA target pairs generated by the correlation strategy with higher median odds ratio were kept. Enrichments for GO (http://geneontology.org/) terms and KEGG and Reactome pathways in miRNA-target was calculated using clusters_to_enrichment.R. Overrepresentation in categories of predicted and validated targets of each miRNA were analyzed separately using the same script. The four scripts degenes_hunter.R, functional_hunter.R, coRmiT.R and cluster_to_enrichment.R are included in the in-house R package ExpHunterSuite [20].

## 3. Results

### 3.1 Identification of DEGs and DEMs in *Lmna*$^{R249W}$ mutant hearts

To elucidate the mechanisms underlying DCM in *Lmna*$^{R249W}$ heterozygous mutant mice, we analyzed both mRNA and miRNA expression profiles by RNA-seq in myocardial tissue of 50-week-old *Lmna*$^{R249W}$ (n = 6) and age-matched wild-typed type mice (n = 6). Among 13,003 expressed genes, we identified 2,148 DEGs in *Lmna*$^{R249W}$ hearts as compared to wild type, of which 1,485 were upregulated whereas 663 were downregulated (Figure 1A and Supplementary File 1). We applied Weighted Correlation Network Analysis (WGCNA) to group the expressed genes in co-expression modules and modules 1, 2 and 3 showed the highest absolute correlation (Pearson |R| of 0.96, 0,96 and 0.87 respectively) with experiment design (Supplementary File 1). A total of 677 miRNAs were quantified of which 53 were differentially expressed (21 upregulated and 32 downregulated) in *Lmna*$^{R249W}$ as compared to wild-type (Figure 1B and Supplementary File 2).

### 3.2 Functional and pathway enrichment analysis of DEGs

To better understand the mechanisms underlying *LMNA*-DCM, GO, KEGG and Reactome pathways analysis were conducted to predict the potential functions of DEGs. The most significantly enriched processes and signal pathways are shown in Figure 2 and 3. Top enriched GO terms in biological processes (BP), molecular function (MF) and cellular components (CC) (Figure 2A and B, and Supplementary File 3, respectively). The GO enriched terms were involved in fatty acid metabolic process, extracellular matrix, muscle contraction and synaptic transmission-related pathways among the main terms involved in BP (Figure 2A). At the MF, DEGs were mainly enriched for voltage-gated ion channel-related pathways (Figure 2B). Furthermore, extracellular matrix, sarcomere, synaptic transmission and ion channel-related terms of CC were significantly enriched (Supplementary File 3). Similarly, the enrichments for Reactome showed relations with lipid metabolism, extracellular matrix, muscle contraction and synaptic transmission-related pathways (Figure 3A). Finally, KEGG pathway

enrichments showed that DEGs were mainly involved in fatty acid metabolism-related pathway, cell adhesion molecules and dilated cardiomyopathy (Figure 3B).

### 3.3 miRNA expression and target gene analyses

To elucidate the regulatory role of DEMs in *Lmna*$^{R249W}$ mutant hearts, the correlation between DEMs and DEGs and their associated co-expression modules was investigated by coRmiT. The strategy that correlates the eigengene value of RNA modules with the hub gene profile of miRNA modules showed the highest coverage, where 5/14 miRNAs had significant overlap FDR < 0.05 with validated pairs databases and a median odds ratio of 1.63 (Figure 4). 108,676 miRNA-target correlated pairs were found, of which 2,197 were previously found on validation databases. These validated targets corresponded to 12 DEMs, four up-regulated (miR-183-5p, miR-690, miR-324-5p and miR-3473a) and eight down-regulated miRNAs (miR-139-5p, miR-196b-5p, miR-3473b, miR-155-5p, miR-133a-5p, miR-1224-5p, miR-3095-5p, miR-149-5p) (Table 1). The functions of validated targets were inspected using over representation analysis for GO terms. Only the targets of four upregulated (miR-183-5p, miR-3473a, miR-324-5p and miR-690) and eight downregulated miRNAs (miR-1224-p, miR-133a-5p, miR-149-5p, miR-155-5p, miR-196b-5p, miR-3095-5p, miR-3473b and miR-139-5p) showed significant enrichments for GO terms (Table 1).

The targets of upregulated miRNAs using the most enriched terms of GO molecular functions and biological processes are shown in Figure 5 and 6. The functional analysis demonstrated that the miR-183-5p targets *Lrp6, Zfmp2* and *Sos1* are involved in crucial processes of the developing heart including mitral valve, ventricular septum, cardiac atrium and embryonic heart tube morphogenesis, and pericardium development (Figure 5). Whereas the miR-183-5p target Lrp6 was related to the Wnt-signaling pathway and apolipoprotein binding, *Taok1* and *Dst* were associated with myosin V-, α-tubulin- and β-tubulin-binding (Figure 6). Similarly to miR-183-5p, the miR-324-5p targets *Akap1* and *Dip2b* affecting α-tubulin and β-tubulin binding. Other miR-324-5p targets such as *Arhgap35*, *Naa30* and *Kmt2c/Kdm3b* were involved in GTPase binding, peptide α-N-acetyltransferase activity and histone methylation/demethylation function (Figure 5).

The miR-3473a targets were related to embryonic heart tube morphogenesis, regulation of cardiomyocyte action potential and cardiac muscle cell contraction through the *Ryr2* gene; this former shares *with* the *Kcnk3* gene the negative modulation of cytosolic calcium ion concentration. *Ryr2* and *Tead1* genes are linked to the embryonic heart tube morphogenesis and, control muscle system process. The regulation of ion transmembrane transporter activity and response to muscle activity is performed targeting *Ryr2* and *Ppargc1a* and, the energy derivation by oxidation of organic compounds through *Ppargc1a, Lyrm7* and *Nhlrc1*. The targets for miR-690 were enriched for terms affected functions as the control of muscle system process, the positive regulation of reactive oxygen species biosynthetic process, the aging, the regulation of cellular response to oxidative stress, the energy derivation by oxidation of organic compounds and, the fatty acid β-oxidation using acyl-CoA dehydrogenase; some of these functions were shared with miR-3473a, but just the *Tead1* gene (Figure 6). The miR-690 targets were also enriched for terms related to glucose binding, and acyl-CoA dehydrogenase activity (Figure 5).

Similarly, the downregulated miRNAs´ targets using the most enriched terms of GO MF and BP are presented in Figures 7 and 8. Among them, the targets of five downregulated miRNAs, miR-1224-5p, miR-133a-5p, miR-139-5p, miR-155-5p and miR-196b-5p involved the cell-substrate adhesion term, whereas cell-matrix adhesion and apoptotic cell clearance were enriched by targets of miR-1224-5p, miR-133a-5p, miR-155-5p and miR-196b-5p. miR-1224-5p, miR-155-5p and miR-196b-5p also affected the positive regulation of cell adhesion and actin filament organization cytoskeleton. In addition, miR-1224-5p and miR-155-5p targets were enriched for terms related to the regulation of actin cytoskeleton reorganization, the regulation of focal adhesion, cortical cytoskeleton organization, regulation of cell-substrate adhesion and apoptotic process involved in development. Besides, miR-155-5p and miR-196b-5p targets also were enriched for terms associated with inflammatory response, wound healing, extracellular matrix organization and muscle cell differentiation. The miR-155-5p and miR-196b-5p target genes were involved in laminin, protease, actinin filament, MHC class II and calcium binding pathways related terms. miR-155-5p together miR-149-5p targets were implicated in the Notch-signaling pathway and metallopeptidase activity regulation, and targets shared by miR-155-5p and miR-1224-5p were involved in the terms GTP binding and lipid transporter function. Finally, miR-3473b targets were associated with calcium and sodium channel activity-related pathways and muscle alpha-actinin binding.

### 3.4 Validation of DEMs by RT-qPCR

Next, we performed an experimental validation using qRT-PCR for certain DEMs, such as miR-133a-5p, miR-139-5p, miR-149-5p, miR-155-5p, miR-183-5p, miR-196b-5p and miR-324-5p. These miRNAs were selected based on their correspondence to validated targets identified through miRNA-target correlated pairs analysis, which encompassed 12 DEMs. Among these DEMs, we prioritized those for which commercial primers were available. qRT-PCR were performed in six *Lmna*$^{R249W}$ and six wild-type myocardial samples. U6 and 5s were used as internal control for miRNA qRT-PCRs. For qRT-PCR confirmation, all seven DEMs showed a strong correlation between RNA-seq and qRT-PCR (Figure 9 and Supplementary File 2).

## 4. Discussion

We have performed for the first time a comprehensive characterization of the mRNA and miRNA transcriptome of myocardial tissue from *Lmna*$^{R249W}$ mutant mice which develop DCM as an isolated phenotype. Although transcriptome works in the heart of *Lmna* mutant mice have been reported in the past [21-26], our study is the first combining small RNA and mRNA sequencing to investigate miRNA-target interactions in an *in vivo* setting of *LMNA*-DCM to unveil the molecular mechanisms behind the pathogenesis and progression of this condition. We report that *Lmna*$^{R249W}$ hearts display distinct mRNA and miRNA expression profiles with 2,148 DEGs and 53 DEMs. We found that DEGs were mainly involved in cardiac-related pathways such as extracellular matrix, muscle contraction, voltage-gated ion channel- and synaptic transmission-related pathways, cell adhesion molecules and fatty acid metabolic process. At this point, we should mention a limitation of our study concerning use of a mouse model and single timepoint.

Myocardial fibrosis, a principal adaptative response in DCM, acts through extracellular matrix changes [27]. The DEGs in *Lmna*$^{R249W}$ hearts were enriched for extracellular matrix in line with recent studies that demonstrated the upregulation of extracellular matrix genes in the heart of distinct *Lmna* mutant mouse lines [28, 29]. Extracellular proteins impairment may alter cell-to-cell interactions which impact the mechanical cues modifying signaling pathways. This mechano-transduction stimulus may lead to a maladaptive remodeling response in *LMNA*-DCM [30]. DCM drives to an abnormal hemodynamic load and mechanical stress to respond to the gene expression, protein synthesis and cardiomyocyte degradation cascade [31].

Besides ventricular dilatation and remodeling, *LMNA*-DCM is also characterized by conduction defects and life-threatening arrhythmogenic events, sometimes before any detectable left ventricular impairment [32, 33]. Voltage-gated ion channels are responsible for action potential generation and its propagation across the myocardium, mainly sodium [34]. Focusing on *LMNA*-DCM, several studies have linked *LMNA* deleterious variants to *SCN5A* regulation and Nav1.5 function [35-39]. In recent research, a deleterious variant in the *LMNA* gene was associated with hyper-polymerization and hyper-acetylation of the tubulin network with concomitant downregulation of Nav1.5 cell expression and activity, leading to disruption of electric transmission [35]. This incorrect communication between cardiomyocytes induces Cx43 remodeling in ventricles [40], and high risk of arrhythmias, hallmark in *LMNA*-DCM. In the same way, our results showed an enrichment of voltage-gated ion channel-related pathways in *Lmna*$^{R249W}$ hearts, accordingly to dysregulation of genes that encode sodium voltage-gated channel subunits in the heart from a different *Lmna* mutant mouse line [24]. On the other hand, although the role of cardiac neurotransmitter system in DCM has not been fully investigated, alterations of presynaptic sympathetic innervation have been associated with idiopathic DCM [41]. Hence, our results show, for the first time, that cardiac synaptic transmission might be implicated in *LMNA*-DCM.

Desmosomes are critical adhesion structures in cardiomyocytes that mediate strong cell-cell contact [42]. While pathogenic variants encoding desmosomal proteins are considered the predominant cause of arrhythmogenic cardiomyopathy [43], pathogenic variants in the *LMNA* gene have also been associated to be possible causes of this disease, especially with severe bradyarrhythmia [44-46]. The pathophysiological mechanisms underlying the arrhythmic phenotype in *LMNA*-DCM are still not well elucidated but the structural nuclei abnormalities, chromatin modifications and the associated transcriptional changes may influence the molecular basis of the *LMNA*-related cardiac phenotypes, including arrhythmogenic events [47]. It has been suggested that the loss of A-type-lamins could have an impact on cytoskeleton organization and cell adhesion [48]. Accordingly, our analysis showed an involvement of DEG genes in cell-cell adhesion. The deregulation of key genes involved in fatty acid metabolism are related to DCM [49, 50], but its involvement in *LMNA*-DCM remains unknown. Hence, we confirm that fatty acid metabolism impairment might be implicated in the pathogenesis of *LMNA*-DCM.

Integrative analysis identified 2,197 validated interactions between 1,892 DEGs and 12 DEMs. Enrichment analysis of target genes for each upregulated miRNA showed that genes regulated by miR-183-5p highlighted heart development processes, and myosin V-, α-tubulin- and β-tubulin-binding, suggesting a previously undescribed role of miR-183-5p in the developing heart. Consistently with the enrichment for myosin V-, α-tubulin- and β-tubulin-binding, Tariq et al. suggested that appropriate

nuclear lamina organization and microtubule network are required for maintaining an adequate nuclear morphology and function [51]. In this regard, Borin et al. found that pathogenic variants in *Lmna* compromise the microtubule network in neonatal rat ventricular myocytes [52]. According to our analysis, miR-183-5p targets *Lrp6*, a gene encoding the coreceptor to Frizzled in the Wnt pathway [53], which is consistent with previous results, demonstrating that dysregulation of Wnt/β-catenin pathway and its downstream target gap junction protein connexin-43 contributes to the pathophysiology of DCM in *Lmna* mutant hearts [54-56]. Therefore, miR-183-5p might be involved in *LMNA*-related malignant arrhythmias leading to electrical conduction disturbances consequence of an incorrect communication between cardiomyocytes due to alterations in microtubule cytoskeleton, acetylation of α-tubulin and subsequent Cx43 remodeling [40]. In the same way occurs in the *Lmna*$^{tm1Stw}$ mutant mouse line, in which LRP6 deficiency led to DCM due to alterations in the autophagic degradation and fatty acid utilization pathways [21, 24].

The target genes of miR-324 were involved in histone methylation/demethylation function which is consistent with the role of Lamin A protein in the epigenomic regulation of chromatin [57]. In addition, its target gene *Atg4* is an autophagic-related gene involved in removing damaged products of the cell, and recycling proteins, glycogen, and fatty acids, thus providing energy for myofibers during stress and/or energy deprivation [58]. Hence, miR-3473a is a principal regulator of heart development and, is involved in heart failure [59]. Our analysis confirms these results. miR-3473a target genes highlighted cardiac muscle cell action potential and contraction, negative regulation of cytosolic calcium ion concentration, regulation of ion transmembrane transporter activity and, response to muscle activity via Ryanodine receptor 2 (*Ryr2*). Abnormal ryanodine function is present in inherited cardiac arrhythmias known as cardiac ryanodinopathies, mainly catecholaminergic polymorphic ventricular tachycardia [60]. Dridi et al. reported biochemical modification of RyR2 protein in the heart tissue in both human patients with *LMNA*-DCM and *Lmna* mutant mice [61]. Similarly to miR-183-5p, miR-3473a could be involved in the progression of arrhythmic events in *LMNA*-DCM. Our analysis showed the transcription factor-encoding gene *Tead1*, a critical component of the Hippo signaling pathway together with the coactivator YAP/TAZ [62], as a potential target of both miR-3473a and miR-690. Agreeing with this, *Tead1* knockout mice showed embryonic lethality and DCM [63], whereas specific ablation of *Tead1* in adult cardiomyocytes after tamoxifen induction led to lethal acute-onset DCM [64]. Hippo signaling activation/YAP-TEAD1 inactivation leads to mitochondrial damage promoting DCM [65]. More recently, single-cell RNA-seq experiments confirmed dysregulated expression of TEAD1 target genes in cardiac tissue from patients with *LMNA*-DCM but not in other DCM patients [66].

Oxidative stress has been linked strongly to cell death and cardiac remodeling processes, both leading to heart failure, being antioxidant treatment a therapeutic approach for cardiomyopathies [67]. miR-690 target genes were also involved in positive regulation of reactive oxygen species biosynthetic process and regulation of cellular response to oxidative stress targeting *Sod2*. Superoxide dismutase (SODs) is antioxidative enzymes that catalyze the degradation of reactive oxygen species. Therefore, they regulate mitochondrial superoxide generation and improve the phenotypes of the DCM and muscle fatigue in mice [68]. In line with this, alterations of SOD2 protein have been linked to DCM progression and ventricular tachycardia both in mice and humans [69, 70].

Regarding the downregulated miRNAs, our analysis showed that many of these miRNAs were enriched in linked BP and MF terms. Concretely, miR-1224-5p, miR-133a-5p, miR-139-5p, miR-155-5p and miR-196b-5p shared the cell-substrate adhesion term, whereas miR-1224-5p, miR-133a-5p, miR-155-5p and miR-196b-5p shared cell-matrix adhesion and apoptotic cell clearance. Many other terms were shared by at least two miRNAs. Prior studies have shown that a mRNA can be targeted by multiple miRNAs simultaneously and regulate the same transcript targets [71]. The synergistic effects of miRNAs are important for distinct biological processes [72-74], including key pathways related to cardiac diseases [75, 76]. miR-1224-5p was mainly involved in the regulation of BP associated with cell adhesion, cytoskeleton organization and apoptosis. Its role in the heart has yet to be explored but, miR-1224-5p has been described to suppress apoptosis and epithelial-to-mesenchymal transition via TGF-ß1/Smad3 signaling pathway; this former is a process characterized by loss of cell-cell adhesions and polarity and the reorganization of the cytoskeleton [77, 78], which might indicate a regulatory role in cell invasion and apoptosis in the heart. In addition, an impairment in the TGF-β pathway has been recently described in DCM patients [31]. Regarding miR-133a-5p, similar findings were reported in other forms of DCM as the miR-133 family was downregulated in cardiac tissue from patients with this condition [79]. Furthermore, studies have demonstrated a role for miR-133a-5p in the pathology of ischemic myocardial diseases inhibiting apoptosis, inflammation, and adverse cardiac remodeling [80], which agrees with the enrichment of the apoptotic cell clearance [81]. The role of miR-139-5p in the heart remains poorly understood; according to our results, its expression is downregulated in the myocardial tissue from patients with both ischemic and hypertrophic cardiomyopathy [82]. miR-139-5p that acts as an anti-hypertrophic miRNA attenuating cardiomyocyte enlargement is involved in *LMNA*-DCM biological process [83]. miR-149-5p in the heart has been linked *in vitro* studies to cardiomyocyte apoptosis [26]. Our results identified miR-149-5p as a potential regulator of the Notch pathway, intercellular signaling that regulates cell fate specification and organogenesis [84]. Both aberrations in Notch signaling and pathogenic variants in *LMNA* are related to left ventricle non-compaction cardiomyopathy [36, 85]. Recently, emerging data have indicated a role of Notch pathway with crosstalk with Hippo signaling in the progression of inherited DCM [86]. In the heart, miR-155-5p was upregulated in inflammatory DCM [87, 88], being critical for immune response in the myocardium [89]. *LMNA*-DCM is characterized by a dysregulated inflammatory response along the contractile and electrical impairment [90], nevertheless, our results showed downregulation of miR-155-5p in *Lmna*$^{R249W}$ hearts. Hence, further studies are needed to understand the role of miR-155-5p in this entity. Finally, to date, nothing is known about the role of miR-196b-5p. Our results show for the first time a dysregulation of these miRNAs in *LMNA*-DCM suggesting a role in heart disease.

## 5. Conclusions

We must note that the gene expression related to pathophysiological changes at different stages of *LMNA* cardiomyopathy will be performed in a next step. Further studies in larger murine cohorts should be achieved to confirm our results and therefore, translate our first *in vivo* approach to clinical studies of *LMNA*-DCM patients. In summary, our study explores, for the first time, the molecular mechanisms behind *LMNA*-DCM through the integration of mRNA and miRNA sequencing data. This integrative approach allowed us to identify novel miRNA-mRNA interaction networks and signaling pathways that

unravel cellular biological processes of *LMNA*-DCM. We suggest our biomarkers might be used as novel therapeutic targets for treating *LMNA*-related DCM.

## Supplementary material

Supplementary materials available at Cardiovascular Research online.


## Funding

This work was supported by grants in the framework of the European Regional Development Fund (ERDF) Integrated Territorial Initiative (ITI0017-2019) and Foundation Progreso y Salud PEER (2020-019). This work was supported by funds from Spanish Ministry of Economy and Competitiveness [PID2022-140047OB-C21]; the Institute of Health Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, “A way to make Europe”). This work was also supported by Bosch i Aymerich Foundation. IDIBGI and Fundació Sant Joan de Dèu are a “CERCA Programme / Generalitat de Catalunya”.


## Author’s contributions

All authors have read and approved the submission of the manuscript. Author Contributions: RT, JAGR, OC and AM conceived the experiments; JCC, FBM and IPdCI recruited the subjects. JCC, FBM and IPdCI conducted the experiments, and IPdCI, BVM and PSZ analyzed the results. JCC, FBM, GSB, OC, and RT wrote the manuscript. All authors reviewed the manuscript.


## Acknowledgments

We thank the European Molecular Biology Laboratory GeneCore, Genomics Core Facility (Heidelberg, Germany) staff for processing our samples.


## Conflicts of interest

We know of no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome.

## Data availability

Data transparency is guaranteed. The datasets generated during and/or analyzed during the current study are available in the supplemental material.

## References


1. Merlo M, Stolfo D, Anzini M, Negri F, Pinamonti B, Barbati G, Ramani F, Lenarda AD and Sinagra G. Persistent recovery of normal left ventricular function and dimension in idiopathic dilated cardiomyopathy during long-term follow-up: does real healing exist? *Journal of the American Heart Association*. 2015;4:e001504.
2. Stehlik J, Edwards LB, Kucheryavaya AY, Benden C, Christie JD, Dobbels F, Kirk R, Rahmel AO and Hertz MI. The Registry of the International Society for Heart and Lung Transplantation: Twenty-eighth Adult Heart Transplant Report--2011. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*. 2011;30:1078-94.
3. Pinto YM, Elliott PM, Arbustini E, Adler Y, Anastasakis A, Bohm M, Duboc D, Gimeno J, de Groote P, Imazio M, Heymans S, Klingel K, Komajda M, Limongelli G, Linhart A, Mogensen J, Moon J, Pieper PG, Seferovic PM, Schueler S, Zamorano JL, Caforio AL and Charron P. Proposal for a revised definition of dilated cardiomyopathy, hypokinetic non-dilated cardiomyopathy, and its implications for clinical practice: a position statement of the ESC working group on myocardial and pericardial diseases. *Eur Heart J*. 2016;37:1850-8.
4. Weintraub RG, Semsarian C and Macdonald P. Dilated cardiomyopathy. *Lancet*. 2017;390:400-414.
5. McNally EM and Mestroni L. Dilated Cardiomyopathy: Genetic Determinants and Mechanisms. *Circ Res*. 2017;121:731-748.
6. Hershberger RE and Morales A. Dilated Cardiomyopathy Overview. In: M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. Stephens and A. Amemiya, eds. *GeneReviews((R))* Seattle (WA); 1993.
7. Parks SB, Kushner JD, Nauman D, Burgess D, Ludwigsen S, Peterson A, Li D, Jakobs P, Litt M, Porter CB, Rahko PS and Hershberger RE. Lamin A/C mutation analysis in a cohort of 324 unrelated patients with idiopathic or familial dilated cardiomyopathy. *Am Heart J*. 2008;156:161-9.
8. Tobita T, Nomura S, Fujita T, Morita H, Asano Y, Onoue K, Ito M, Imai Y, Suzuki A, Ko T, Satoh M, Fujita K, Naito AT, Furutani Y, Toko H, Harada M, Amiya E, Hatano M, Takimoto E, Shiga T, Nakanishi T, Sakata Y, Ono M, Saito Y, Takashima S, Hagiwara N, Aburatani H and Komuro I. Genetic basis of cardiomyopathy and the genotypes involved in prognosis and left ventricular reverse remodeling. *Scientific reports*. 2018;8:1998.
9. Captur G, Arbustini E, Bonne G, Syrris P, Mills K, Wahbi K, Mohiddin SA, McKenna WJ, Pettit S, Ho CY, Muchir A, Gissen P, Elliott PM and Moon JC. Lamin and the heart. *Heart*. 2018;104:468-479.
10. Malhotra R and Mason PK. Lamin A/C deficiency as a cause of familial dilated cardiomyopathy. *Curr Opin Cardiol*. 2009;24:203-8.
11. van Rijsingen IA, Nannenberg EA, Arbustini E, Elliott PM, Mogensen J, Hermans-van Ast JF, van der Kooi AJ, van Tintelen JP, van den Berg MP, Grasso M, Serio A, Jenkins S, Rowland C, Richard P, Wilde AA, Perrot A, Pankuweit S, Zwinderman AH, Charron P, Christiaans I and Pinto YM.

Gender-specific differences in major cardiac events and mortality in lamin A/C mutation carriers. *European journal of heart failure*. 2013;15:376-84.

12. Ollila L, Nikus K, Holmstrom M, Jalanko M, Jurkko R, Kaartinen M, Koskenvuo J, Kuusisto J, Karkkainen S, Palojoki E, Reissell E, Piirila P and Helio T. Clinical disease presentation and ECG characteristics of LMNA mutation carriers. *Open heart*. 2017;4:e000474.
13. Behnoush AH, Khalaji A, Naderi N, Ashraf H and von Haehling S. ACC/AHA/HFSA 2022 and ESC 2021 guidelines on heart failure comparison. *ESC heart failure*. 2023;10:1531-1544.
14. Rosario KF, Karra R, Amos K, Landstrom AP, Lakdawala NK, Brezitski K, Kim H and Devore AD. LMNA Cardiomyopathy: Important Considerations for the Heart Failure Clinician. *Journal of cardiac failure*. 2023.
15. Kreutzer FP, Fiedler J and Thum T. Non-coding RNAs: key players in cardiac disease. *J Physiol*. 2020;598:2995-3003.
16. Gebert LFR and MacRae IJ. Regulation of microRNA function in animals. *Nat Rev Mol Cell Biol*. 2019;20:21-37.
17. Alonso-Villa E, Bonet F, Hernandez-Torres F, Campuzano O, Sarquella-Brugada G, Quezada-Feijoo M, Ramos M, Mangas A and Toro R. The Role of MicroRNAs in Dilated Cardiomyopathy: New Insights for an Old Entity. *International journal of molecular sciences*. 2022;23.
18. Cordoba-Caballero J, Perkins JR, Garcia-Criado F, Gallego D, Navarro-Sanchez A, Moreno-Estelles M, Garces C, Bonet F, Roma-Mateo C, Toro R, Perez B, Sanz P, Kohl M, Rojano E, Seoane P and Ranea JAG. Exploring miRNA-target gene pair detection in disease with coRmiT. *Brief Bioinform*. 2024;25.
19. Ben Yaou R, Yun P, Dabaj I, Norato G, Donkervoort S, Xiong H, Nascimento A, Maggi L, Sarkozy A, Monges S, Bertoli M, Komaki H, Mayer M, Mercuri E, Zanoteli E, Castiglioni C, Marini-Bettolo C, D'Amico A, Deconinck N, Desguerre I, Erazo-Torricelli R, Gurgel-Giannetti J, Ishiyama A, Kleinsteuber KS, Lagrue E, Laugel V, Mercier S, Messina S, Politano L, Ryan MM, Sabouraud P, Schara U, Siciliano G, Vercelli L, Voit T, Yoon G, Alvarez R, Muntoni F, Pierson TM, Gomez-Andres D, Reghan Foley A, Quijano-Roy S, Bonnemann CG and Bonne G. International retrospective natural history study of LMNA-related congenital muscular dystrophy. *Brain Commun*. 2021;3:fcab075.
20. Jabato FM, Cordoba-Caballero J, Rojano E, Roma-Mateo C, Sanz P, Perez B, Gallego D, Seoane P, Ranea JAG and Perkins JR. Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite. *Scientific reports*. 2021;11:15062.
21. Chen SN, Lombardi R, Karmouch J, Tsai JY, Czernuszewicz G, Taylor MRG, Mestroni L, Coarfa C, Gurha P and Marian AJ. DNA Damage Response/TP53 Pathway Is Activated and Contributes to the Pathogenesis of Dilated Cardiomyopathy Associated With LMNA (Lamin A/C) Mutations. *Circ Res*. 2019;124:856-873.
22. Coste Pradas J, Auguste G, Matkovich SJ, Lombardi R, Chen SN, Garnett T, Chamberlain K, Riyad JM, Weber T, Singh SK, Robertson MJ, Coarfa C, Marian AJ and Gurha P. Identification of Genes

and Pathways Regulated by Lamin A in Heart. *Journal of the American Heart Association*. 2020;9:e015690.

23. Onoue K, Wakimoto H, Jiang J, Parfenov M, DePalma S, Conner D, Gorham J, McKean D, Seidman JG, Seidman CE and Saito Y. Cardiomyocyte Proliferative Capacity Is Restricted in Mice With Lmna Mutation. *Frontiers in cardiovascular medicine*. 2021;8:639148.
24. Shao Z, Koh W, Ni Y, Li W, Agatisa-Boyle B, Merkurjev D and Tang WHW. RNA Sequence Analyses throughout the Course of Mouse Cardiac Laminopathy Identify Differentially Expressed Genes for Cell Cycle Control and Mitochondrial Function. *Scientific reports*. 2020;10:6632.
25. Shemer Y, Mekies LN, Ben Jehuda R, Baskin P, Shulman R, Eisen B, Regev D, Arbustini E, Gerull B, Gherghiceanu M, Gottlieb E, Arad M and Binah O. Investigating LMNA-Related Dilated Cardiomyopathy Using Human Induced Pluripotent Stem Cell-Derived Cardiomyocytes. *International journal of molecular sciences*. 2021;22.
26. Zhang L, Liu T, Wang P, Shen Y and Huang T. Overexpression of Long Noncoding RNA H19 Inhibits Cardiomyocyte Apoptosis in Neonatal Rats with Hypoxic-Ischemic Brain Damage Through the miR-149-5p/LIF/PI3K/Akt Axis. *Biopreserv Biobank*. 2021;19:376-385.
27. Louzao-Martinez L, Vink A, Harakalova M, Asselbergs FW, Verhaar MC and Cheng C. Characteristic adaptations of the extracellular matrix in dilated cardiomyopathy. *Int J Cardiol*. 2016;220:634-46.
28. Cai ZJ, Lee YK, Lau YM, Ho JC, Lai WH, Wong NL, Huang D, Hai JJ, Ng KM, Tse HF and Siu CW. Expression of Lmna-R225X nonsense mutation results in dilated cardiomyopathy and conduction disorders (DCM-CD) in mice: Impact of exercise training. *Int J Cardiol*. 2020;298:85-92.
29. Chang L, Huang R, Chen J, Li G, Shi G, Xu B and Wang L. An alpha-helix variant p.Arg156Pro in LMNA as a cause of hereditary dilated cardiomyopathy: genetics and bioinfomatics exploration. *BMC medical genomics*. 2023;16:229.
30. Lyon RC, Zanella F, Omens JH and Sheikh F. Mechanotransduction in cardiac hypertrophy and failure. *Circ Res*. 2015;116:1462-1476.
31. Tsuru H, Yoshihara C, Suginobe H, Matsumoto M, Ishii Y, Narita J, Ishii R, Wang R, Ueyama A, Ueda K, Hirose M, Hashimoto K, Nagano H, Tanaka R, Okajima T, Ozono K and Ishida H. Pathogenic Roles of Cardiac Fibroblasts in Pediatric Dilated Cardiomyopathy. *Journal of the American Heart Association*. 2023;12:e029676.
32. Arbustini E, Pilotto A, Repetto A, Grasso M, Negri A, Diegoli M, Campana C, Scelsi L, Baldini E, Gavazzi A and Tavazzi L. Autosomal dominant dilated cardiomyopathy with atrioventricular block: a lamin A/C defect-related disease. *J Am Coll Cardiol*. 2002;39:981-90.
33. Brodt C, Siegfried JD, Hofmeyer M, Martel J, Rampersaud E, Li D, Morales A and Hershberger RE. Temporal relationship of conduction system disease and ventricular dysfunction in LMNA cardiomyopathy. *Journal of cardiac failure*. 2013;19:233-9.
34. Kleber AG and Rudy Y. Basic mechanisms of cardiac impulse propagation and associated arrhythmias. *Physiol Rev*. 2004;84:431-88.
35. De Zio R, Pietrafesa G, Milano S, Procino G, Bramerio M, Pepe M, Forleo C, Favale S, Svelto M, Gerbino A and Carmosino M. Role of Nuclear Lamin A/C in the Regulation of Nav1.5 Channel and

Microtubules: Lesson From the Pathogenic Lamin A/C Variant Q517X. *Front Cell Dev Biol*. 2022;10:918760.

36. Liu Z, Shan H, Huang J, Li N, Hou C and Pu J. A novel lamin A/C gene missense mutation (445 V > E) in immunoglobulin-like fold associated with left ventricular non-compaction. *Europace*. 2016;18:617-22.
37. Markandeya YS, Tsubouchi T, Hacker TA, Wolff MR, Belardinelli L and Balijepalli RC. Inhibition of late sodium current attenuates ionic arrhythmia mechanism in ventricular myocytes expressing LaminA-N195K mutation. *Heart Rhythm*. 2016;13:2228-2236.
38. Olaopa MA, Ai T, Chao B, Xiao X, Vatta M and Habecker BA. Phosphorylation of Lamin A/C at serine 22 modulates Na(v) 1.5 function. *Physiological reports*. 2021;9:e15121.
39. Salvarani N, Crasto S, Miragoli M, Bertero A, Paulis M, Kunderfranco P, Serio S, Forni A, Lucarelli C, Dal Ferro M, Larcher V, Sinagra G, Vezzoni P, Murry CE, Faggian G, Condorelli G and Di Pasquale E. The K219T-Lamin mutation induces conduction defects through epigenetic inhibition of SCN5A in human cardiac laminopathy. *Nature communications*. 2019;10:2267.
40. Macquart C, Juttner R, Morales Rodriguez B, Le Dour C, Lefebvre F, Chatzifrangkeskou M, Schmitt A, Gotthardt M, Bonne G and Muchir A. Microtubule cytoskeleton regulates Connexin 43 localization and cardiac conduction in cardiomyopathy caused by mutation in A-type lamins gene. *Hum Mol Genet*. 2019;28:4043-4052.
41. Bengel FM, Permanetter B, Ungerer M, Nekolla SG and Schwaiger M. Relationship between altered sympathetic innervation, oxidative metabolism and contractile function in the cardiomyopathic human heart; a non-invasive study using positron emission tomography. *Eur Heart J*. 2001;22:1594-600.
42. Zhang X, Shao X, Zhang R, Zhu R and Feng R. Integrated analysis reveals the alterations that LMNA interacts with euchromatin in LMNA mutation-associated dilated cardiomyopathy. *Clin Epigenetics*. 2021;13:3.
43. Austin KM, Trembley MA, Chandler SF, Sanders SP, Saffitz JE, Abrams DJ and Pu WT. Molecular mechanisms of arrhythmogenic cardiomyopathy. *Nat Rev Cardiol*. 2019;16:519-537.
44. Kato K, Takahashi N, Fujii Y, Umehara A, Nishiuchi S, Makiyama T, Ohno S and Horie M. LMNA cardiomyopathy detected in Japanese arrhythmogenic right ventricular cardiomyopathy cohort. *Journal of cardiology*. 2016;68:346-51.
45. Patel V, Asatryan B, Siripanthong B, Munroe PB, Tiku-Owens A, Lopes LR, Khanji MY, Protonotarios A, Santangeli P, Muser D, Marchlinski FE, Brady PA and Chahal CAA. State of the Art Review on Genetics and Precision Medicine in Arrhythmogenic Cardiomyopathy. *International journal of molecular sciences*. 2020;21.
46. Quarta G, Syrris P, Ashworth M, Jenkins S, Zuborne Alapi K, Morgan J, Muir A, Pantazis A, McKenna WJ and Elliott PM. Mutations in the Lamin A/C gene mimic arrhythmogenic right ventricular cardiomyopathy. *Eur Heart J*. 2012;33:1128-36.
47. Crasto S, My I and Di Pasquale E. The Broad Spectrum of LMNA Cardiac Diseases: From Molecular Mechanisms to Clinical Phenotype. *Front Physiol*. 2020;11:761.

48. Corne TDJ, Sieprath T, Vandenbussche J, Mohammed D, Te Lindert M, Gevaert K, Gabriele S, Wolf K and De Vos WH. Deregulation of focal adhesion formation and cytoskeletal tension due to loss of A-type lamins. *Cell Adh Migr*. 2017;11:447-463.
49. Son NH, Park TS, Yamashita H, Yokoyama M, Huggins LA, Okajima K, Homma S, Szabolcs MJ, Huang LS and Goldberg IJ. Cardiomyocyte expression of PPARgamma leads to cardiac dysfunction in mice. *J Clin Invest*. 2007;117:2791-801.
50. West JA, Beqqali A, Ament Z, Elliott P, Pinto YM, Arbustini E and Griffin JL. A targeted metabolomics assay for cardiac metabolism and demonstration using a mouse model of dilated cardiomyopathy. *Metabolomics*. 2016;12:59.
51. Tariq Z, Zhang H, Chia-Liu A, Shen Y, Gete Y, Xiong ZM, Tocheny C, Campanello L, Wu D, Losert W and Cao K. Lamin A and microtubules collaborate to maintain nuclear morphology. *Nucleus*. 2017;8:433-446.
52. Borin D, Pena B, Chen SN, Long CS, Taylor MRG, Mestroni L and Sbaizero O. Altered microtubule structure, hemichannel localization and beating activity in cardiomyocytes expressing pathologic nuclear lamin A/C. *Heliyon*. 2020;6:e03175.
53. Li M, Chen X, Chen L, Chen K, Zhou J and Song J. MiR-1-3p that correlates with left ventricular function of HCM can serve as a potential target and differentiate HCM from DCM. *J Transl Med*. 2018;16:161.
54. Ai Z, Fischer A, Spray DC, Brown AM and Fishman GI. Wnt-1 regulation of connexin43 in cardiac myocytes. *J Clin Invest*. 2000;105:161-71.
55. Le Dour C, Macquart C, Sera F, Homma S, Bonne G, Morrow JP, Worman HJ and Muchir A. Decreased WNT/beta-catenin signalling contributes to the pathogenesis of dilated cardiomyopathy caused by mutations in the lamin a/C gene. *Hum Mol Genet*. 2017;26:333-343.
56. Olson DJ, Christian JL and Moon RT. Effect of wnt-1 and related proteins on gap junctional communication in Xenopus embryos. *Science*. 1991;252:1173-6.
57. Perovanovic J, Dell'Orso S, Gnochi VF, Jaiswal JK, Sartorelli V, Vigouroux C, Mamchaoui K, Mouly V, Bonne G and Hoffman EP. Laminopathies disrupt epigenomic developmental programs and cell fate. *Science translational medicine*. 2016;8:335ra58.
58. Singh KK, Yanagawa B, Quan A, Wang R, Garg A, Khan R, Pan Y, Wheatcroft MD, Lovren F, Teoh H and Verma S. Autophagy gene fingerprint in human ischemia and reperfusion. *The Journal of thoracic and cardiovascular surgery*. 2014;147:1065-1072 e1.
59. Zhou X, Zhang S, Zhao Y, Wang W and Zhang H. A multi-omics approach to identify molecular alterations in a mouse model of heart failure. *Theranostics*. 2022;12:1607-1620.
60. Steinberg C, Roston TM, van der Werf C, Sanatani S, Chen SRW, Wilde AAM and Krahn AD. RYR2-ryanodinopathies: from calcium overload to calcium deficiency. *Europace*. 2023;25.
61. Dridi H, Wu W, Reiken SR, Ofer RM, Liu Y, Yuan Q, Sittenfeld L, Kushner J, Muchir A, Worman HJ and Marks AR. Ryanodine receptor remodeling in cardiomyopathy and muscular dystrophy caused by lamin A/C gene mutation. *Hum Mol Genet*. 2021;29:3919-3934.
62. Zhao B, Ye X, Yu J, Li L, Li W, Li S, Yu J, Lin JD, Wang CY, Chinnaiyan AM, Lai ZC and Guan KL. TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev*. 2008;22:1962-71.

63. Chen Z, Friedrich GA and Soriano P. Transcriptional enhancer factor 1 disruption by a retroviral gene trap leads to heart defects and embryonic lethality in mice. *Genes Dev*. 1994;8:2293-301.
64. Liu R, Lee J, Kim BS, Wang Q, Buxton SK, Balasubramanyam N, Kim JJ, Dong J, Zhang A, Li S, Gupte AA, Hamilton DJ, Martin JF, Rodney GG, Coarfa C, Wehrens XH, Yechoor VK and Moulik M. Tead1 is required for maintaining adult cardiomyocyte function, and its loss results in lethal dilated cardiomyopathy. *JCI insight*. 2017;2.
65. Wu Y, Qian R, Yang Y, Sheng Y, Li W and Wang W. Activation Pathways and Free Energy Landscapes of the SARS-CoV-2 Spike Protein. *ACS Omega*. 2021;6:23432-23441.
66. Yamada S, Ko T, Ito M, Sassa T, Nomura S, Okuma H, Sato M, Imasaki T, Kikkawa S, Zhang B, Yamada T, Seki Y, Fujita K, Katoh M, Kubota M, Hatsuse S, Katagiri M, Hayashi H, Hamano M, Takeda N, Morita H, Takada S, Toyoda M, Uchiyama M, Ikeuchi M, Toyooka K, Umezawa A, Yamanishi Y, Nitta R, Aburatani H and Komuro I. TEAD1 trapping by the Q353R-Lamin A/C causes dilated cardiomyopathy. *Sci Adv*. 2023;9:eade7047.
67. Tsutsui H, Kinugawa S and Matsushima S. Oxidative stress and heart failure. *Am J Physiol Heart Circ Physiol*. 2011;301:H2181-90.
68. Koyama H, Nojiri H, Kawakami S, Sunagawa T, Shirasawa T and Shimizu T. Antioxidants improve the phenotypes of dilated cardiomyopathy and muscle fatigue in mitochondrial superoxide dismutase-deficient mice. *Molecules*. 2013;18:1383-93.
69. Almomani R, Herkert JC, Posafalvi A, Post JG, Boven LG, van der Zwaag PA, Willems P, van Veen-Hof IH, Verhagen JMA, Wessels MW, Nikkels PGJ, Wintjes LT, van den Berg MP, Sinke RJ, Rodenburg RJ, Niezen-Koning KE, van Tintelen JP and Jongbloed JDH. Homozygous damaging SOD2 variant causes lethal neonatal dilated cardiomyopathy. *J Med Genet*. 2020;57:23-30.
70. Sharma S, Bhattarai S, Ara H, Sun G, St Clair DK, Bhuiyan MS, Kevil C, Watts MN, Dominic P, Shimizu T, McCarthy KJ, Sun H, Panchatcharam M and Miriyala S. SOD2 deficiency in cardiomyocytes defines defective mitochondrial bioenergetics as a cause of lethal dilated cardiomyopathy. *Redox Biol*. 2020;37:101740.
71. Friedman RC, Farh KK, Burge CB and Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*. 2009;19:92-105.
72. Pons-Espinal M, de Luca E, Marzi MJ, Beckervordersandforth R, Armirotti A, Nicassio F, Fabel K, Kempermann G and De Pietri Tonelli D. Synergic Functions of miRNAs Determine Neuronal Fate of Adult Neural Stem Cells. *Stem cell reports*. 2017;8:1046-1061.
73. Sahu M and Mallick B. Deciphering synergistic regulatory networks of microRNAs in hESCs and fibroblasts. *Int J Biol Macromol*. 2018;113:1279-1286.
74. Wu S, Huang S, Ding J, Zhao Y, Liang L, Liu T, Zhan R and He X. Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. *Oncogene*. 2010;29:2302-8.
75. Porrello ER, Mahmoud AI, Simpson E, Johnson BA, Grinsfelder D, Canseco D, Mammen PP, Rothermel BA, Olson EN and Sadek HA. Regulation of neonatal and adult mammalian heart regeneration by the miR-15 family. *Proc Natl Acad Sci U S A*. 2013;110:187-92.

76. Tijsen AJ, van der Made I, van den Hoogenhof MM, Wijnen WJ, van Deel ED, de Groot NE, Alekseev S, Fluiter K, Schroen B, Goumans MJ, van der Velden J, Duncker DJ, Pinto YM and Creemers EE. The microRNA-15 family inhibits the TGFbeta-pathway in the heart. *Cardiovasc Res*. 2014;104:61-71.
77. Jin B, Jin D, Zhuo Z, Zhang B and Chen K. MiR-1224-5p Activates Autophagy, Cell Invasion and Inhibits Epithelial-to-Mesenchymal Transition in Osteosarcoma Cells by Directly Targeting PLK1 Through PI3K/AKT/mTOR Signaling Pathway. *Onco Targets Ther*. 2020;13:11807-11818.
78. Yao X, Cui X, Wu X, Xu P, Zhu W, Chen X and Zhao T. Tumor suppressive role of miR-1224-5p in keloid proliferation, apoptosis and invasion via the TGF-beta1/Smad3 signaling pathway. *Biochem Biophys Res Commun*. 2018;495:713-720.
79. Wang Y, Li M, Xu L, Liu J, Wang D, Li Q, Wang L, Li P, Chen S and Liu T. Expression of Bcl-2 and microRNAs in cardiac tissues of patients with dilated cardiomyopathy. *Molecular medicine reports*. 2017;15:359-365.
80. Xiao Y, Zhao J, Tuazon JP, Borlongan CV and Yu G. MicroRNA-133a and Myocardial Infarction. *Cell Transplant*. 2019;28:831-838.
81. Poon IK, Lucas CD, Rossi AG and Ravichandran KS. Apoptotic cell clearance: basic biology and therapeutic potential. *Nat Rev Immunol*. 2014;14:166-80.
82. Saddic LA, Chang TW, Sigurdsson MI, Heydarpour M, Raby BA, Shernan SK, Aranki SF, Body SC and Muehlschlegel JD. Integrated microRNA and mRNA responses to acute human left ventricular ischemia. *Physiol Genomics*. 2015;47:455-62.
83. Ming S, Shui-Yun W, Wei Q, Jian-Hui L, Ru-Tai H, Lei S, Mei J, Hui W and Ji-Zheng W. miR-139-5p inhibits isoproterenol-induced cardiac hypertrophy by targetting c-Jun. *Biosci Rep*. 2018;38.
84. Penton AL, Leonard LD and Spinner NB. Notch signaling in human development and disease. *Semin Cell Dev Biol*. 2012;23:450-7.
85. Rojanasopondist P, Nesheiwat L, Piombo S, Porter GA, Jr., Ren M and Phoon CKL. Genetic Basis of Left Ventricular Noncompaction. *Circulation Genomic and precision medicine*. 2022;15:e003517.
86. Langa P, Shafaattalab S, Goldspink PH, Wolska BM, Fernandes AA, Tibbits GF and Solaro RJ. A perspective on Notch signalling in progression and arrhythmogenesis in familial hypertrophic and dilated cardiomyopathies. *Philos Trans R Soc Lond B Biol Sci*. 2023;378:20220176.
87. Besler C, Urban D, Watzka S, Lang D, Rommel KP, Kandolf R, Klingel K, Thiele H, Linke A, Schuler G, Adams V and Lurz P. Endomyocardial miR-133a levels correlate with myocardial inflammation, improved left ventricular function, and clinical outcome in patients with inflammatory cardiomyopathy. *European journal of heart failure*. 2016;18:1442-1451.
88. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, Fang L, Li Z, Lin L, Liu R, Zhang Y, Xu H, Li S, Zhou Y, Davies RW, Liu Q, Walters RG, Lin K, Ju J, Korneliussen T, Yang MA, Fu Q, Wang J, Zhou L, Krogh A, Zhang H, Wang W, Chen Z, Cai Z, Yin Y, Yang H, Mao M, Shendure J, Wang J, Albrechtsen A, Jin X, Nielsen R and Xu X. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell*. 2018;175:347-359 e14.

89. Lewandowski P, Golawski M, Baron M, Reichman-Warmusz E and Wojnicz R. A Systematic Review of miRNA and cfDNA as Potential Biomarkers for Liquid Biopsy in Myocarditis and Inflammatory Dilated Cardiomyopathy. *Biomolecules*. 2022;12.
90. Gerbino A, Forleo C, Milano S, Piccapane F, Procino G, Pepe M, Piccolo M, Guida P, Resta N, Favale S, Svelto M and Carmosino M. Pro-inflammatory cytokines as emerging molecular determinants in cardiolaminopathies. *Journal of cellular and molecular medicine*. 2021;25:10902-10915.

## Table legend

**Table 1. miRNA and summary of their targets.** This table summarizes the coRmiT output showing the DEMs with correlated targets, their log2FC. The table also shows how many correlated targets of each miRNA have been previously validated (according to databases) and the RNA-seq co-expression modules that includes those targets.

| miRNA | miRNA log2FC | Validated targets | RNA-seq modules |
|---|---|---|---|
| mmu-miR-690 | 1,66 | 46 | 1,10 |
| mmu-miR-183-5p | 1,33 | 9 | 1,10 |
| mmu-miR-324-5p | 1,30 | 9 | 1,10 |
| mmu-miR-3473a | 1,12 | 21 | 1,10 |
| mmu-miR-182-5p | 2,01 | 0 | 5 |
| mmu-miR-758-5p | 1,44 | 0 | 1,10 |
| mmu-miR-375-5p | 1,40 | 0 | 1 |
| mmu-miR-5619-5p | 1,10 | 0 | 1,10 |
| mmu-miR-139-5p | -1,08 | 22 | 2,3,39 |
| mmu-miR-196b-5p | -1,10 | 199 | 2,3,39 |
| mmu-miR-3473b | -1,14 | 6 | 2,3,39 |
| mmu-miR-155-5p | -1,26 | 1718 | 2,3,39 |
| mmu-miR-133a-5p | -1,27 | 109 | 2,3,39 |
| mmu-miR-1224-5p | -1,32 | 2 | 2,3,39 |
| mmu-miR-3095-5p | -1,38 | 2 | 2,3,39 |
| mmu-miR-149-5p | -1,53 | 61 | 2,3,39 |
| mmu-miR-1249-5p | -1,16 | 0 | 2,3,39 |
| mmu-miR-133b-5p | -1,29 | 0 | 2,3,39 |
| mmu-miR-5132-5p | -1,36 | 0 | 2,3 |

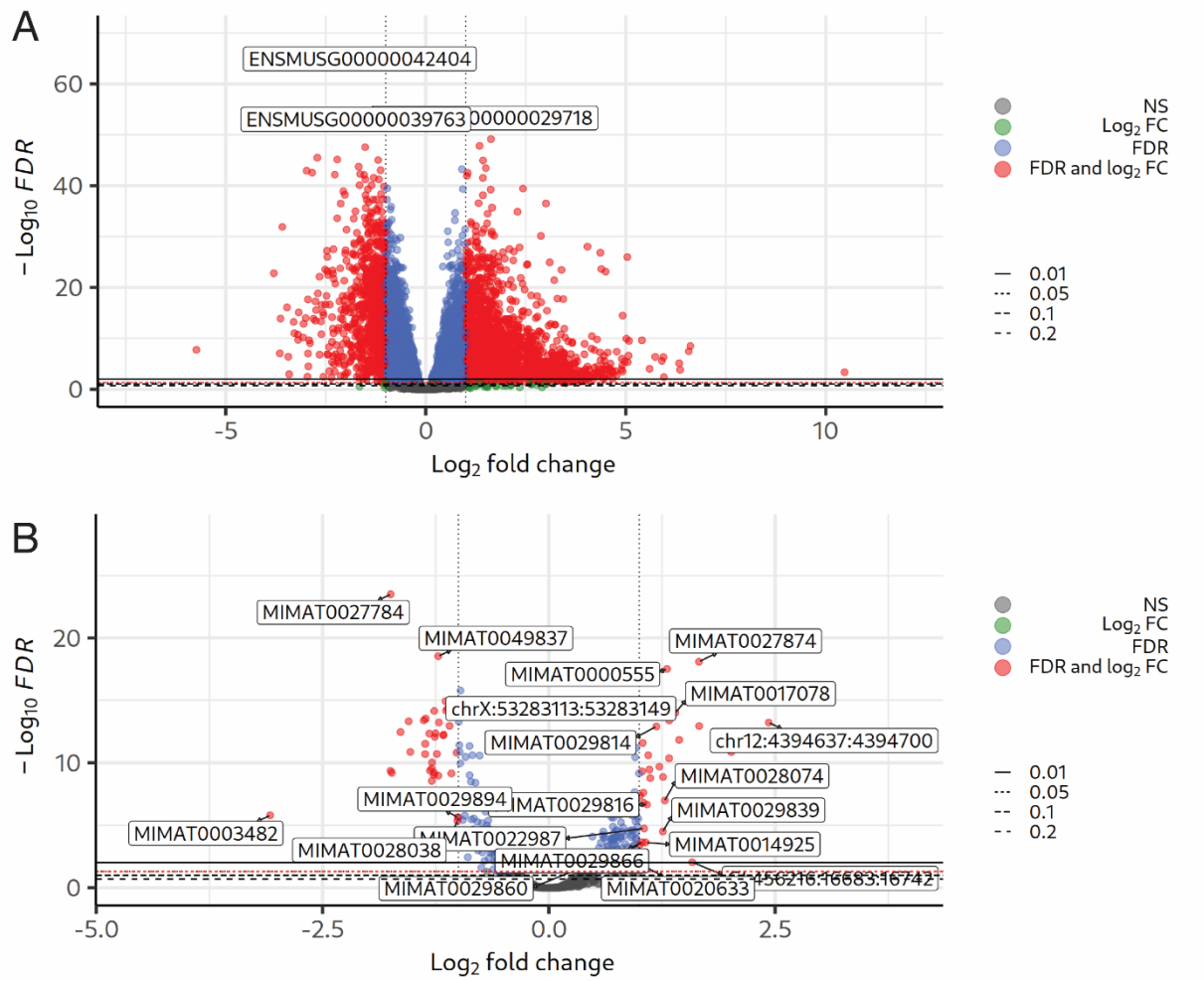


**Figure 1. DEGs and DEMs in *LMNA*$^{R249W}$ tissue samples. (A)** The volcano plots showing DEGs in *LMNA*$^{R249W}$ samples compared to wild-type. **(B)** The volcano plots showing DEMs in *LMNA*$^{R249W}$ samples compared to wild-type. The X axis represent the log2 fold change and the Y axis correspond to -logFDR. The dots are genes/miRNA which have been classified in colors: red dots represent transcripts with |log2FC| > 1 and FDR <= 0.05, blue dots are those with |log2FC| <= 1 and FDR <= 0.05, green are those with |log2FC| > 1 and FDR > 0.05 and grey the remaining. Dashed lines show the different FDR thresholds. RNA was obtained from myocardial tissue of 50-week-old *Lmna*$^{+/R249W}$ (n = 6) and age-matched wild-typed type mice (n = 6).

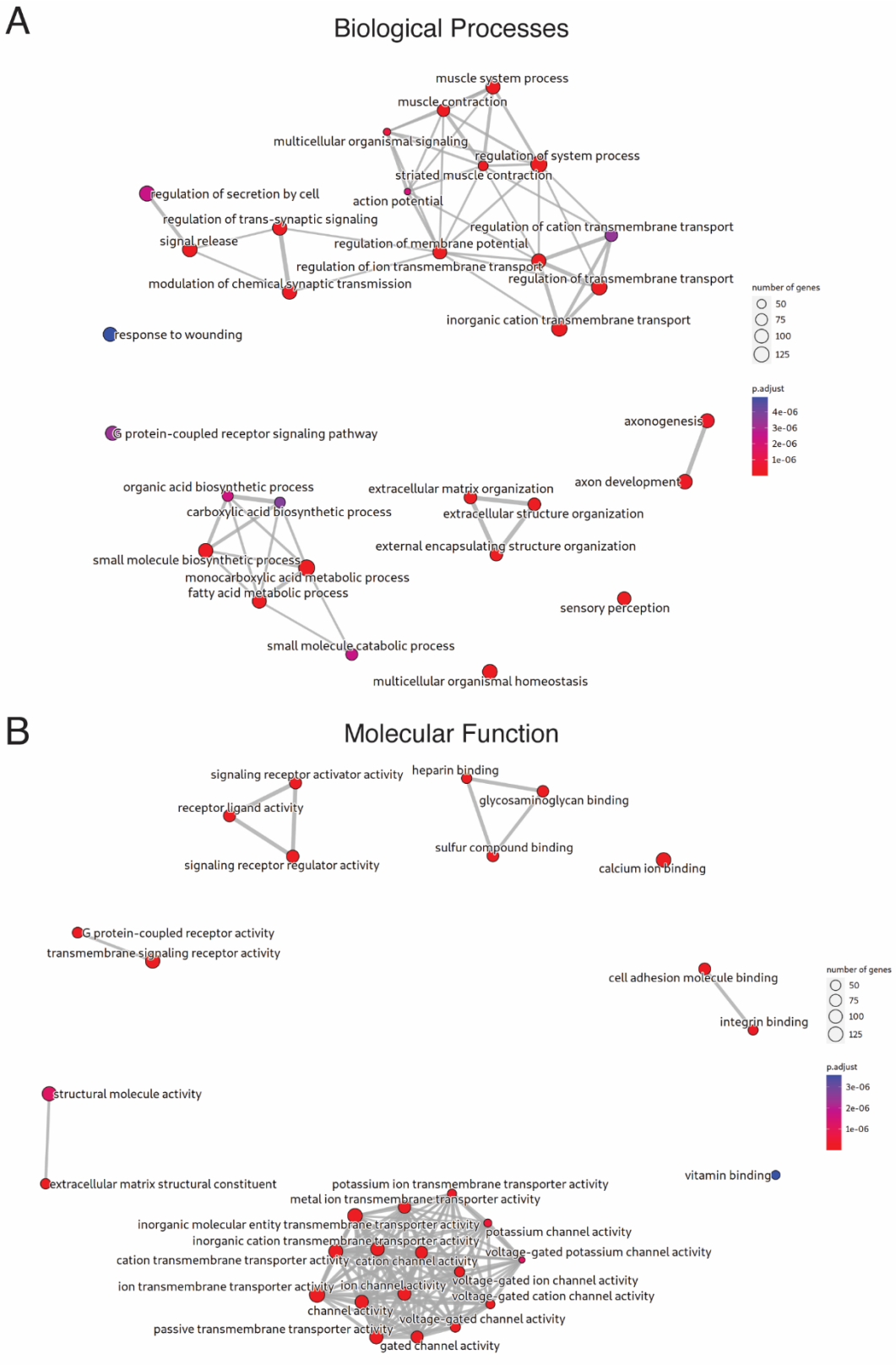


**Figure 2. Enriched GO Biological Process terms in DEGs**. **(A)** Top 30 enriched GO BP terms in DEGs. The color of the dots represents the FDR and their size represent the number of DEGs annotated for each term. Two terms are linked when they shared annotated DEGs. The wider and shorter links correspond to more DEGs shared between terms. **(B)** Enriched GO Molecular Function terms in DEGs.

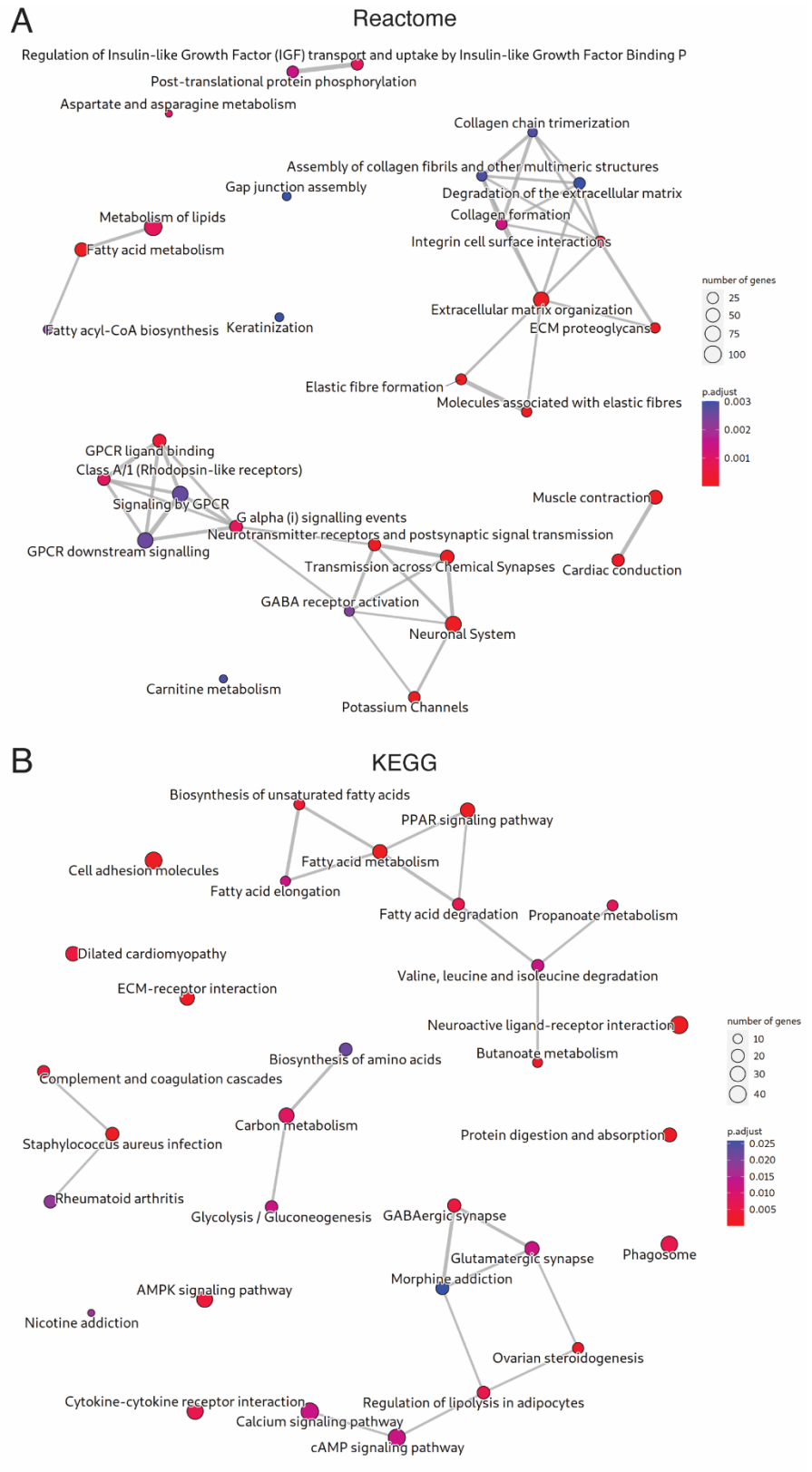


**Figure 3. Enriched pathways in differentially expressed genes**. **(A)** Reactome pathways. **(B)** KEGG pathways. This plot shows as dots the top 30 enriched pathways in DEGs. The color of the dots represents the FDR and their size represent the number of DEGs annotated for each pathway. Two pathways are linked when they shared annotated DEGs. The wider and shorter links correspond to more DEGs shared between pathways.

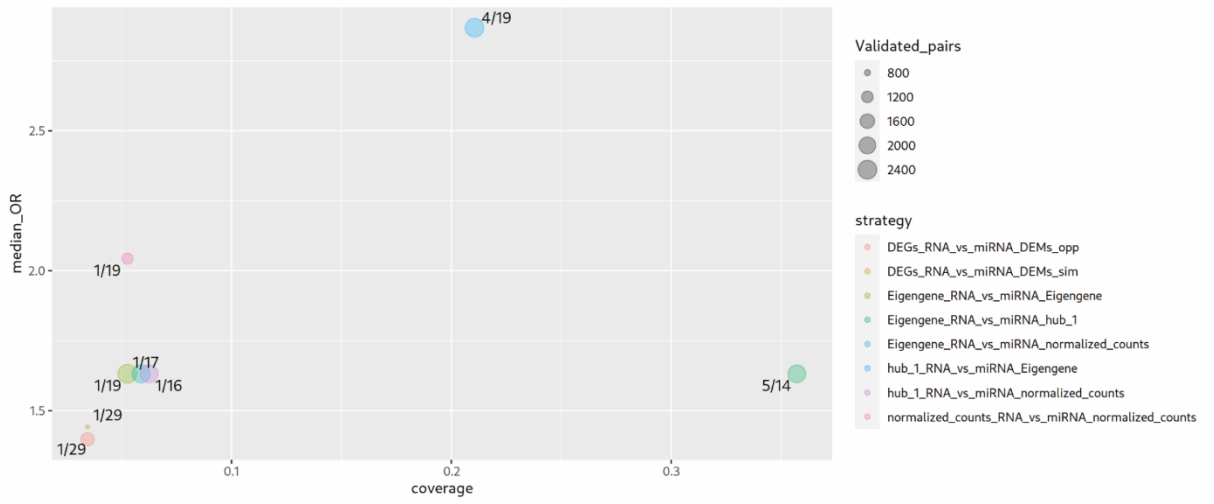


**Figure 4. miRNA-target correlation strategies**. This plot shows the ranking of the different correlation strategies by their odds ratio (Y axis) and their coverage (X axis). The higher odds ratio represents more representation of the correlated pairs in database. The coverage is the proportion of miRNA with correlated targets within strategy that showed significant overlap with databases according to Fisher test FDR < 0.05. The dots represent different correlation strategies where the color indicates the name of the strategy and the size correspond to the number of miRNA-target pairs found in databases.

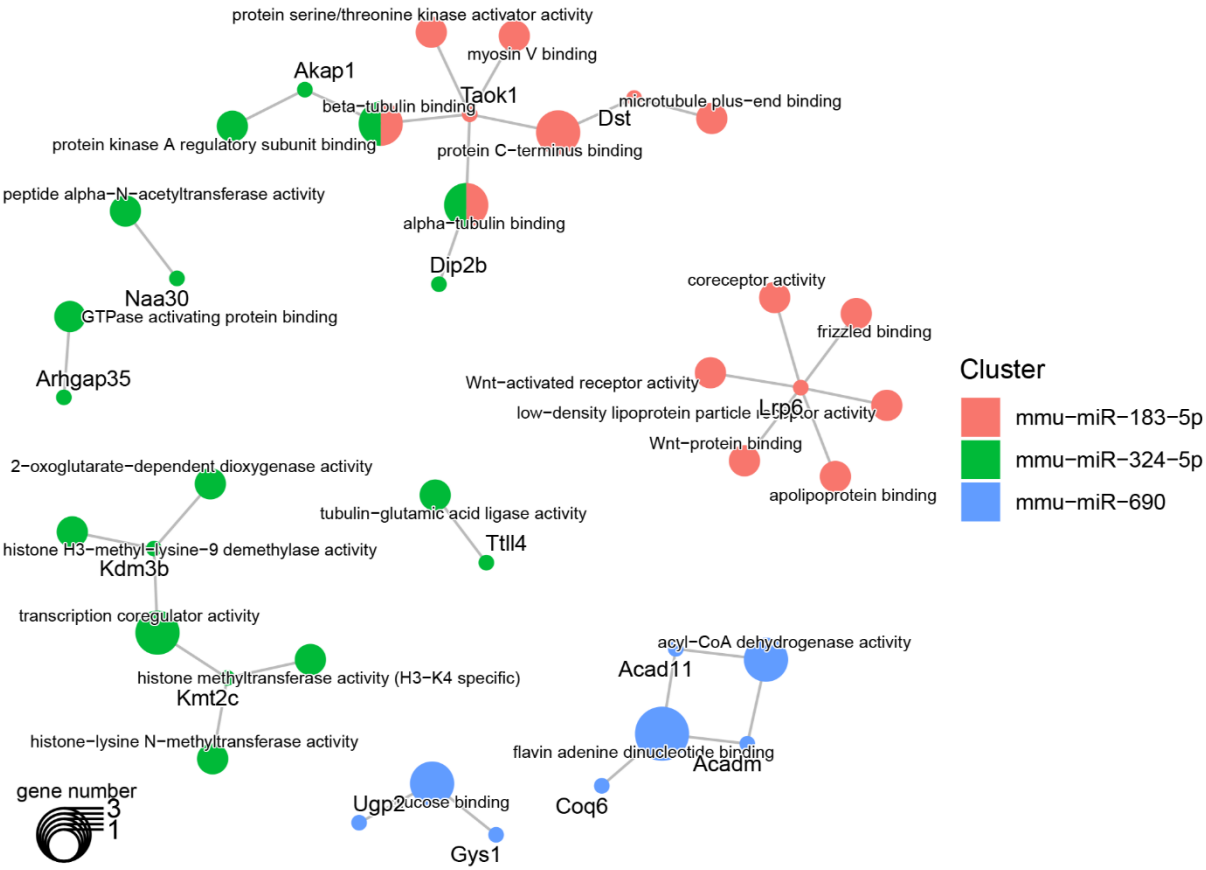

**Figure 5**. **Enriched GO Molecular Function terms in targets of the upregulated miRNA.** This plot shows the significant enriched GO MF terms (larger dots) in target genes (smaller dots). The color of the dots represents the miRNA that targets the gene and hence, the enriched term. The dots with two colors are targeted by two different miRNAs. The size of the term dot represents how many targets are annotated with it.

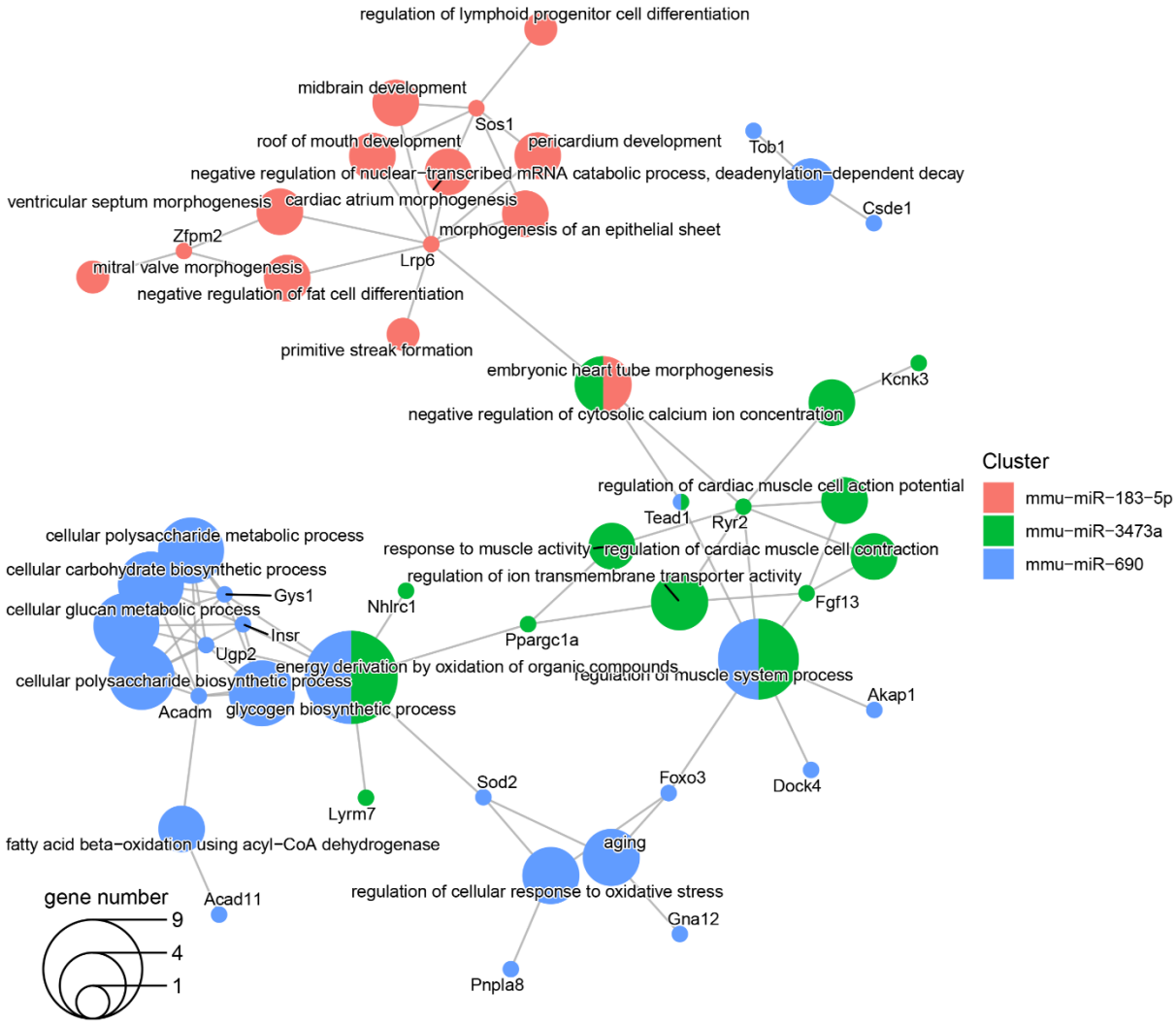


**Figure 6. Enriched GO Biological Process terms in targets of the upregulated miRNA**. This plot shows the significant enriched GO BP terms (larger dots) in target genes (smaller dots). The color of the dots represents the miRNA that targets the gene and hence, the enriched term. The dots with two colors are targeted by two different miRNAs. The size of the term dot represents how many targets are annotated with it.

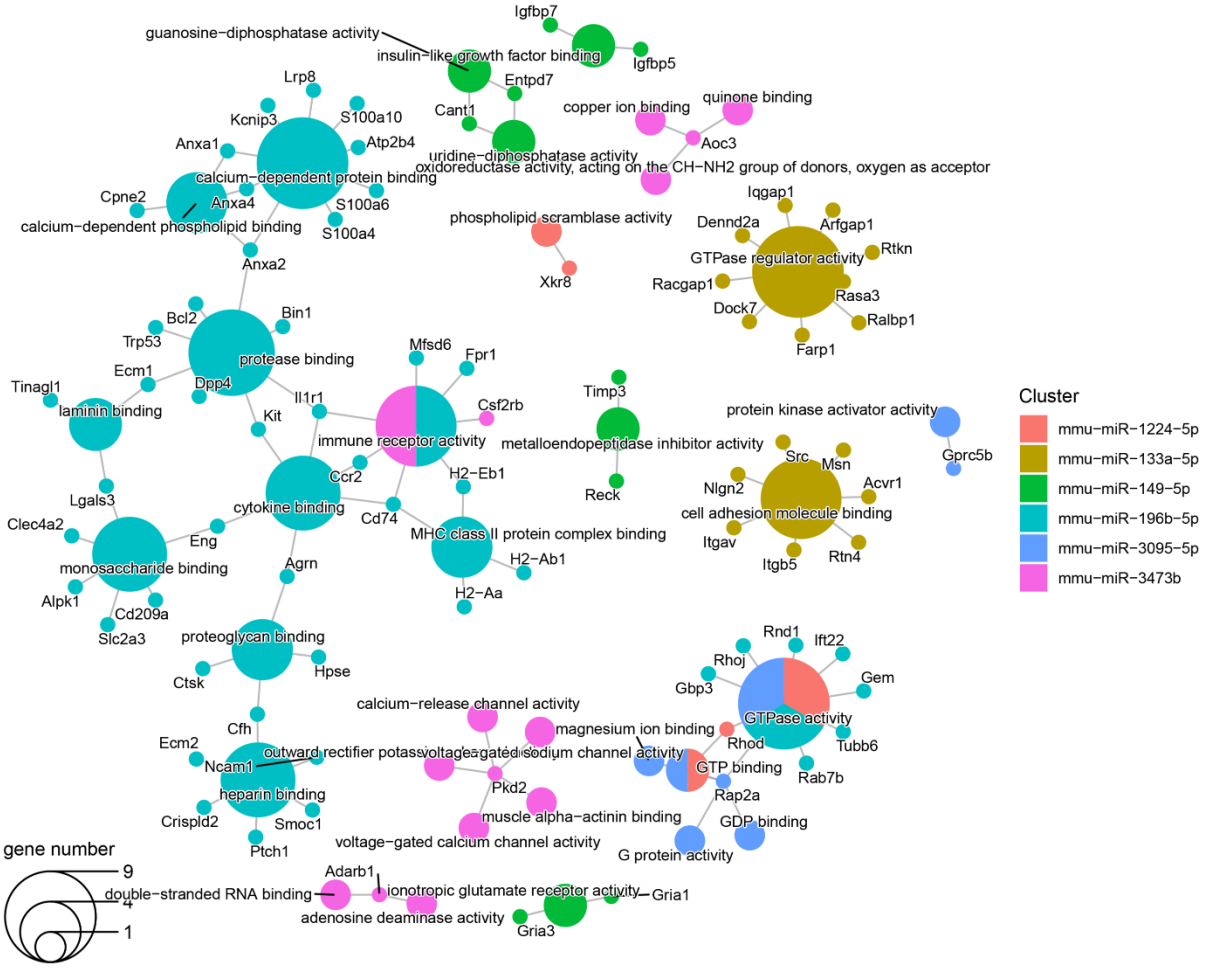


**Figure 7. Enriched Gene Ontology Molecular Function terms in targets of the downregulated miRNAs**. This plot shows the significant enriched GO MF terms (larger dots) in target genes of each miRNA (smaller dots). The color of the dots represents the miRNA that targets the gene and hence, the enriched term. The dots with two colors are targeted by two different miRNAs. The size of the term dot represents how many miRNAs targeted genes annotated with each term.

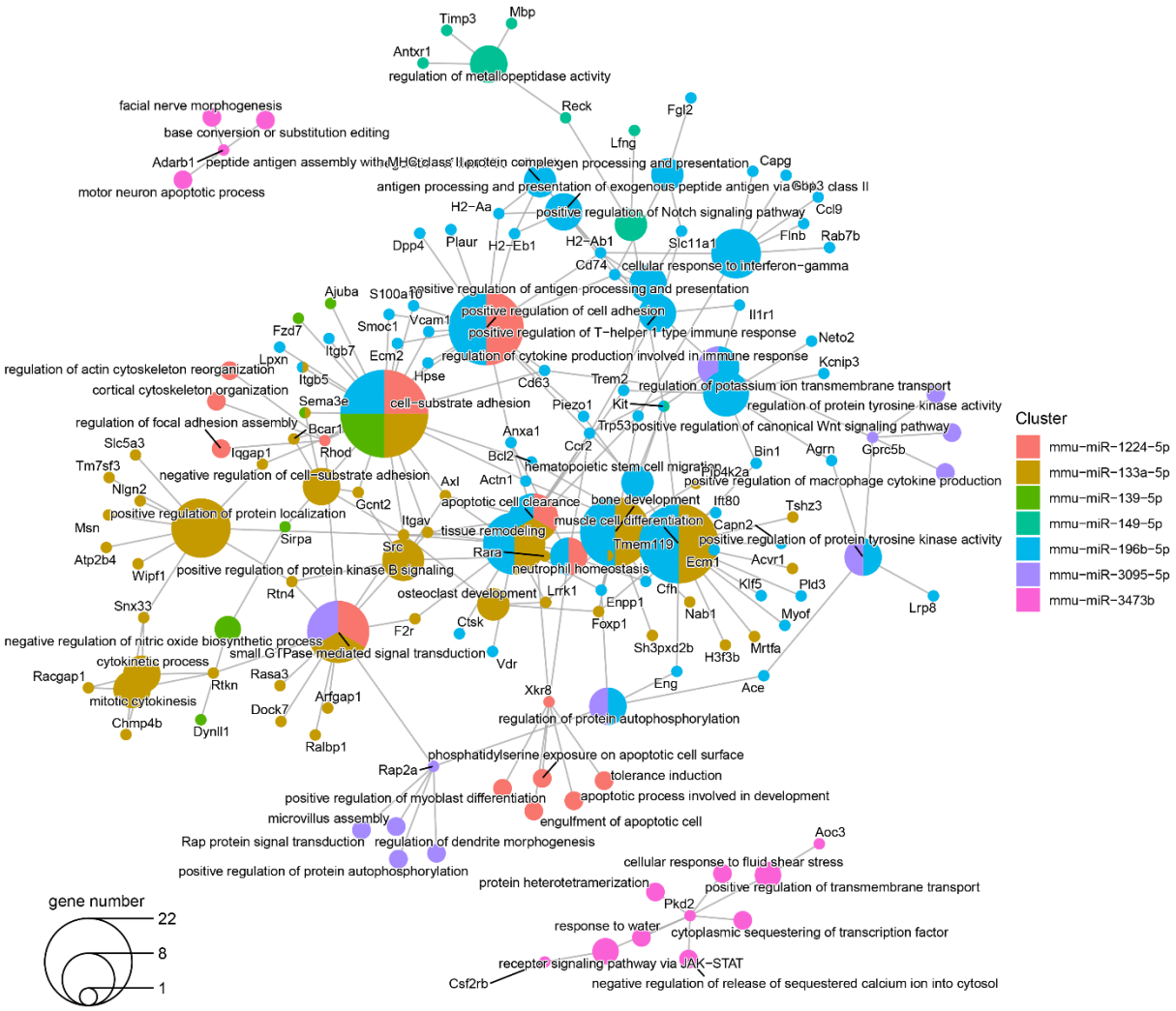


**Figure 8. Enriched GO Biological Process terms in targets of the downregulated miRNA**. This plot shows the significant enriched GO BP terms (larger dots) in target genes of each miRNA (smaller dots). The color of the dots represents the miRNA that targets the gene and hence, the enriched term. The dots with two colors are targeted by two different miRNAs. The size of the term dot represents how many miRNAs targeted genes annotated with each term.

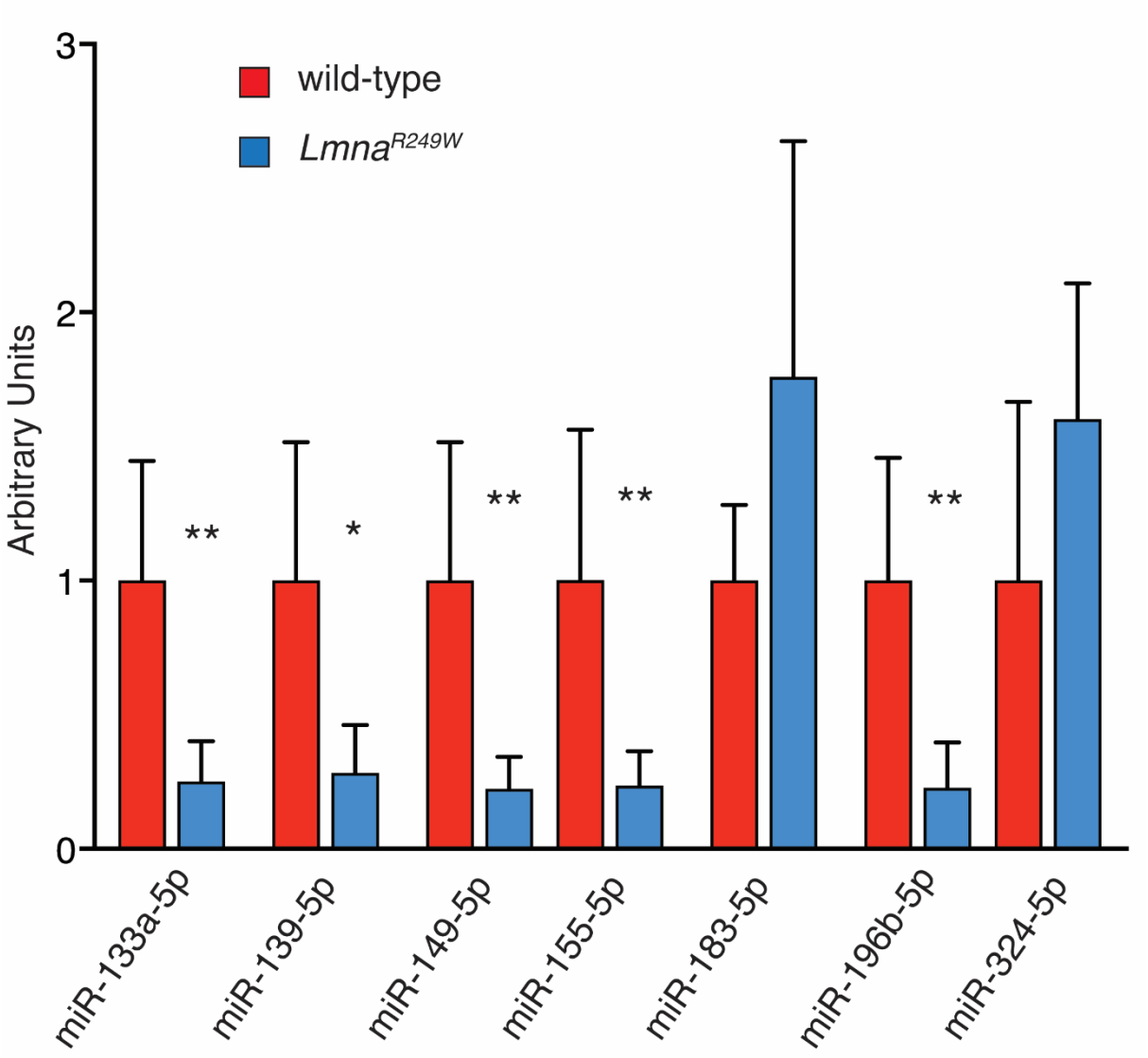


**Figure 9. qRT-PCR validation**. Expression levels of miRNAs in *LMNA*$^{R249W}$ (n = 6) vs wild-type (n = 6) analyzed by qRT-PCR. Statistical significance was determined using a two-tailed t test. * $p < 0.05$; ** $p < 0.01$.

# Capítulo 9

# Integración de transcriptómica y otros factores experimentales basado en el Análisis de Componentes Principales

## 9.1. Introducción

Los datos de expresión génica obtenidos por RNA-seq se incluyen en una tabla de conteos como se detalló en la Sección 2.5. Esta tabla contiene un valor de expresión por cada gen, siendo un claro ejemplo de elevada dimensionalidad de los datos. Por lo tanto, la aplicación de técnicas de reducción de la dimensionalidad, como el Análisis de Componentes Principales (ACP), permite resumir la complejidad de los datos sin perder información relevante [91, 92, 93].

El ACP es comúnmente utilizado para hacer el control de calidad de las muestras, ya que permite revisar visualmente si se agrupan de acuerdo con el diseño experimental y detectar posibles sesgos en los protocolos utilizados [94]. Estos sesgos pueden ser efectos de lote, diferencias a la hora de preparar la librería de secuenciación o errores en el proceso de secuenciación [95]. Además, pueden existir factores técnicos y biológicos que alteren de manera inesperada la expresión génica, los cuales deben ser considerados al interpretar los resultados [96, 94]. En consecuencia, el ACP posee un gran potencial para facilitar la interpretación de los experimentos y las causas subyacentes de los resultados observados, especialmente cuando se complementa con metodologías adicionales [97].

El ACP permite reducir la dimensionalidad de los datos al transformar un conjunto amplio de variables en un conjunto más reducido de componentes principales (CP), los cuales capturan la mayor parte de la variabilidad de los datos originales

[98], mediante fórmulas de transición [97]. En el contexto de la expresión génica, el ACP agrupa la varianza de la expresión de los genes en las muestras en CP.

Los CP son dimensiones ortogonales entre sí que se ordenan por el porcentaje de la varianza total que resumen, siendo los CP1 y CP2 los que condensan la mayor varianza [98]. La representación de las muestras en los dos primeros CP permite a los investigadores encontrar patrones de expresión desconocidos, detectar el origen de las variaciones e identificar factores de distorsión y valores atípicos [99, 100, 101, 102, 103]. Sin embargo, puede haber más componentes que expliquen la varianza de algunos factores [104]. En consecuencia, se pueden realizar análisis estadísticos posteriores para analizar cuáles de los CP relevantes resumen la variabilidad de las muestras de forma significativa [98] e investigar las relaciones entre estos CP, los genes y los factores experimentales en mayor profundidad [97]. Además, existen paquetes del lenguaje de programación *R* que facilitan el análisis de los resultados del ACP, como *pcaExplorer* [105] y *PCAtools* [106].

Debido a la naturaleza cuantitativa de la expresión de los genes, al aplicar el ACP se puede calcular la correlación entre los genes y cada CP, así como realizar las pruebas estadísticas correspondientes [106, 107]. Asimismo, las formulas de transición se pueden usar para proyectar factores experimentales y factores biológicos de las muestras (tratados como variables suplementarias) en el espacio de CP sin alterarlo, asistiendo a la interpretación de dichos CP [97]. Tras la proyección de estos factores, se puede calcular la asociación entre éstos y los CP. En el caso de los factores cuantitativos se puede calcular la correlación con los CP y se puede medir su significancia [97]. En el caso de los factores cualitativos, en los que cada individuo está etiquetado con una categoría, se puede calcular las coordenadas promedio de los individuos para asignarle un valor a cada categoría. Este promedio se puede comparar con las coordenadas medias de todos los individuos usando una prueba $t$ de Student [97]. De forma paralela, se puede aplicar un Agrupamiento Jerárquico basado en los Componentes Principales (AJCP) usando sólo estos CP relevantes para determinar grupos de individuos o variables [97]. Además, se puede comparar dichos grupos con el diseño experimental para esbozar conclusiones biológicas.

Por todas estas razones, he implementado el ACP en el programa de análisis de expresión *degenes_hunter.R* dentro del paquete *ExpHunterSuite* [81]. Esta integración ha sido acompañada de la implementación de pruebas estadísticas para relacionar los CP y los grupos derivados del AJCP con los factores experimentales. Este análisis lo he aplicado al experimento de cultivos de fibroblastos derivados de pacientes con la enfermedad PMM2-CDG, comparando pacientes enfermos con distintos grados de gravedad de la enfermedad entre sí. Además, lo he aplicado a un experimento de organoides derivados de pacientes con el síndrome de Schaaf-Yang a distintos tiempos de cultivo. Para ambos casos se ha explorado cómo las muestras se agrupan en base a los patrones de expresión de todos los genes y de

sólo los genes con expresión diferencial, y se ha determinado qué factores experimentales tienen un mayor impacto en cada grupo de muestras.

## 9.2. Material y Métodos

### 9.2.1. Conjuntos de datos experimentales

En esta sección se describen las características de los conjuntos de datos analizados. La información técnica detallada al respecto del análisis de expresión diferencial de cada caso se describe en la Sección 9.2.2.

**PMM2-CDG:** La cohorte de PMM2-CDG consta de 27 pacientes fenotipados con las anotaciones de la *HPO* cuya descripción y análisis se desarrolló en el Capítulo 4. A diez de estos pacientes se les realizó una secuenciación de ARN tal como se describe en los Capítulos 7 y 9 y en Gallego y colaboradores [108]. Estos pacientes fueron clasificados en dos grupos "LOW" (baja gravedad, cinco pacientes) y "HIGH" (alta gravedad, cinco pacientes) según la escala internacional cooperativa de la ataxia (ICARS, por sus siglas en inglés).

En el análisis de expresión diferencial se compararon los pacientes clasificados con baja gravedad (grupo de control) contra los de alta gravedad (grupo de tratamiento) de la enfermedad. Esta nomenclatura está incluida en el análisis como el factor cualitativo "*treat*". La escala de gravedad ICARS, la escala de puntuación pediátrica de desordenes congénitos de la glucosilación de Nijmegen (NPCRS por sus siglas en inglés) y el diámetro relativo medio-sagital del vermis (MVRD por sus siglas en inglés) se incluyeron asimismo como factores cuantitativos suplementarios.

**Síndrome de Schaaf-Yang:** El síndrome de Schaaf-Yang (SSY) es un desorden genético causado por mutaciones sin sentido puntuales en el gen *MAGEL2* que está localizado en el cromosoma 15q11-q13 [109]. A nivel clínico, este síndrome presenta un amplio espectro de síntomas incluyendo una limitación intelectual que puede variar desde leve a severa, un desorden del espectro autista, caracterizado por comportamientos repetitivos y dificultades a la hora de interacciones sociales, y problemas motores como hipotonía muscular, contracturas y retraso en el desarrollo [110]. Para investigar las consecuencias moleculares de las mutaciones en el gen *MAGEL2* asociadas a SSY, un grupo colaborador del grupo investigador donde se realizó esta tesis doctoral desarrolló dos tipos de organoides de cerebro: esferoides corticales humanos (hCS, por sus siglas en inglés) que representan el prosencéfalo dorsal, y esferoides del subpalio humanos (hSS, por sus siglas en inglés) que representan el prosencéfalo ventral. Estos organoides derivan de células de un

único paciente con SSY y de un único individuo control. Las muestras se tomaron a tres tiempos distintos; 30, 60 y 90 días. Para aumentar la reproducibilidad, las muestras consistieron en una mezcla de tres a cinco organoides de diferentes placas de cultivo. El ARN de las muestras se secuenció en una plataforma NovaSeq para producir lecturas pareadas de una longitud de 150 pb con un promedio de 100 millones de lecturas por muestra.

En el análisis de expresión diferencial se compararon los organoides derivados del individuo sano (control) y del paciente (tratamiento). Esta nomenclatura se incluye en el análisis como el factor cualitativo "*treat*". Los tipos de organoides (hCS y hSS), el individuo de origen (siendo S_135 el control y S_66 el paciente) y el tiempo de cultivo de los organoides (30d, 60d y 90d) se usaron como factores cualitativos.

### 9.2.2. Configuración del análisis de expresión

La cuantificación de los genes de cada conjunto de datos se realizó mediante la aplicación del flujo de trabajo *DEG_workflow*[1], descrito en el Capítulo 7. La tabla de conteos obtenida se analizó con el programa *degenes_hunter.R* del paquete *ExpHunterSuite* [81] (Capítulo 6) para calcular la expresión diferencial. Los genes expresados diferencialmente (GEDs) fueron aquellos detectados por los paquetes seleccionados: DESeq2 [52] y edgeR [53], con un valor de $|log_2FC| > 1$ y un *false discovery rate* ($FDR$) $< 0{,}05$.

### 9.2.3. Descripción del ACP y de su implementación en *ExpHunterSuite*

Se ha implementado en el paquete *ExpHunterSuite* [81] (Capítulo 6) un análisis de componentes principales (ACP) para aplicarlo a los datos de expresión. Además, se ha complementado con procedimientos para facilitar la interpretación de los resultados, asociándolos con variables experimentales. Estos pasos adicionales son unas pruebas estadísticas sobre la relación entre los componentes principales (CP) y los datos experimentales de las muestras, un Agrupamiento Jerárquico basado en los Componentes Principales (AJCP) usando los CP relevantes y un análisis de sobrerrepresentación de los grupos de muestras obtenidos con AJCP y los grupos experimentales.

El ACP se ha aplicado a los datos de expresión para transformar las variables (en este caso los valores de expresión de los genes) en CP mediante operaciones lineales, llamadas formulas de transición. Los CP tienen asociado un valor propio

[1] https://github.com/seoanezonjic/DEG_workflow

que está relacionado con la proporción de varianza de los datos originales que representan [97]. Los CP se priorizan por su valor propio de mayor a menor, y por tanto por la varianza que representan, lo que ayuda a la reducción de la dimensionalidad [97]. Normalmente, se usan los CP1 y CP2 para representar visualmente las muestras, pero no siempre capturan todos los cambios relevantes en la expresión de genes entre distintas condiciones biológicas [104].

Para seleccionar los CP relevantes se identificaron aquellos que eran significativos en los datos. Esto se hizo usando el cuantil 0,95 de la distribución de los porcentajes de varianza, derivada de simulaciones de distribuciones normales aleatorias con la misma cantidad de datos que la tabla de conteos [111]. Se usaron las funciones del paquete *FactoMineR* [97], como se detalla en su documentación, para obtener los datos relevantes del espacio de CP y de las muestras analizadas en el mismo.

La interpretación de los CP se suele hacer de forma manual al examinar la disposición de las muestras al representar gráficamente los CP [97]. No obstante, se quiso realizar una interpretación objetiva de cada CP relevante para comprender mejor su significado. Por ello, se incorporó al análisis información adicional de las muestras, como por ejemplo características clínicas, variables ambientales o medidas biológicas que pueda contribuir a la separación de las muestras. Aunque esta información (usada como variables suplementarias) no contribuye al cálculo de los CP, sí puede dar pistas sobre los patrones de expresión de las muestras analizadas. Las variables suplementarias se pueden clasificar, según su naturaleza, en cuantitativas o cualitativas/categóricas y se pueden proyectar sobre el espacio de CP mediante la aplicación de funciones del paquete *FactoMineR* [97].

En el caso de las variables suplementarias cuantitativas, se aplican las fórmulas de transición precomputadas para proyectarlas dentro del espacio de CP. Al proyectar una variable cuantitativa sobre el espacio de CP, se obtiene un vector con las coordenadas de dicha variable en los CP. Posteriormente, se calcula el coeficiente de correlación de Pearson entre el vector de coordenadas de cada variable y las coordenadas de los individuos en los CP [97]. Esta correlación es una medida de asociación entre las variables y los CP. De estas correlaciones se obtiene una significancia estadística asociada y se ordenan de mayor a menor según su $|R|$ para priorizar las variables suplementarias con mayor asociación con los patrones de expresión que engloban los CP [97].

En el caso de las variables suplementarias cualitativas, se realiza un análisis de varianza de una vía (ANOVA) en las coordenadas de los individuos etiquetados con cada variable. Además se calcula el valor de ajuste $R^2$ de la regresión lineal de las coordenadas de los individuos en los CP como valor de asociación entre cada variable cualitativa y los CP. Posteriormente, se realizan pruebas $t$ de Student considerando si las coordenadas promedio de los individuos de cada categoría están

por encima o por debajo de la media global [97].

En relación al AJCP, se consideran las coordenadas de los individuos en todos los CP relevantes para determinar grupos [97]. El árbol jerárquico se calcula con el método de Ward [112] y el número de grupos se determina de forma automática, basado en el incremento de la inercia. Dado un número de grupos $Q$, las reglas de selección de $Q$ basadas en inercia sugieren que su incremento acumulado entre $Q-1$ y $Q$ es mucho mayor que el incremento entre $Q$ y $Q+1$ [113]. Para cada valor de $Q$ se calcula la proporción de inercia acumulada siguiendo la ecuación 9.1:

$$\frac{\Delta(Q)}{\Delta(Q+1)}, \tag{9.1}$$

, donde $\Delta(Q)$ es el incremento de inercia acumulada entre $Q-1$ y $Q$ [113]. El valor de $Q$ con la menor proporción es seleccionado como el número óptimo de particiones del árbol. Este método equivale a aplicar el *método del codo* [114] de forma automática.

Para consolidar los grupos de individuos, las coordenadas promedio de los individuos dentro de cada partición del árbol se usan como las posiciones iniciales de grupo para el algoritmo de *K-means* [113]. Esta estrategia se ha implementado usando el paquete *FactoMineR* [97], y dispone de una representación para comprender la estructura de los datos e identificar patrones de expresión que pueden pasar desapercibidos cuando se analiza cada CP por separado [115]. Al incorporar la información de múltiples CP de forma simultanea, el AJCP puede mejorar la comprensión de las interacciones entre las variables y la segregación de las muestras.

Una vez agrupados los individuos usando el AJCP, los grupos se pueden anotar con las categorías de las variables cualitativas. En el caso de los experimentos de RNA-seq se dispone de un diseño experimental para calcular la expresión diferencial donde las muestras se dividen en los grupos control y tratamiento. Este etiquetado experimental que se ha incluido en el análisis como la variable cualitativa "treat" puede usarse para asociar de forma estadística los grupos de muestras del AJCP con los grupos del diseño experimental original. Para medir esta asociación, se aplica una prueba exacta de Fisher para calcular la sobrerrepresentación de las muestras de cada condición experimental en los grupos del AJCP. Los valores $p$ obtenidos se ajustan con el método de Benjamini-Hochberg para corregir las pruebas múltiples y los grupos del AJCP se anotan con cada condición experimental sobrerrepresentada con un $FDR$ menor que 0,05.

En este estudio, se ha aplicado el ACP, la proyección de factores experimentales sobre los CP, el AJCP y la anotación de los grupos de muestras dos veces, primero usando todos los genes expresados y luego, usando la expresión de los GEDs como variables activas.

Esta doble aplicación ayuda a comprender cómo contribuyen las variables experimentales a la separación de las muestras. El uso de variables complementarias a la expresión de genes permite explicar la separación de las muestras observada en el análisis de expresión diferencial. De forma específica, se puede analizar si las variables están asociadas al CP1, indicando si la distribución de las muestras en este CP encaja con el diseño experimental. Además, se pueden identificar patrones de expresión secundarios de los GEDs que pueden ser capturados por el CP2 y los componentes subsecuentes. Posteriormente, se puede investigar qué variables experimentales afectan a estos patrones secundarios pudiendo explicar los posibles mecanismos biológicos subyacentes.

Esta metodología se aplico al estudio de los dos conjuntos de datos de enfermedades raras descritos anteriormente: los casos de PMM2-CDG y SSY.

## 9.3. Resultados

En este estudio se ha aplicado la metodología completa del ACP implementada en *degenes_hunter.R* a dos casos reales de enfermedades raras. Este algoritmo, junto con las posteriores fases se aplicó usando dos grupos de genes distintos en cada conjunto de datos como variables activas: la expresión de todos los genes y la de los GED en específico.

### 9.3.1. PMM2-CDG

Este conjunto de datos consiste en cultivos de fibroblastos de pacientes de la cohorte de PMM2-CDG (Capítulo 4). Se comparó la expresión génica de los pacientes con distinta gravedad de la enfermedad, siendo los pacientes con enfermedad leve (*LOW*) el grupo control y los que tenían una enfermedad grave (*HIGH*) el grupo de tratamiento. El análisis de expresión diferencial detectó 12.368 genes expresados, de los cuales 25 son GED.

Cuando el PCA se aplicó a la expresión de todos los genes, se determinaron cuatro CP relevantes que explicaban la mayoría de la varianza sumando un 68,01 % del total (Figura 9.1A). A pesar de que los CP1 y CP2 englobaban casi la mitad de la varianza total, no fueron capaces de diferenciar las muestras según el diseño experimental. No obstante, las muestras con distinta gravedad de la enfermedad se separan levemente en el CP3 (Figura 9.1A). Las variables de edad (*Age*), las medidas y escalas que miden la gravedad de la enfermedad (NCPRS, ICARS y MVRD) y el vector *treat* que describe el diseño experimental se proyectaron sobre el espacio de CP. Se observó que el CP4 es el único componente que correlaciona únicamente con el factor edad (*Age*) con una R de Pearson de 0,64 y que el CP3 es el único componente que se asocia de manera significativa con la variable *treat*,

con un $R^2$ de 0,443 y $p < 0{,}05$ (Tabla 9.1). Usando los CP relevantes se hizo un agrupamiento de los individuos usando AJCP, y se observó que los grupos no son coherentes con el diseño experimental ya que no se pueden diferenciar los pacientes con distinta gravedad (Figura 9.1B). Posteriormente, se intentó asociar de forma estadística los grupos de pacientes y las categorías del diseño experimental (control y tratamiento) pero no se encontró ninguna asociación significativa (Tabla 9.2).

Al usar los niveles de expresión de los GED como variables activas en el ACP, se observó que hay una clara diferenciación entre las muestras de distintos grupos de gravedad. Además, se observa que hay más cohesión entre los pacientes con baja gravedad ($LOW$) que entre los pacientes con alta gravedad ($HIGH$), que se encuentran más dispersos. Como se muestra en la Figura 9.2A, el ACP indica que los dos primeros CP son los más relevantes, englobando el 76,8 % de la varianza. El diseño experimental se asoció, esta vez de forma significativa, al CP1, correspondiendo con la variable *treat*, con un $R^2$ of 0,914 (Tabla 9.1).

En cuanto a las variables cuantitativas, hay una correlación entre el factor edad ($Age$) y el CP1 que, aunque tiene un valor de R de Pearson bajo, es significativo. Además, todas las medidas de gravedad de la enfermedad correlacionan con el CP1 con valores de R altos y significativos (Figura 9.2C).

Al aplicar AJCP se obtuvieron tres grupos de pacientes como se muestra en la Figura 9.2B. Se observó una clara diferenciación entre los pacientes con distinta gravedad ya que los pacientes con baja gravedad se agrupan y se consiguió asociar este grupo con su condición experimental (Tabla 9.2). Además, se puede observar que los pacientes con alta gravedad se dividen en dos subgrupos (Figura 9.2B).

**Tabla 9.1:** Asociación de las variables suplementarias y los Componentes principales (**CP**) de todos los genes expresados (**Exp**) y los genes con expresión diferencial (**GED**) en los conjuntos de datos de PMM2-CDG y el síndrome de Schaaf-Yang (**SSY**). Se calculó un ajuste $R^2$ asociado a un análisis de la varianza para cada variable categórica y se calculó el coeficiente de correlación R de Pearson para cada variable cuantitativa. Solo se muestran asociaciones con un valor P de las respectivas pruebas menor que 0,05.

| **Caso** | **Conjunto de genes** | **CP** | **Variable** | **Medida de asociación** | **Valor** |
|---|---|---|---|---|---|
| PMM2-CDG | Exp | 4 | Age | R de Pearson | 0,64 |
| | Exp | 3 | treat | $R^2$ | 0,443 |
| | GED | 1 | treat | | 0,914 |
| SSY | Exp | 3 | treat | | 0,704 |
| | GED | 1 | treat | | 0,979 |

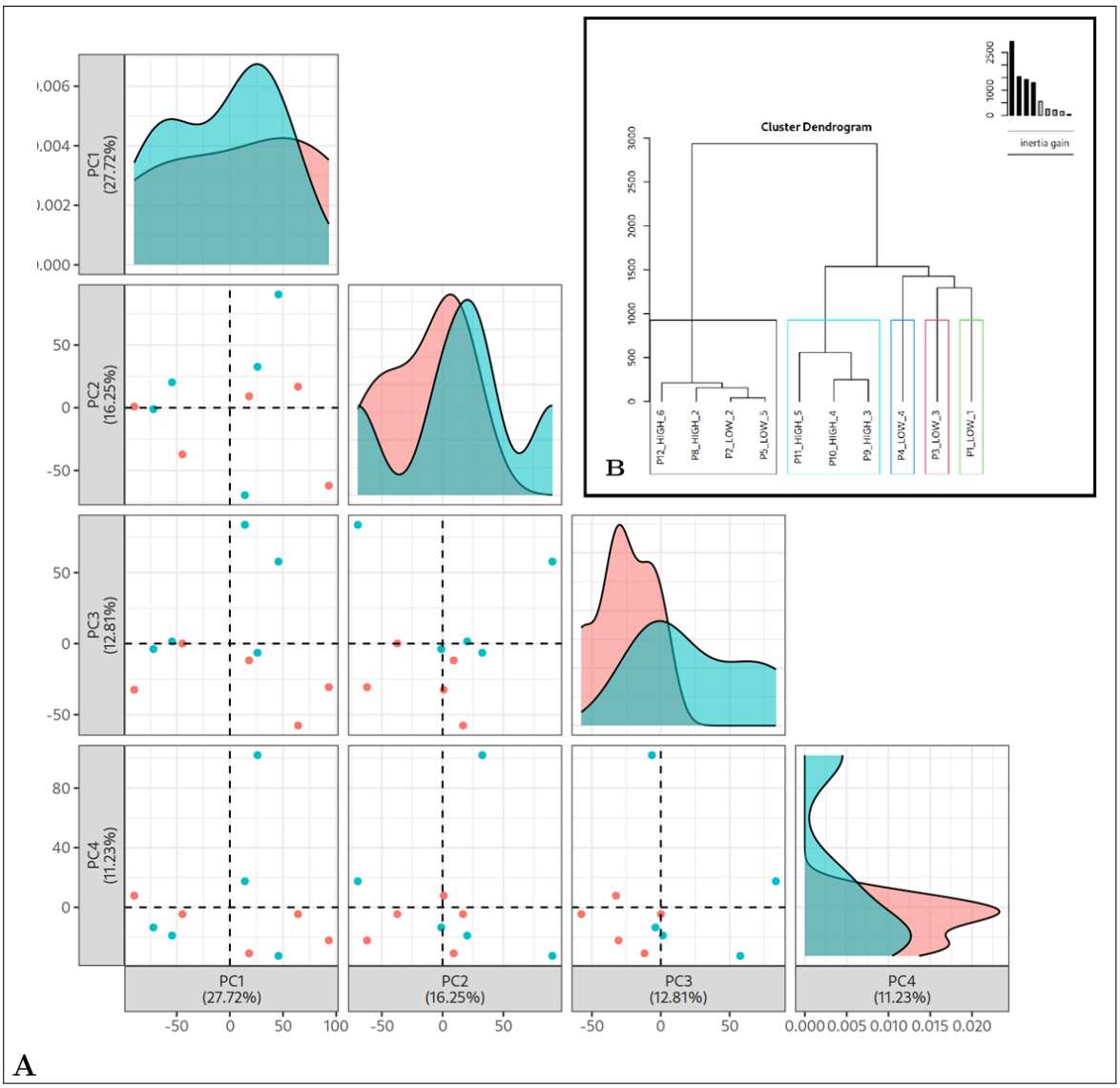


**Figura 9.1:** Análisis de Componentes Principales (ACP) y Agrupamiento Jerárquico de los Componentes Principales (AJCP) del conjunto de datos de PMM2-CDG usando todos los genes. **A:** Las muestras dispuestas en los Componentes Principales (PC) que resumen de manera significativa la mayoría de la varianza total. El porcentaje de varianza resumida por cada CP se muestra entre paréntesis y la distribución de los pacientes con baja y alta gravedad se representa en azul y rojo, respectivamente. **B:** Dendrograma y grupos de muestras resultantes del AJCP.

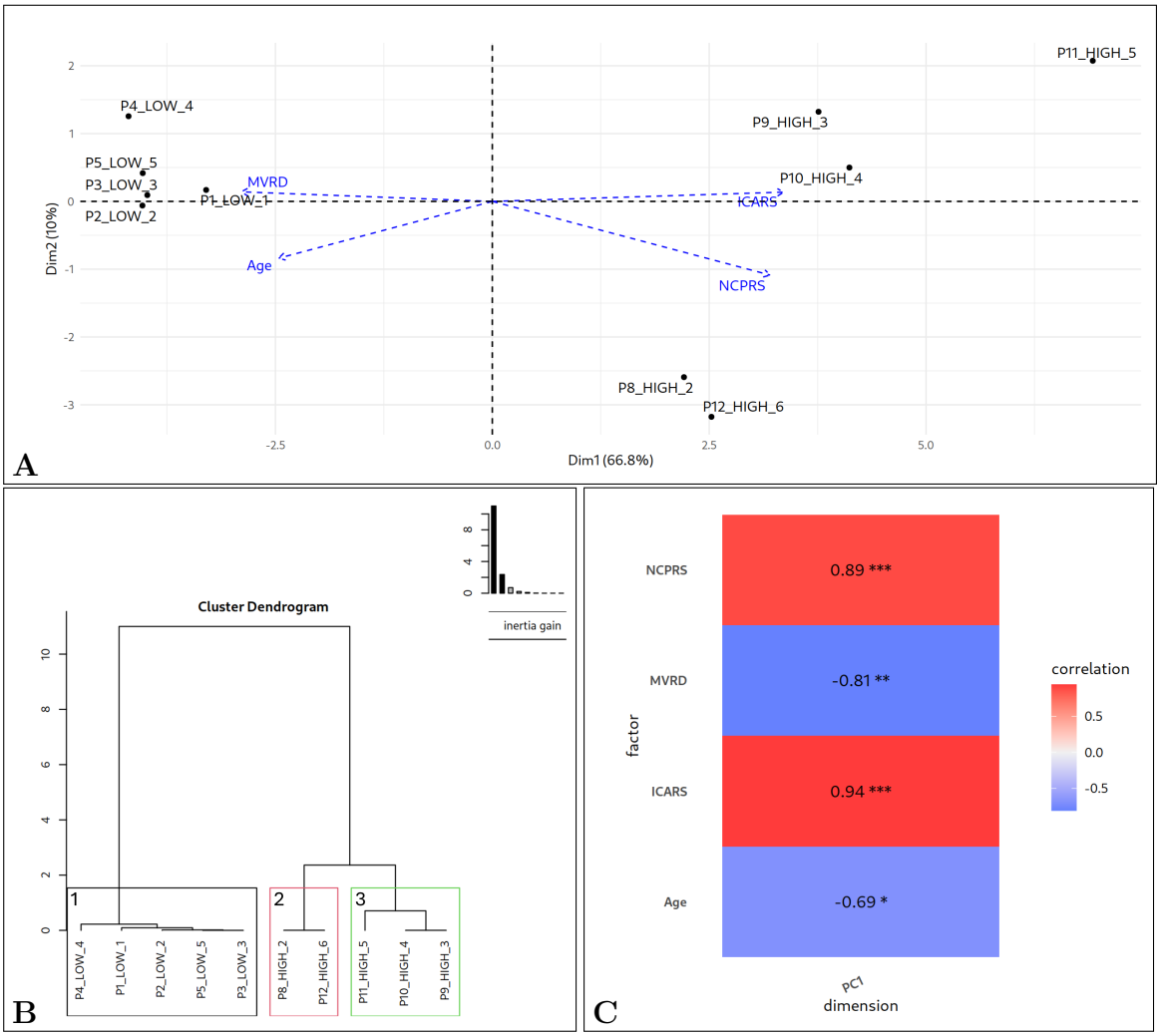


**Figura 9.2:** Análisis de Componentes Principales (ACP) y Agrupamiento Jerárquico de los Componentes Principales (AJCP) del conjunto de datos de PMM2-CDG usando los 25 genes con expresión diferencial. **A:** La disposición de las muestras y las variables suplementarias cuantitativas proyectadas sobre los dos primeros Componentes Principales (CP). El porcentaje de varianza que resume cada CP se muestra entre paréntesis. **B:** Dendrograma y grupos de muestras resultantes del AJCP. **C:** Asociación y significancia de las variables cuantitativas suplementarias y los CP. La significancia se representa con etiquetas siendo *: $0{,}01 < P \leq 0{,}05$, **: $0{,}001 < P \leq 0{,}01$ y ***: $P \leq 0{,}001$

**Tabla 9.2:** Relación entre los grupos de muestras derivados del Agrupamiento Jerárquico de Componentes Principales y las categorías de las variables cualitativas suplementarias. Esta tabla resume los resultados de los casos de **PMM2-CDG** y el síndrome de Schaaf-Yang (**SSY**). Se realizó una prueba exacta de Fisher para obtener el valor P y las pruebas múltiples se corrigieron con el **FDR** de Benjamini-Hochberg. Los distintos **conjuntos de genes** sobre los que se aplicó el ACP son todos los genes expresados (**Exp**) y los expresados de forma diferencial (**GED**). Las variables cualitativas y sus categorías se muestran en la columna **Categoría** separadas por ":" y los grupos del AJCP se muestran en la columna **Grupo**

| Caso | Conjunto de genes | Grupo | Categoría | Valor P | FDR |
|---|---|---|---|---|---|
| PMM2-CDG | Exp | 1 | treat:Treat | 3,97e-03 | 2,38e-02 |
| | | 1 | condition:LOW | 3,97e-03 | 2,38e-02 |
| SSY | Exp | 3 | time:30 | 2,02e-03 | 6,06e-03 |
| | | 2 | time:60 | 2,02e-03 | 6,06e-03 |
| | | 1 | time:90 | 2,02e-03 | 6,06e-03 |
| | GED | 1 | treat:Treat | 1,08e-03 | 6,49e-03 |

### 9.3.2. Síndrome de Schaaf-Yang

El conjunto de datos del síndrome de Schaaf-Yang consiste en 12 muestras que comprende la diferenciación de distintas zonas de cerebro (hCS y hSS), individuos control (S_135) y enfermo (S_66)y tiempos de crecimiento de 30, 60 y 90 días. El análisis de expresión, comparando las muestras de los individuos control y enfermo, identificó 14.745 genes expresados, de los cuales 338 fueron GED.

Se aplicó el ACP a los niveles de expresión de todos los genes y se encontraron tres CP relevantes que contienen un 74,01 % de la varianza total (Figura 9.3A). Se puede observar que la combinación de los CP1 y CP2 separan las muestras por el tiempo de cultivo y que el CP3 separa las muestras por el individuo de origen (Figura 9.3A y B) y se asocia con las condiciones experimentales (variable *treat*) con un $R^2$ de 0,704 (Tabla 9.1).

Este patrón se refuerza con el AJCP que agrupa las muestras en base al tiempo de cultivo y cada grupo corresponde de forma significativa con un tiempo de cultivo del diseño experimental (Figura 9.3C y Tabla 9.2).

El ACP se aplicó a la expresión de los 338 GED y resultó en dos CP relevantes que resumen un 73,94 % de la varianza total (Figura 9.3D). Por un lado, CP1 diferencia los distintos individuos y se asocia al diseño experimental con un $R^2$ de 0,979 (Tabla 9.1). Por otro lado, el CP2 separa las muestras por el tiempo de cultivo (Figura 9.3D). Al aplicar el AJCP usando estos dos CP relevantes se obtuvieron tres grupos de muestras. El primer grupo contiene todas las muestras

derivadas del paciente enfermo, el grupo dos corresponde con las muestras de los organoides del organismo control cultivados 30 días y el grupo tres corresponde con las muestras de los organoides del individuo control cultivados 60 o 90 días (Figura 9.3E).

## 9.4. Discusión

En este estudio se demuestra el potencial del ACP aplicado a datos de expresión génica derivados de la secuenciación, más allá de su uso convencional como control de calidad de muestras, para obtener una perspectiva detallada de ciertos patrones presentes en estos datos. Además, el ACP le da un contexto en el que interpretar estos patrones gracias a los factores experimentales asociadas a las muestras.

Cuando el ACP se aplica a un conjunto de datos, se realiza un proceso de reducción de la dimensionalidad por el cual el número de variables originales, como la expresión de todos los genes o de los GED, se reducen por combinaciones lineales a unos poco ejes ortogonales (los conocidos como componentes principales o CP). Sobre estos CP se pueden proyectar variables cuantitativas y cualitativas. Ambos procesos de reducción y proyección permiten que las variables puedan integrarse y estudiarse dentro del espacio de CP. Esta integración facilita la identificación intuitiva de conexiones ocultas gracias a la visualización de los datos, y permite la aplicación de pruebas estadísticas y técnicas de agrupamiento para la confirmación de relaciones importantes.

En el caso de estudio de PMM2-CDG, cuando se aplicó ACP a todos los genes expresados, los cuatro CP relevantes que se obtuvieron (Figura 9.1A) y su correspondiente AJCP (Figura 9.1B) no diferenciaron los pacientes de distintos grupos experimentales, con la excepción del CP3 que se asocia de forma significativa con los grupos experimentales, aunque con un valor bajo de $R^2$ de 0,443 (Tabla 9.1). Sin embargo, la poca varianza que engloba el CP3 (12,81 %, Figura 9.1A) sugiere que la mayoría de la varianza de los datos se debe a diferencias intragrupales. En consecuencia, dada la separación confusa de las muestras en los CP y grupos del AJCP, se necesita una exploración más exhaustiva para descartar la heterogeneidad de las muestras como la causa principal.

Al aplicar el ACP a la expresión de los GED, se obtuvieron dos CP relevantes que explican el 76.8 % de la varianza total. En el CP1 se puede ver de forma clara la separación entre los pacientes con baja y alta gravedad de la enfermedad (Figura 9.2A). En este caso, el ICARS es la variable cuantitativa con mayor correlación con el CP1, tal y como se esperaba, ya que las muestras se clasificaron en alta o baja gravedad siguiendo esta escala (Figura 9.2C). Un resultado a destacar es que la medida MVRD sigue el mismo patrón, pero en sentido contrario, que el ICARS, siendo ambas variables casi perpendiculares al PC2 (Figura 9.2A), lo que confirma

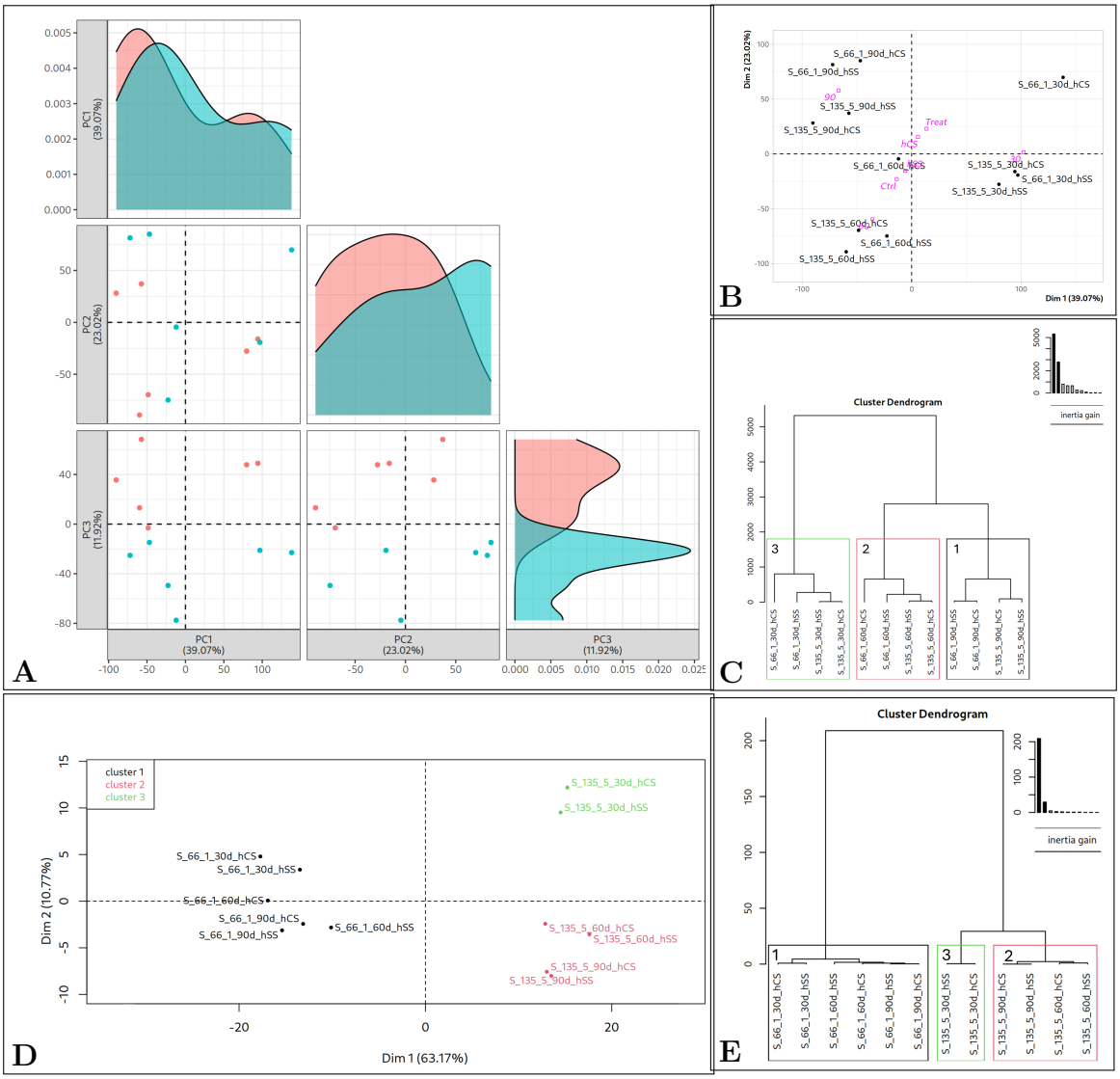


**Figura 9.3:** Análisis de Componentes Principales (ACP) y Agrupamiento Jerárquico de los Componentes Principales (AJCP) del caso de estudio del síndrome de Schaaf-Yang (SSY) usando todos los genes (A,B,C) y los 338 genes expresados de forma diferencial (D,E). **A:** Las muestras se disponen en los CP que resumen de forma significativa la varianza total. El porcentaje de varianza resumido por cada CP se muestra entre paréntesis y la distribución de las muestras de organoides derivadas de los pacientes sano y enfermo se muestran en rojo y azul, respectivamente. **B:** Las muestras se disponen en los dos primeros Componentes Principales (CP) y el porcentaje de varianza resumido por cada CP se muestra entre paréntesis. Los cuadrados fucsia corresponden con cada categoría de las variables cualitativas suplementarias y la posición es la media de coordenadas de las muestras de dicha categoría. **C:** Dendrograma y grupos de muestras resultantes del AJCP. **D:** Las muestras se disponen en los dos primeros Componentes Principales (CP) y el porcentaje de varianza resumido por cada CP se muestra entre paréntesis. El color de las muestras coincide con los grupos determinados por el AJCP **E:** Dendrograma y grupos de muestras resultantes del AJCP.

las observaciones vistas en estudios previos [116]. Esto es relevante para priorizar genes con alta contribución al CP1 para validar experimentalmente su relación con el diámetro medio-sagital del vermis, asociado a la escala de gravedad ICARS. En el agrupamiento con AJCP, no solo se separaron los pacientes de baja gravedad de los de alta gravedad, sino que se se revelaron dos subgrupos de pacientes con alta gravedad (Figura 9.2B y Tabla 9.2). El CP2 es el que muestra de manera clara las diferencias entre los dos subgrupos de pacientes con alta gravedad. El NCPRS y la edad fueron las variables cuantitativas con la mayor asociación al CP2 (Figura 9.2A), lo que sugiere una relación interesante entre ambas variables y el fenotipo de los pacientes. Sin embargo, se necesitan más muestras y la inclusión de datos fenotípicos específicos dado que ninguna de estas relaciones fue significativa.

En el caso de SSY, el ACP sobre los genes expresados en los organoides (diferenciados de la región dorsal y a la región ventral del prosencéfalo para dos individuos distintos y con muestras tomadas a tres tiempos distintos), revela que el tiempo de cultivo tiene un mayor impacto en la diferenciación de las muestras que el individuo de origen o el tipo de organoide (Figura 9.3B). Esto se apoya con que cada grupo del AJCP (Figura 9.3C) corresponde con uno de los tiempos de cultivo (Tabla 9.2). Además, la separación entre las muestras derivadas de los individuos control y enfermo no se observa hasta el CP3, lo que apunta a que esta condición tiene un impacto menor en la varianza total (Figura 9.3A).

Sin embargo, cuando se aplica el ACP a la expresión de los GED, se necesitan solo dos CP para explicar alrededor del 75 % de la varianza total. El CP1 y el AJCP son capaces de separar las muestras derivadas del individuo sano y del enfermo (Figura 9.3D y E). Es interesante comentar que el AJCP es capaz de diferenciar las muestras derivadas del paciente sano por el tiempo de cultivo, pero agrupa las muestras del paciente enfermo juntas. Esto indica que las diferencias que pueda haber en el desarrollo del organoide a distintos tiempos de cultivo en el individuo sano, no se observan en el individuo enfermo. Este ejemplo demuestra la utilidad del ACP como primera aproximación para realizar estudios piloto y generar propuestas para futuros estudios basados en los factores identificados, ya que el tiempo de cultivo es un factor mucho más interesante a tener en cuenta que el tipo de organoide.

La aplicación del ACP en dos conjuntos de datos distintos de enfermedades raras ha facilitado la integración de datos de RNA-seq con una serie de variables experimentales de interés. Con su uso, los investigadores pueden identificar genes asociados a los fenotipos de una enfermedad susceptibles de ser posibles biomarcadores o dianas terapéuticas. La combinación del ACP y AJCP permite un estudio exhaustivo y fácil de interpretar de las relaciones complejas entre los factores genéticos, medioambientales o clínicos y la expresión de los genes. Este tipo de análisis es especialmente relevante en el estudio de enfermedades ra-

ras, ya que la búsqueda de los factores causantes de la enfermedad necesita tener en cuenta un gran espectro de influencias clínicas, genéticas y medioambientales [117]. En estudios futuros, se podría incluir variables adicionales como información demográfica, exposición a la contaminación medioambiental, fenotipos de los pacientes u otras influencias biológicas potenciales. En consecuencia, a modo de adelanto, en el Capítulo 11 se desarrollará la integración de los datos de expresión y fenotípicos del caso de estudio de PMM2-CDG mediante técnicas de reducción de la dimensionalidad derivadas del ACP.

# Capítulo 10

# Estudio de los ARNcirc en pacientes de displasia arritmogénica

## 10.1. Introducción

La displasia arritmogénica engloba un conjunto de trastornos cardíacos con una prevalencia estimada de entre uno y cinco casos por cada 10,000 individuos. Esta enfermedad se caracteriza por la sustitución progresiva del miocardio ventricular por tejido graso o fibroso, lo que puede provocar arritmias potencialmente mortales y un incremento significativo del riesgo de muerte súbita, especialmente en individuos jóvenes [118]. A fin de tomar medidas preventivas personalizadas que no agraven la situación del paciente conforme avanza su edad, la identificación de biomarcadores en pacientes con displasia arritmogénica es crucial [119].

En la última década, el estudio de los ARN circulares (ARNcirc) ha ido en aumento debido a su papel regulador de procesos moleculares ligados a enfermedades [120], así como por sus aplicaciones en biomedicina [44, 121, 122]. Estas moléculas son ARN no codificantes en los que los extremos 5' y 3' se unen de forma covalente mediante un proceso llamado ayuste inverso (también conocido como *back-splicing junction* en inglés) [44]. Los ARNcirc participan en la modulación post-transcripcional de genes diana mediante la unión competitiva a miARN y anulando sus efectos [123].

Los ARNcirc generalmente se analizan mediante protocolos de secuenciación de ARN total, donde no se efectúa una selección específica de los ARNm. La identificación de los ARNcirc a partir de datos de secuenciación consiste en la detección de lecturas generadas por ayuste inverso que mapean de forma quimérica sobre el genoma. Este mapeo quimérico se realiza de forma fraccionada en distintas

regiones de la referencia genómica, y se distingue del ayuste canónico porque una de las fracciones de la lectura mapea de forma inversa a la otra.

Existen numerosas herramientas específicas para identificar ayustes inversos, como *circRNA_finder* [124], *DCC* [125], *CIRCexplorer* [126], *CIRI2* [127] y *Find-Circ* [128], junto a configuraciones específicas de los programas de mapeo, como *TopHat-Fusion* [129], *STAR* [48], *Segemehl* [130] o *BWA-MEM* [131]. Sin embargo, debido a la escasa cantidad de lecturas que suelen pertenecer al ayuste inverso y a la tasa de falsos positivos, hay estudios que demuestran que la mejor opción es combinar dichas herramientas [132].

En el marco del desarrollo de esta tesis doctoral, he desarrollado y aplicado un flujo de trabajo para analizar la expresión e interacciones entre genes, miARN y ARNcirc a partir de muestras de secuenciación de ARN total y de ARN pequeños de pacientes con displasia arritmogénica.

## 10.2. Material y Métodos

Para la búsqueda y estudio de los ARNcirc que puedan tener un potencial impacto en el desarrollo de la displasia arritmogénica, se ha desarrollado un flujo de trabajo que comienza a partir de datos de secuenciación de ARN total y ARN pequeños. Este flujo detecta los genes, ARNcirc y miARN con expresión diferencial, establece las relaciones entre los tres tipos de moléculas, prioriza los ARNcirc mediante su similitud fenotípica con la displasia arritmogénica y realiza un análisis funcional de los genes asociados a ARNcirc (Figura 10.1).

### 10.2.1. Selección de pacientes y secuenciación de ARN

Se secuenciaron tanto el ARN total como los ARN pequeños de muestras del ventrículo derecho de cuatro pacientes fallecidos diagnosticados con displasia arritmogénica, y de otros cuatro individuos que fallecieron por una razón ajena a enfermedades cardíacas que sirvieron como grupo control.

La selección de pacientes, la extracción de tejidos, la preparación de muestras y su secuenciación fue llevada a cabo por colaboradores pertenecientes a grupos de investigación del Instituto de Investigación Biomédica de Girona Josep Trueta (IDIBGI), Instituto de Investigación Sanitaria La Fe de Valencia (IIS La Fe) y el Instituto de Investigación e Innovación Biomédica de Cádiz (INiBICA).

Para el aislamiento y purificación del ARN total se aplicó el protocolo *TRI Reagent*$^{TM}$ (Sigma-Aldrich, St Louis, MO, EE.UU.) según las instrucciones del fabricante. Posteriormente, se aplicó un tratamiento con desoxiribonucleasa usando el kit *RNA clean & concentrator-5* (Zymo Research, Irvine, CA, EE.UU.). El ARN se cuantificó usando el kit *Qubit RNA High-Sensitivity Assay* en el fluorómetro

*Qubit®2.0* (Life Technologies, Carlsbad, CA, EE.UU.). La calidad e integridad del ARN total se controló en el *Agilent Technologies 2100 Bioanalyzer* (Agilent Technologies, Santa Clara, CA, EE.UU.).

La secuenciación del ARN total y de los ARN pequeños se llevó a cabo en el servicio GeneCore (EMBL Heidelberg, Alemania). Por un lado, las librerías *single-end* de ARN total se prepararon con el kit *NEBNext Ultra II Directional RNA Library Prep* para plataformas de *Illumina*, aplicando el kit de *rRNA Depletion* (New England Biolabs, Ipswich, MA, EE.UU.) para ser secuenciadas posteriormente en la plataforma *Illumina SE100*, generando archivos fastq con lecturas de 130 pb. Por otro lado, las librerías *single-end* de ARN pequeños se generaron con el kit *NEXTFLEX small RNA-seq v3* (Perkin Elmer, Waltham, MA, EE.UU.) y se secuenciaron empleando la plataforma *Illumina SE60*, generando archivos fastq con lecturas de 70 pb.

### 10.2.2. Análisis de miARN-genes diana

Las interacciones entre los miARN y sus genes dianas se obtuvieron mediante el flujo de expresión y de detección de dianas de miARN desarrollado en los capítulos 6 y 7. La cuantificación de los miARN, a partir de las muestras de secuenciación de ARN pequeños, y la de los genes, a partir de las muestras de secuenciación de ARN total, se hizo por separado usando el flujo de trabajo *DEG_workflow* [133]. El análisis de la expresión diferencial y de la coexpresión se llevó a cabo con el *script degenes_hunter.R* de *ExpHunterSuite* [89], configurado para que los miARN y genes expresados de forma diferencial fueran aquellos detectados por los paquetes de análisis de expresión *DESeq2* [52] y *edgeR* [53], con la siguiente configuración: $|Log_2FC| > 1$ y FDR $< 0{,}05$. El análisis de coexpresión se llevó a cabo empleando el paquete *WGCNA* [55], configurando los parámetros por defecto de *degenes_hunter.R* [89]. La identificación de las relaciones miARN-gen diana se realizó con *coRmiT.R* (cuyo algoritmo se describe al detalle en el Capítulo 7), configurado para establecer relaciones entre los miARN y genes expresados de forma diferencial (GED) y devolver aquellas asociaciones detectadas por el método de selección e integración. Además, *coRmiT.R* se ejecutó para obtener las parejas con correlación negativa y positiva, usando umbrales de $|R|$ de Pearson de 0,5 a 0,95 en rangos de 0,05. Los detalles de la ejecución del flujo se pueden encontrar en el repositorio https://github.com/JoseCorCab/coRmiT_ceRNA.

### 10.2.3. Detección y análisis de la expresión de ARNcirc

El análisis de ARNcirc se llevó a cabo empleando los archivos de lecturas obtenidos a partir de las muestras de secuenciación de ARN total (con lecturas de 130 pb). Un de las características de las herramientas de detección de ARNcirc es

que devuelven una elevada tasa de falsos positivos. A fin de reducirla y obtener resultados más fiables, se utilizó el programa *CirComPara2*, que combina los resultados de varios algoritmos especializados en la identificación de ARNcirc, tales como *circRNA_finder*, *DCC*, *CIRCexplorer*, *CIRI2* y *FindCirc*, para minimizar la tasa de falsos positivos [134]. De los ARNcirc encontrados por *CirComPara2* se seleccionaron aquellos detectados por más de uno de los algoritmos de identificación. Los ARNcirc seleccionados se cuantificaron con *CirComPara2* teniendo en cuenta exclusivamente las lecturas correspondientes a la región del ayuste inverso.

La expresión diferencial de los ARNcirc se calculó con *ExpHunterSuite* usando el paquete *limma* [46]. Los ARNcirc expresados de forma diferencial fueron aquellos con $|Log_2FC| > 1$ y FDR $< 0{,}05$.

### 10.2.4. Estudio de los ARNcirc y sus genes diana

Debido a la escasez de información sobre relaciones entre miARN y ARNcirc, se usó la herramienta *Circr* [135] para predecir las relaciones entre los miARN y los ARNcirc expresados de forma diferencial. *Circr* hace uso de tres predictores *de novo* de interacciones entre miARN y ARN, tales como *miRanda*, *RNAhybrid* y *TargetScan*. Se seleccionaron aquellas interacciones entre miARN y ARNcirc obtenidas por al menos dos de los tres predictores.

Las relaciones miARN-genes detectadas con *coRmiT* y las relaciones miARN-ARNcirc de *Circr* se combinaron para establecer relaciones genes-ARNcirc. Finalmente, para priorizar ARNcirc para su validación experimental en futuros ensayos, se decidió hacer estudios funcionales y fenotípicos de los genes asociados a cada ARNcirc. Con el fin de estudiar la relevancia fenotípica de los ARNcirc, se calculó la similitud semántica entre los genes asociados a cada ARNcirc y el perfil fenotípico descrito para la displasia arritmogénica. En primer lugar, se empleó la base de datos Monarch Initiative [83] para obtener la información correspondiente a los fenotipos patológicos descritos para cada uno de los subtipos de displasia arritmogénica conocidos. Estos fenotipos aparecen descritos en esta base de datos como términos de la *Human Phenotype Ontology* (*HPO*) [53], y se seleccionaron aquellos que coincidían con el fenotipo de los pacientes analizados en este estudio.

En segundo lugar, se empleó de nuevo la información almacenada en la *Monarch Initiative* para obtener los términos *HPO* descritos de cada uno de los genes asociados a los ARNcirc. Finalmente, se calculó con la herramienta *semtools* (Capítulo 4) la similitud semántica empleando el método de Lin entre los *HPO* de cada gen y los *HPO* de la enfermedad. Para cada ARNcirc, se calculó la media de la similitud semántica entre todos los genes asociados, designándolo bajo el término "similitud fenotípica". Se escogieron los diez primeros ARNcirc con mayor similitud fenotípica y se llevó a cabo el análisis de enriquecimiento de los genes asociados a cada uno de ellos en *Gene Ontology* (*GO*), empleando el programa incluido en

*ExpHunterSuite*, *clusters_to_enrichments.R*, considerando como enriquecimientos significativos aquellos con una $FDR < 0{,}05$. Se activó el parámetro *-c* que elimina los parentales de un término GO que ha sido asociado de forma significativa a los genes de un ARNcirc, simplificando los resultados.

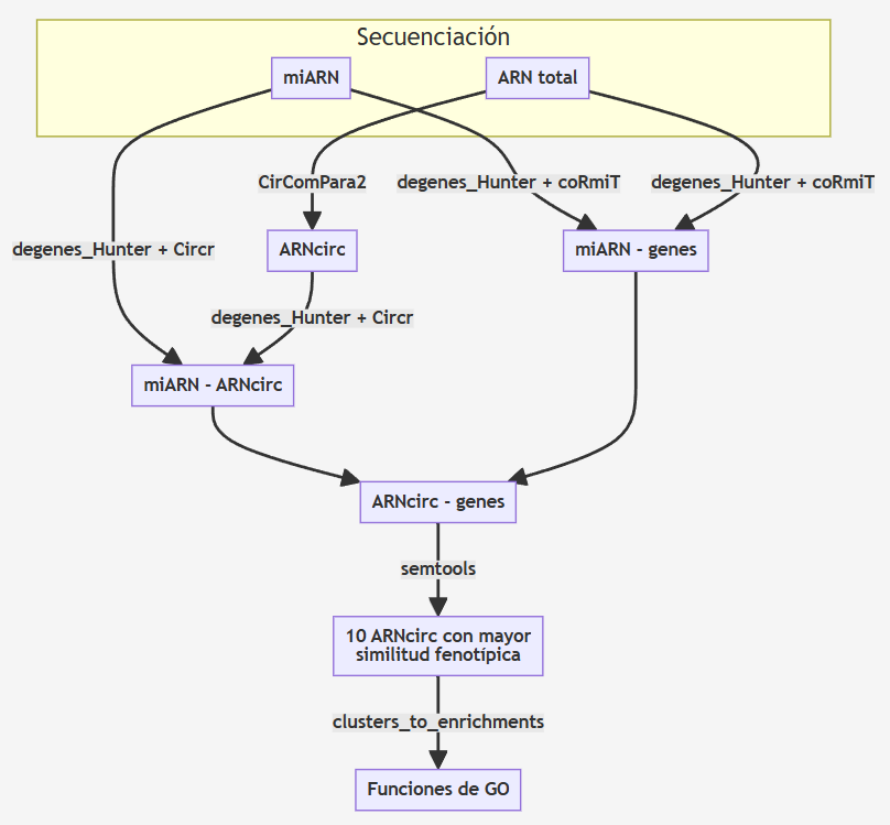


**Figura 10.1:** Diagrama que resume el flujo de trabajo para analizar los ARNcirc. En morado se muestran los distintos datos que se usan o generan en diferentes pasos del flujo y en gris se indican los programas usados.

## 10.3. Resultados

La búsqueda de ARNcirc asociados a la displasia arritmogénica del ventrículo derecho consistió en el análisis independiente de la expresión de miARN, genes y ARNcirc, hasta el análisis de las funciones de los genes dianas de ARNcirc, como se muestra en la figura 10.1.

Por un lado, el análisis de la expresión de miARN y de genes reveló 9 miARN y 1793 genes con expresión diferencial en la comparación hecha entre los controles

con respecto a los enfermos. La estrategia de selección e integración de *coRmiT* identificó 3627 relaciones entre los miARN y GEDs, de las cuales 1363 parejas correlacionaron de forma negativa y 2264 de forma positiva.

Por otro lado, el empleo de *CirComPara2* sobre los datos de secuenciación de ARN total identificó 3822 ARNcirc expresados, de los cuales 44 lo hicieron de forma diferencial entre controles y enfermos. La aplicación de *Circr* determinó 197 parejas miARN-ARNcirc detectadas por un mínimo de dos predictores. Las parejas miARN-genes y miARN-ARNcirc se combinaron para obtener un total de 33 770 relaciones entre 41 ARNcirc y 1029 genes.

De los genes asociados a ARNcirc se identificaron 242 con anotaciones en términos fenotípicos de la *HPO* empleando la base de datos *Monarch Initiative*. Estas anotaciones se emplearon para calcular la similitud semántica entre los perfiles fenotípicos de los genes y el perfil fenotípico de la enfermedad. Se pudo calcular la similitud fenotípica de todos los ARNcirc con la excepción de hsa_circ_0056810 y hsa_circ_0030254, cuyos genes asociados no tenían anotaciones en *HPO* (Tabla 10.1). Además, es interesante destacar que los siete ARNcirc con mayor similitud fenotípica presentaron anotaciones en la *HPO* únicamente para uno de sus genes asociados: *ALG10B*.

Todos los genes asociados a los diez ARNcirc con mayor similitud fenotípica, independientemente de si presentaban o no anotaciones en la *HPO*[1], se emplearon para llevar a cabo un enriquecimiento funcional en términos *GO* (Tabla 10.1, columna **Nº genes**). Con respecto a los resultados obtenidos para la subontología *Biological Processes* de *GO*, el análisis de enriquecimiento funcional reveló que dos de los diez ARNcirc tuvieron enriquecimientos significativos. En concreto, chr2:178793403-178793541 tiene un impacto en las funciones relacionadas con la regulación del potencial de acción del músculo cardíaco, del ensamblaje de la cromatina, de la hidroxilación de proteínas y de la morfogénesis de células del sistema nervioso (Figura 10.2). Por otro lado, el hsa_circ_0028899 afecta a la regulación de los procesos metabólicos de ARN ribosómico, de la traducción en respuesta al estrés ligada a p53 y del ayuste de ARN (Figura 10.2).

[1]Acalaración: Los ARNcirc tienen asociados genes, algunos con HPO (que se usaron para el calculo de la similitud) y otros no. Para el enriquecimiento se usaron todos

**Tabla 10.1:** Listado de los ARNcirc expresados diferencialmente y asociados con genes. La lista incluye los ARNcirc ordenados de forma descendente según su valor de similitud fenotípica. Se muestra la posición priorizada, el nombre del ARNcirc (**ARNcirc**), el número de genes asociados (**Nº genes**), el número de genes con anotación fenotípica (**Nº genes con *HPO***) y la similitud fenotípica (**Similitud fenotípica**). Los 10 ARNcirc seleccionados para el análisis funcional están marcados con una línea horizontal.

| Posición | ARNcirc | Nº genes | Nº genes con *HPO* | Similitud fenotípica |
|---|---|---|---|---|
| 1 | hsa_circ_0030378 | 26 | 1 | 0,689 |
| 2 | hsa_circ_0009024 | 21 | 1 | 0,689 |
| 3 | hsa_circ_0007444 | 19 | 1 | 0,689 |
| 4 | hsa_circ_0002266 | 21 | 1 | 0,689 |
| 5 | hsa_circ_0001492 | 24 | 1 | 0,689 |
| 6 | hsa_circ_0001414 | 21 | 1 | 0,689 |
| 7 | chr6:129366218-129427854 | 24 | 1 | 0,689 |
| 8 | hsa_circ_0000095 | 10 | 3 | 0,384 |
| 9 | chr2:178793403-178793541 | 5 | 3 | 0,384 |
| 10 | hsa_circ_0028899 | 354 | 69 | 0,234 |
| 11 | hsa_circ_0142214 | 1228 | 270 | 0,232 |
| 12 | hsa_circ_0001189 | 1221 | 270 | 0,232 |
| 13 | chr9:109080838-109091155 | 1223 | 270 | 0,232 |
| 14 | chr1:247156405-247159813 | 1223 | 270 | 0,232 |
| 15 | chrX:129553235-129554446 | 881 | 199 | 0,232 |
| 16 | hsa_circ_0005900 | 1214 | 267 | 0,231 |
| 17 | hsa_circ_0001329 | 862 | 198 | 0,229 |
| 18 | chr19:5604582-5604936 | 335 | 68 | 0,227 |
| 19 | chr12:12244261-12244655 | 333 | 68 | 0,227 |
| 20 | hsa_circ_0142312 | 1362 | 304 | 0,227 |
| 21 | chr8:140846259-140890769 | 1362 | 304 | 0,227 |
| 22 | chr8:140830471-140890769 | 1362 | 304 | 0,227 |
| 23 | chr7:16258389-16278226 | 1362 | 304 | 0,227 |
| 24 | chr6:136694139-136698682 | 1362 | 304 | 0,227 |
| 25 | chr4:105424195-105456745 | 1362 | 304 | 0,227 |
| 26 | chr3:47067069-47098081 | 1362 | 304 | 0,227 |
| 27 | chr2:178650750-178689896 | 1362 | 304 | 0,227 |
| 28 | chr1:8465924-8614686 | 1362 | 304 | 0,227 |
| 29 | chr16:57164000-57173869 | 1362 | 304 | 0,227 |
| 30 | chr15:59195460-59218087 | 1362 | 304 | 0,227 |
| 31 | chr14:23404295-23431481 | 1362 | 304 | 0,227 |
| 32 | chr10:49915916-49935142 | 1362 | 304 | 0,227 |
| 33 | chr10:45939360-45962994 | 1362 | 304 | 0,227 |
| 34 | chr4:148152468-148154901 | 1341 | 303 | 0,226 |
| 35 | chr17:4066446-4072756 | 1343 | 303 | 0,226 |
| 36 | hsa_circ_0009964 | 1352 | 301 | 0,226 |
| 37 | hsa_circ_0000284 | 1001 | 232 | 0,223 |
| 38 | hsa_circ_0000701 | 481 | 105 | 0,219 |
| 39 | hsa_circ_0005955 | 134 | 34 | 0,187 |
| 40 | hsa_circ_0056810 | 2 | 0 | N/A |
| 41 | hsa_circ_0030254 | 2 | 0 | N/A |

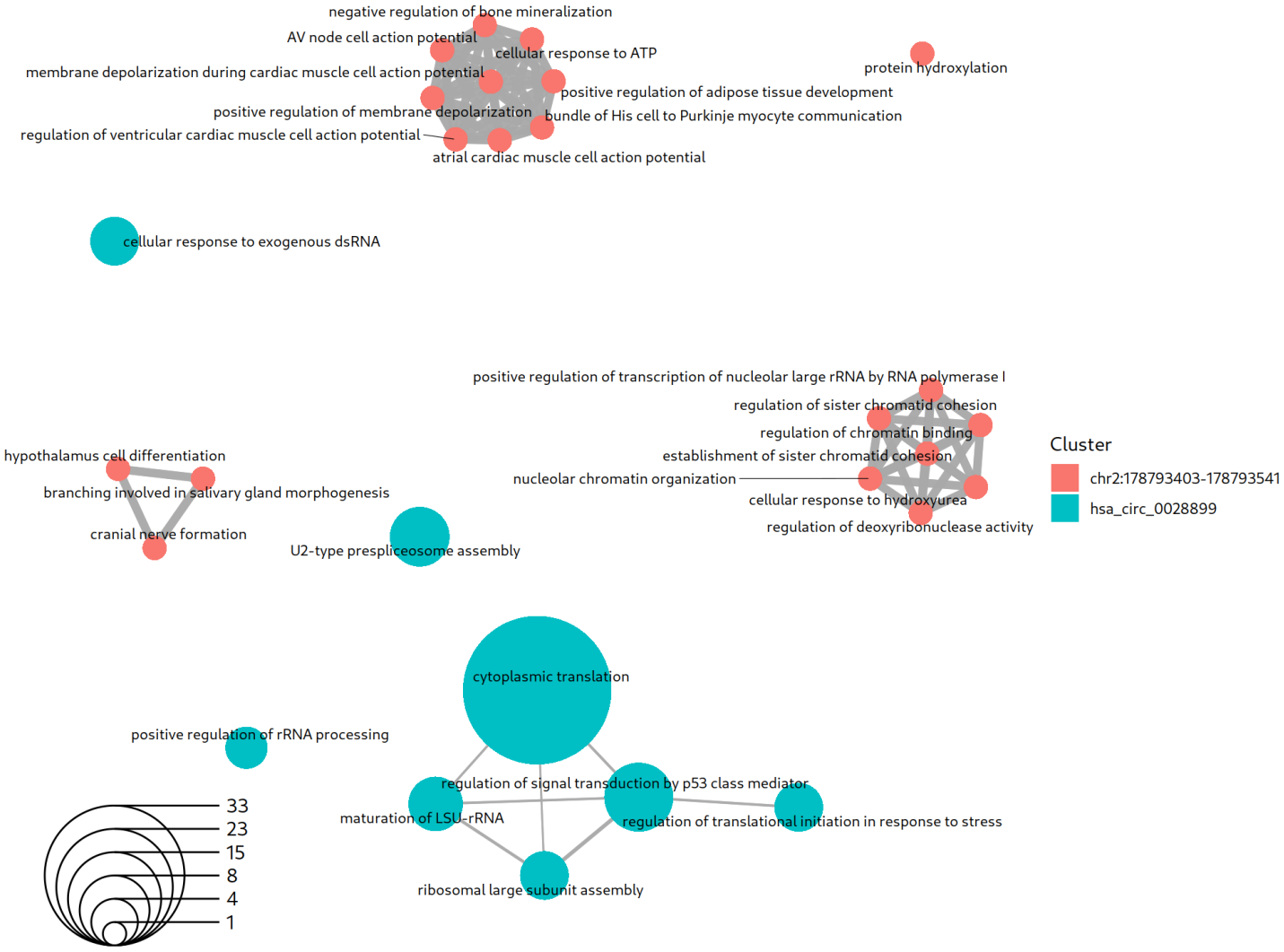


**Figura 10.2:** Términos de la *Gene Ontology* para la subontología *Biological Processes* y sus relaciones con los genes asociados a los diez ARNcirc con mayor similitud fenotípica a la displasia arritmogénica del ventrículo derecho. Los círculos representan los términos de la ontología, donde el color indica la proporción de genes asociados a cada ARNcirc, y el tamaño refleja la cantidad de genes (pertenecientes a cualquier ARNcirc) anotados en dicho término, tal y como se indica en las leyendas. Los enlaces entre los términos indican la presencia de genes anotados en ambos términos, siendo el grosor de los enlaces y la proximidad entre los términos indicativos de la cantidad de genes compartidos.

## 10.4. Discusión

Las herramientas de detección de ARNcirc y los predictores de parejas miARN-ARNcirc suelen presentar una alta tasa de falsos positivos, por lo que es necesario obtener un consenso a partir de los resultados de varias de ellas [135, 134, 132]. Esto es incluso más importante de considerar en estudios donde se dispone de escasas muestras a analizar, como es el caso de las enfermedades raras, en las cuales se suele observar una tendencia a ausencia de ARNcirc asociados a la patología.

Se debe considerar que la existencia de un trío de regulación ARNcirc-miARN-gen, con el gen y el ARNcirc compitiendo de forma endógena en la unión con el miARN, aún con cambio en los niveles de expresión de ambos miARN y ARNcirc, no implica necesariamente un cambio de expresión en el gen [136]. Sin embargo, se optó por buscar asociaciones con los GEDs para localizar aquellos implicados en procesos celulares con un posible impacto en el fenotipo patológico de la enfermedad. Además, se decidió buscar parejas miARN-gen con correlación tanto positiva como negativa ya que la intervención del ARNcirc puede disminuir el efecto inhibitorio de los miARN sobre los genes o neutralizarlo por completo. Cuando el ARNcirc actúa sobre el mecanismo de represión de miARN-gen, no es posible conocer el efecto real del miARN en la expresión del gen. Por eso se exploran las correlaciones positivas y negativas.

La priorización de los genes asociados a ARNcirc mediante sus fenotipos relacionados depende, en gran medida, de la información disponible en las bases de datos. En este caso de estudio, tan solo el 23,5 % de los genes asociados a ARNcirc presentaban anotaciones en la *HPO*. Se espera que con el aumento de la información disponible en estas fuentes, promovidas mayoritariamente por el esfuerzo de estudios clínicos e iniciativas internacionales como la *Monarch Initiative*, mejoren los resultados obtenidos por este tipo de aproximaciones. No obstante, el cálculo de la similitud fenotípica ha permitido reducir la lista de ARNcirc y genes asociados candidatos para su validación experimental (Tabla 10.1, Columna **Nº genes**).

Con respecto al análisis funcional de los genes asociados a los ARNcirc cabe destacar que el chr2:178793403-178793541 podría considerarse como una posible diana terapéutica gracias a las funciones de regulación del potencial de acción del músculo cardíaco que presentan sus genes asociados, ya que es un proceso directamente involucrado en el desarrollo de la displasia arritmogénica del ventrículo derecho.

El hsa_circ_0028899 (también llamado circRNF10) afecta a procesos moleculares generales como el ayuste de ARN y la traducción ligada a p53, que coincide con los estudios que relacionan a este ARNcirc con la regulación de tumores [137, 138]. Sin embargo, a pesar de que no se ha encontrado una relación clara entre las funciones asociadas y la displasia arritmogénica, se puede explorar en un futuro tanto la relación entre genes específicos asociados a este ARNcirc como su relación con la

enfermedad. También hay que destacar que la mayoría de los ARNcirc con mayor similitud fenotípica no han sido enriquecidos de forma significativa en términos *GO*. No obstante, es interesante señalar que hay un único gen anotado en *HPO* asociado a los siete primeros ARNcirc, *ALG10B*. Este gen funciona como regulador del potencial de acción cardíaco en enfermedades como el síndrome de QT largo [139].

El desarrollo y aplicación de este flujo de trabajo ha resultado ser una buena opción para encontrar ARNcirc expresados de forma diferencial en los pacientes de displasia arritmogénica. Esto es especialmente útil para priorizarlos y seleccionar candidatos de cara a una validación experimental en el contexto de identificar posibles biomarcadores o dianas terapéuticas.

# Parte IV

# Búsqueda de patrones de expresión causantes de fenotipos patológicos

# Capítulo 11

# Integración de los datos de pacientes con PMM2-CDG para evaluar la severidad de la enfermedad

## 11.1. Introducción

El análisis de los fenotipos y la expresión génica de pacientes de PMM2-CDG, descritos en las partes II y III, respectivamente, de esta tesis doctoral, son dos aproximaciones independientes capaces de ampliar el conocimiento de las enfermedades raras. Por un lado, los fenotipos patológicos son términos que se usan para caracterizar a una enfermedad o la gravedad de la misma. La naturaleza categórica de los fenotipos ha permitido buscar parecidos y diferencias entre pacientes o enfermedades (Capítulos 4 y 5) mediante el cálculo de la similitud semántica para, posteriormente, proceder a su estratificación en subgrupos. Por otro lado, en esta tesis doctoral se han analizado datos de expresión para identificar genes expresados diferencialmente en comparaciones entre casos y controles (Capítulos 6, 7 y 10) que permiten comprender el desarrollo de la enfermedad. Además, estos genes con expresión diferencial se han usado, de forma análoga a los fenotipos, para encontrar subgrupos de pacientes o muestras experimentales al aplicar ACP (Capítulo 9). Esto se puede considerar un análisis muy similar al realizado con los fenotipos pero estratificando los pacientes por los patrones de expresión.

En el caso de la enfermedad PMM2-CDG se dispone de datos fenotípicos y transcriptómicos de los mismos pacientes (Capítulos 4, 7 y 9). Por un lado, existen métodos que específicamente integran la expresión génica y los fenotipos, tales como *ATHENA* [140], *SKMsmo* [141] y *Weka 3* [142]. Además de estos tres, hay

otros métodos que llevan a cabo la misma función pero que requieren de una gran cantidad de datos para hacer la correcta integración de los mismos [46]. Por otro lado, se han desarrollado métodos para la integración de múltiples fuentes de datos ómicos provenientes de las mismas muestras, como *mixOmics* [143] o *BCC* [144]. Estos enfoques se basan, respectivamente, en la selección de variables y en el uso de estadística bayesiana [145]. No obstante, las herramientas mencionadas se centran principalmente en la detección de asociaciones entre variables y de muestras o pacientes, pero suelen sacrificar la facilidad de interpretación de los resultados, en contraste con el Análisis de Factores Múltiples (AFM) [146, 145]. La integración de datos fenotípicos y transcriptómicos puede realizarse mediante la técnica de reducción de dimensionalidad implementada en el AFM, lo que permite explorar la contribución de cada tipo de dato en un espacio de dimensiones reducidas. Este enfoque facilita la estratificación de los pacientes al considerar simultáneamente los datos de expresión génica y los fenotipos, proporcionando una visión más completa de la enfermedad.

En base a todo lo expuesto, he diseñado y aplicado un flujo de trabajo que aplica la herramienta *Cohort Analyzer* y nuestro paquete de R *ExpHunterSuite*. En el caso de este último, he expandido el código para incluir los análisis basados en ACP (Capítulo 9) pero modificados para tratar datos de distinta naturaleza. He incorporado el Análisis de Correspondencias Múltiples (ACM) para diferenciar grupos de pacientes según su fenotipo, y el AFM para el estudio integrado de fenotipos y genes con expresión diferencial. Esto nos permitirá analizar la agrupación de los pacientes según su perfil de expresión génica, los fenotipos observados y la gravedad de la enfermedad. Finalmente, se podrán identificar los fenotipos y genes más relevantes en la determinación de la severidad de la enfermedad.

## 11.2. Material y métodos

### 11.2.1. Preparación de los datos de pacientes

La cohorte de PMM2-CDG consta de 27 pacientes fenotipados con las anotaciones de la *HPO*, cuya descripción y análisis se desarrolló en el Capítulo 4. A diez de estos pacientes se les realizó una secuenciación de ARN tal como se describe en los Capítulos 7 y 9. Estos pacientes fueron clasificados en dos grupos "LOW" y "HIGH" según la gravedad de la enfermedad en la escala internacional cooperativa de la ataxia (ICARS por sus siglas en inglés). Los valores de expresión génica fueron cuantificados mediante la aplicación del flujo de trabajo *DEG_workflow*[1] (Capítulo 7 y 9) y el análisis de expresión diferencial se llevó a cabo con el paquete

[1] https://github.com/seoanezonjic/DEG_workflow

de R *ExpHunterSuite*, haciendo la comparación entre los grupos de muestras etiquetadas como "LOW" y "HIGH", tal y como se describe en el Capítulo 9. En este caso, el análisis se realizó teniendo en cuenta las variantes patogénicas características de ambos alelos del gen *PMM2*, así como factores cualitativos suplementarios. Dado que todos los pacientes están diagnosticados con PMM2-CDG, se eliminaron aquellos fenotipos patológicos comunes con el fin de maximizar las diferencias entre ellos. Tanto los fenotipos de los pacientes como los genes con expresión diferencial (GEDs) fueron el punto de inicio de este estudio.

### 11.2.2. Estudio de los pacientes y sus fenotipos

El estudio de los pacientes a nivel fenotípico se realizó en dos pasos. En primer lugar se aplicó *Cohort Analyzer* (descrito en Capítulo 4) para evaluar las diferencias fenotípicas entre los pacientes de ambos grupos. Este programa se configuró para agrupar a los pacientes según el cálculo de la similitud semántica por el método de Lin [14] entre todos los perfiles fenotípicos. Además, esta herramienta se utilizó para representar los fenotipos asociados a los grupos de pacientes. La representación se llevó a cabo mediante la creación de un perfil fenotípico grupal, que integra todos los fenotipos de los pacientes pertenecientes al grupo. Posteriormente, se calculó la similitud semántica entre cada fenotipo individual de los pacientes y el perfil fenotípico grupal, proporcionando una caracterización más precisa y representativa de cada agrupación.

En segundo lugar se aplicó un Análisis de Correspondencias Múltiples (ACM) sobre los fenotipos. El ACM es un análisis de reducción de la dimensionalidad que se puede definir como una extensión del ACP para datos categóricos. El procedimiento completo del ACM se encuentra descrito en el libro *Multiple Factor Analysis by Example Using R* [146], no obstante, a continuación se describe dicho procedimiento de forma simplificada: En lugar de utilizar datos cuantitativos como el ACP, el ACM convierte cada categoría de las variables cualitativas en una tabla de posesión binaria, donde cada fila representa una categoría, con valores 0 o 1 según la pertenencia de cada individuo, que están representados en las columnas (Figura 11.1). Esta tabla de posesión se usa para obtener la proporción de individuos que poseen cada categoría, que se calcula dividiendo el número de individuos con dicha categoría (la suma de la fila de la tabla binaria) por el total de individuos. La tabla de posesión se transforma en una tabla de posesión relativa al dividir cada uno de sus valores (0 o 1) por la proporción de individuos de su categoría y restándole el valor 1 para centrar los datos. Estas posesiones relativas serán mayores para las categorías menos frecuentes, dándoles mayor peso en términos de varianza, y 0 para aquellas que posean todos los individuos (Figura 11.1) [146, *Multiple Correspondence Analysis*]. Sobre la tabla de posesión relativa, el ACM aplica un proceso similar al ACP. Se buscan las combinaciones lineales de

las categorías, conocidas como componentes principales (CP), que maximicen la varianza. Este análisis geométrico permite encontrar patrones de asociación entre individuos y categorías en un espacio de dimensiones reducidas ya que la distancia entre puntos en este espacio refleja la similitud entre individuos y la co-ocurrencia de categorías [146, *Multiple Correspondence Analysis*]. Al igual que ocurre con el ACP, se pueden proyectar factores suplementarios sobre el espacio de los CP. Sin embargo, cuando el ACP proyecta las categorías calculando la media de coordenadas de los individuos que la poseen (Capítulo 9), el ACM proyecta las frecuencias de posesión de cada categoría. El cálculo de asociaciones entre los factores suplementarios y los CP, y el análisis estadístico asociado es igual al descrito para el ACP en el Capítulo 9.

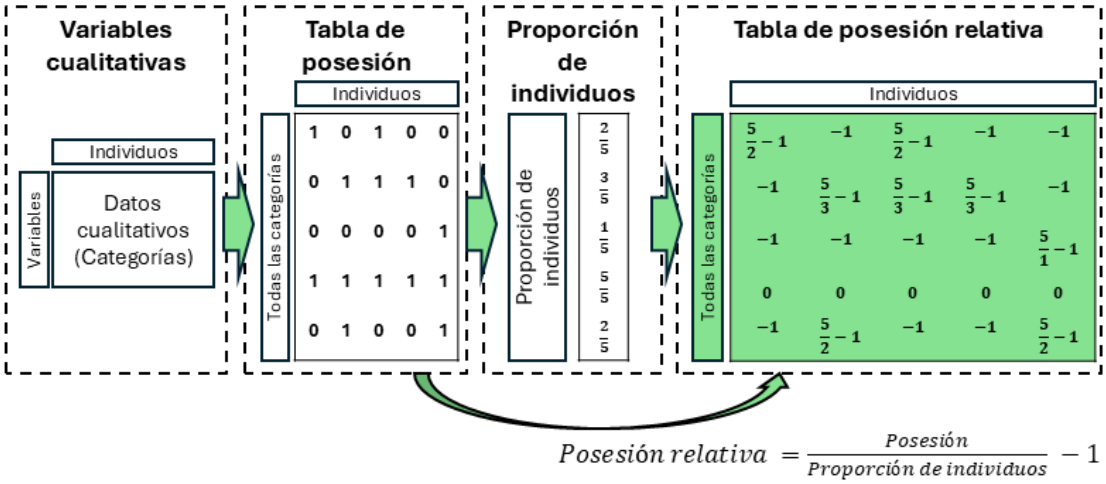


**Figura 11.1:** Ejemplo del cálculo de la posesión relativa que realiza el Análisis de Correspondencias Múltiples (ACM). A partir de una tabla de variables cuantitativas se construye una tabla de posesión que se emplea para calcular la proporción de individuos que poseen cada categoría. Los valores de posesión se dividen entre la proporción de individuos y se les resta el valor 1 para obtener la posesión relativa.

El ACM del paquete *FactoMineR* [97] se implementó en el programa *multivar_mine.R*, integrado en el paquete *ExpHunterSuite* [89, 81]. Esta implementación permite proyectar factores experimentales sobre los CP, realizar un agrupamiento jerárquico de componentes principales (AJCP) y establecer relaciones estadísticas entre los grupos de pacientes y el diseño experimental. En este caso, se utilizó *multivar_mine.R* con los datos fenotípicos de los pacientes, configurando su ejecución para usar la clasificación de grupos de gravedad ("LOW" y "HIGH"). Las variantes patogénicas de ambos alelos del gen *PMM2* se incluyeron como factores cualitativos suplementarios, y las escalas de gravedad ICARS, la escala de puntuación pediátrica de desordenes congénitos de la glucosilación de Nijmegen (NPCRS

por sus siglas en inglés) y el diámetro relativo medio-sagital del vermis (MVRD por sus siglas en inglés) como factores cuantitativos suplementarios.

### 11.2.3. Análisis integrado de fenotipos y genes

El AFM es una técnica de reducción de dimensionalidad diseñada para analizar datos de individuos compuestos por dos o más grupos de variables. Estos grupos pueden ser cuantitativos o cualitativos[2].

El ACP se centra en variables cuantitativas, mientras que el ACM trabaja con variables cualitativas. En cambio, el AFM permite combinar grupos de variables tanto cuantitativas como cualitativas, lo que posibilita analizar simultáneamente datos de distinta naturaleza en un espacio común de dimensiones reducidas o CP [146]. Este enfoque puede aplicarse a múltiples grupos de variables, ya sean cuantitativas, cualitativas o una combinación de ambas.

Para comprender este proceso, es importante señalar que en ambos análisis de reducción de dimensionalidad (ACP o ACM) se genera una matriz con las coordenadas de las variables y los individuos en cada CP. A cada uno de los $n$ CP se le asocia un valor propio $\lambda^n$, que es proporcional al porcentaje de varianza explicada por dicho componente (Figura 11.2 A y B). Aunque los detalles del procedimiento del AFM están descritos en el libro *Multiple Factor Analysis by Example Using R* [146], a continuación se expone una breve descripción de dicho procedimiento.

El AFM aplica un análisis de reducción de dimensionalidad individual (ACP o ACM, según la naturaleza de los datos) a cada grupo de variables. Posteriormente, utiliza las tablas de datos resultantes y los valores propios de los CP de cada grupo para integrarlos en un análisis conjunto. En primer lugar, se calcula un peso por cada grupo de variables. Este peso equivale a $1/\lambda_i^1$, donde $\lambda_i^1$ es el valor propio más grande (valor propio del CP1) del grupo de variables $i$. En segundo lugar, se construye una tabla completa que integra la tabla original para los grupos cuantitativos y, en el caso de los grupos cualitativos, utiliza una tabla de frecuencias que representa la pertenencia de las observaciones a cada categoría. Los grupos de variables se ponderan con los pesos mencionados para equilibrar la influencia de cada grupo. Por último, a la tabla ponderada se le aplica un ACP final para obtener los CP integrados y las coordenadas de individuos y variables en este nuevo espacio [146]. Esto facilita la identificación de patrones y relaciones tanto entre individuos como entre grupos de variables en un espacio común de dimensiones reducidas. Al igual que ocurre con el ACP y el ACM, se pueden proyectar grupos de variables suplementarios sobre los CP y aplicar el AJCP para agrupar a los individuos. Los

[2]El AFM puede trabajar con grupos de variables de naturaleza mixta pero este tipo de datos no se ha analizado en el marco de esta tesis doctoral.

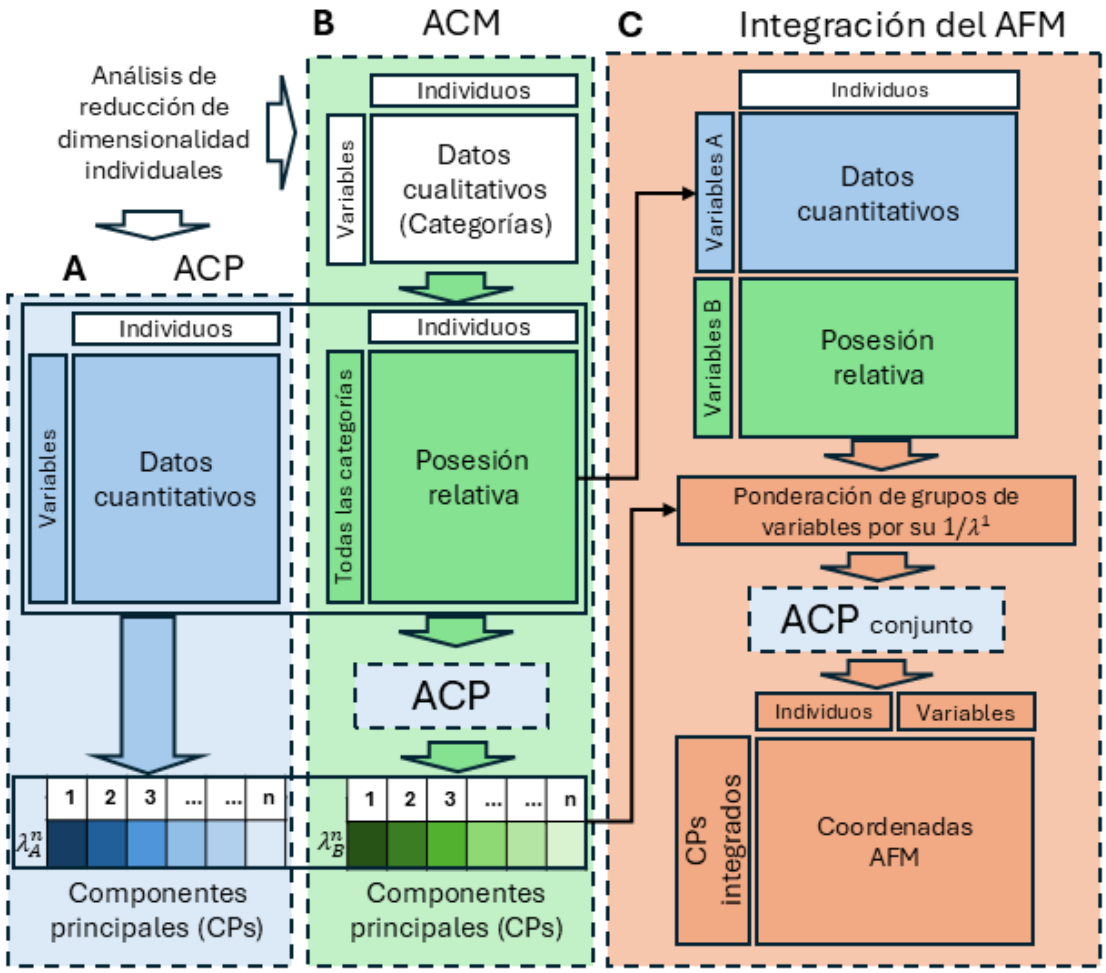


**Figura 11.2:** Diagrama resumen de los pasos que realiza el Análisis de Factores Múltiples (AFM). Este diagrama muestra el caso de dos grupos de variables siendo el grupo **A** cuantitativo y el grupo **B** cualitativo, ejemplificando los datos que se analizan en este estudio. Este análisis integra grupos de variables que pueden ser cualitativas o cuantitativas. Cuando el grupo de variables es cuantitativo se aplica un Análisis de Componentes Principales (**A - ACP**) y cuando es cualitativo se aplica un Análisis de Correspondencias Múltiples (**B - ACM**). En **B**, los datos cualitativos se transforman a una tabla de frecuencias de posesión sobre la que se aplica un ACP. Ambos análisis devuelven las coordenadas de los individuos y las variables en un espacio de componentes principales (**CP**) con sus valores propios $\lambda_i^n$ asociados, siendo **n** el numero de CPs y i el nombre del grupo de variables. En **C**, se combinan las tablas de datos cuantitativos y la tabla de frecuencias de posesión para generar la tabla completa del AFM. Las variables de cada grupo se ponderan por el inverso del mayor valor propio de cada grupo $\lambda^1$. Posteriormente, los datos ponderados se usarán para generar los CP integrados mediante ACP.

grupos de variables cualitativas se proyectan de manera similar al ACM, calculando previamente las frecuencias de pertenencia de las categorías.

Sobre el espacio de dimensiones reducidas se pueden proyectar grupos de variables suplementarios que no afectan al cálculo de los CP integrados, permitiendo interpretar dichos CP Además, los CP de los análisis de reducción de dimensionalidad individuales pueden ser proyectados como variables cuantitativas. El cálculo de las asociaciones entre los factores suplementarios y los CP, así como el análisis estadístico asociado, siguen el mismo procedimiento descrito para el ACP en el Capítulo 9.

En el flujo de trabajo se ha incluido el programa *mfa_degs_phen.R*, desarrollado para implementar el AFM de *FactoMineR* [97], con el objetivo de proyectar grupos de variables suplementarias y los CP individuales sobre los CP globales, y agrupar los individuos mediante AJCP. Se ha aplicado este programa a los datos fenotípicos (grupo "*phenotypes*") y de expresión diferencial de los diez pacientes de PMM2-CDG (grupo "*genes*"). Además, se ha configurado para usar las variantes patogénicas de ambos alelos del gen *PMM2* (grupo "*variants*"), la clasificación de grupos de gravedad ("LOW" y "HIGH", grupo "*severity*") y las escalas de gravedad NPCRS, ICARS y MVRD (grupo "*severity_scales"*) como grupos de variables suplementarios (Tabla 11.1).

**Tabla 11.1:** Descripción de los grupos de variables utilizados en el Análisis de Factores Múltiples (AFM). En esta tabla se describen los distintos grupos de variables incluidas en el AFM y se indica el **nombre** de los grupos de variables, las descripciones de las variables, la naturaleza de los **datos** y el **tipo** de grupo. Un grupo **activo** se usa para construir el espacio de CP integrado y un grupo **suplementario** se proyecta sobre este espacio, posteriormente.

| Nombre | Descripción | Datos | Tipo |
|---|---|---|---|
| *phenotypes* | Fenotipos | Cualitativo | Activo |
| *genes* | Genes con expresión diferencial | Cuantitativo | |
| *variants* | Variantes patogénicas | Cualitativo | Suplementario |
| *severity* | Clasificación de grupos de gravedad | Cualitativo | |
| *severity_scales* | Escalas y medidas de gravedad | Cuantitativo | |

## 11.3. Resultados

Este estudio realiza un análisis integrado de los datos fenotípicos y transcriptómicos de pacientes diagnosticados con la enfermedad PMM2-CDG. De la cohorte inicial de 27 pacientes analizada en el Capítulo 4 se secuenció el ARN de diez pacientes con distinta severidad.

Primero se estudió la relación entre las variantes del gen *PMM2* presentes en los pacientes estudiados. La proyección de las variantes patogénicas de ambos alelos de cada paciente sobre los CP de todos los genes expresados y los GED (calculados en el Capítulo 9) reveló que la mayoría de pacientes posee variantes propias. Esto se observa en ambos espacios de CP donde las coordenadas del paciente y las de las variantes proyectadas son las mismas en la mayoría de los casos (Figura 11.3). De las variantes compartidas entre varios pacientes, se puede observar que a2_p.Leu32Arg esta localizada cerca de los individuos con alta gravedad ("HIGH", Figura 11.3B). Sin embargo no se ha detectado ninguna relación significativa entre las variantes y los CP de los GED.

Los fenotipos de estos diez pacientes se analizaron con el programa *Cohort Analyzer*, con el que se determinó un único agrupamiento que engloba a todos los pacientes por su similitud semántica (Figura 11.4A). Sin embargo, al estudiar los pacientes según la similitud semántica de cada uno de sus fenotipos con el perfil fenotípico grupal, se encontraron diferencias entre los pacientes con distinta gravedad de la enfermedad (Figura 11.4B). En base a estos resultados, se realizó un análisis de reducción de dimensionalidad (ACM en este caso) para centrar el estudio en los fenotipos con mayor varianza entre las muestras.

Se aplicó el ACM (Figura 11.2B) a los fenotipos de los pacientes para evaluar su capacidad de diferenciar a los pacientes según la gravedad de la enfermedad en un espacio de dimensiones reducidas. Se observó que el 60,88 % de la varianza de los fenotipos se resume en tres CP relevantes (Figura 11.5). El primer CP es capaz de separar los pacientes con distinta gravedad (Figura 11.6A) y correlaciona de forma significativa con las escalas de gravedad (Figura 11.6B). En concreto, el primer CP correlaciona con NCPRS, ICARS y MVRD con valores -0,9, -0,89 y 0,68, respectivamente. La proyección de las variantes patogénicas sobre los CP de los fenotipos revela que hay variantes que se separan en el CP1 y, por consiguiente, se asocian a la gravedad de los pacientes (Figura 11.6A). Sin embargo, ninguna asociación es significativa. Por otro lado, existen fenotipos asociados de manera significativa a los tres CP relevantes. Sin embargo, son el primer y tercer CP los que se asocian a fenotipos con un valor de $R^2$ mayor que 0,8 (Tabla 11.2). En base a los CP relevantes, los individuos se dividen en cuatro grupos y, aunque las muestras P8_H_2 y P2_L_2 constituyen un grupo en sí mismos, el dendrograma diferencia a los individuos por su gravedad (Figura 11.7).

Se aplicó el AFM (Figura 11.2) sobre los GED y fenotipos como grupos de variables distintos (Tabla 11.1). Además, se incluyeron las variantes patogénicas en ambos alelos del gen *PMM2* de los pacientes, la clasificación de gravedad y sus escalas como grupos de variables suplementarios (Tabla 11.1).

El 52,9 % de la varianza se resumió en dos CP integrados relevantes. Se observó que la dirección de los CP1 individuales de los genes y los fenotipos y del

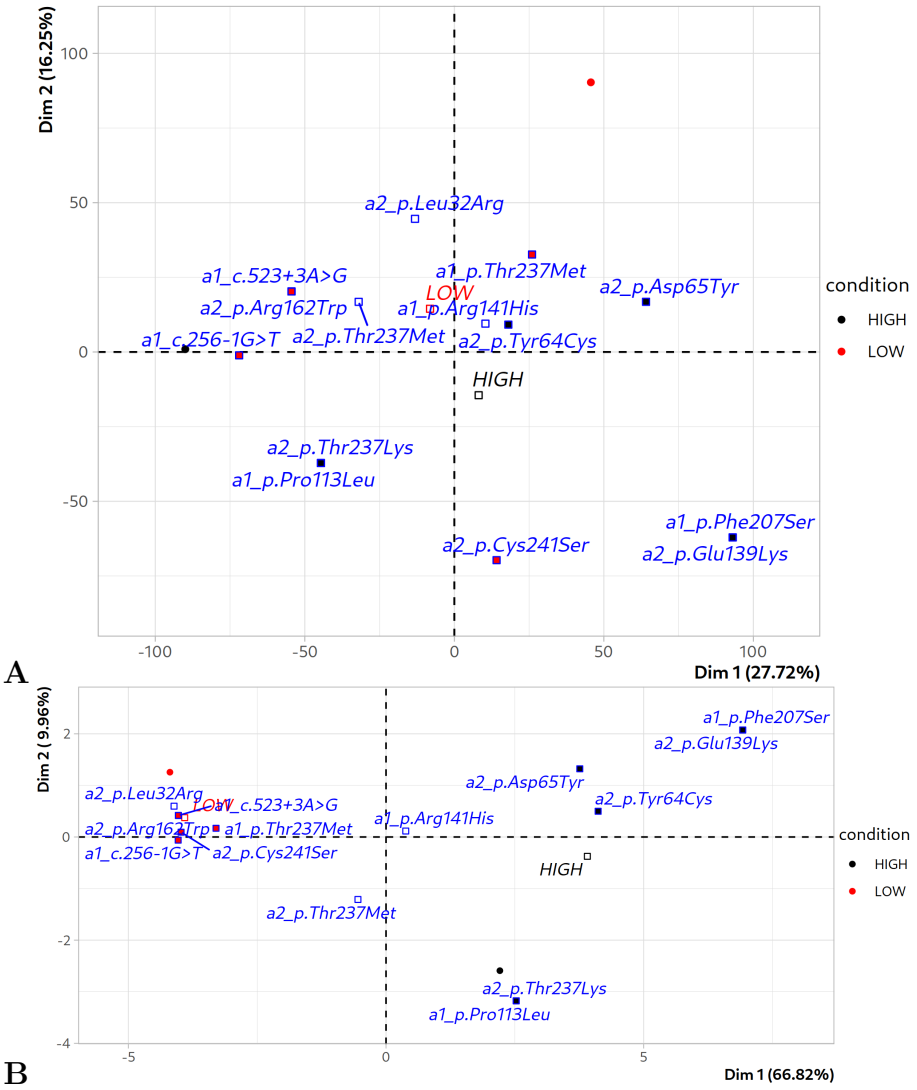


**Figura 11.3:** Análisis de componentes principales de la expresión génica de los pacientes con PMM2-CDG. En este gráfico se representan los dos primeros componentes (Dim 1 y Dim 2) del ACP aplicado a la expresión de todos los genes (**A**) y los GED (**B**). En ambos gráficos, los puntos negros y rojos representan a los pacientes con alta gravedad (**HIGH**) y baja gravedad (**LOW**), y los cuadrados y etiquetas del mismo color muestran las coordenadas medias de los individuos de cada grupo. Los cuadros y texto en azul representan las variantes patogénicas de ADN (**c.**) o proteína **p.** del gen *PMM2* que presentan los pacientes en ambos alelos (**a1** y **a2**). Estas variantes están descritas por el nucleótido/aminoácido que presentaría un individuo sano, la posición y el nucleótido/aminoácido resultante del cambio.

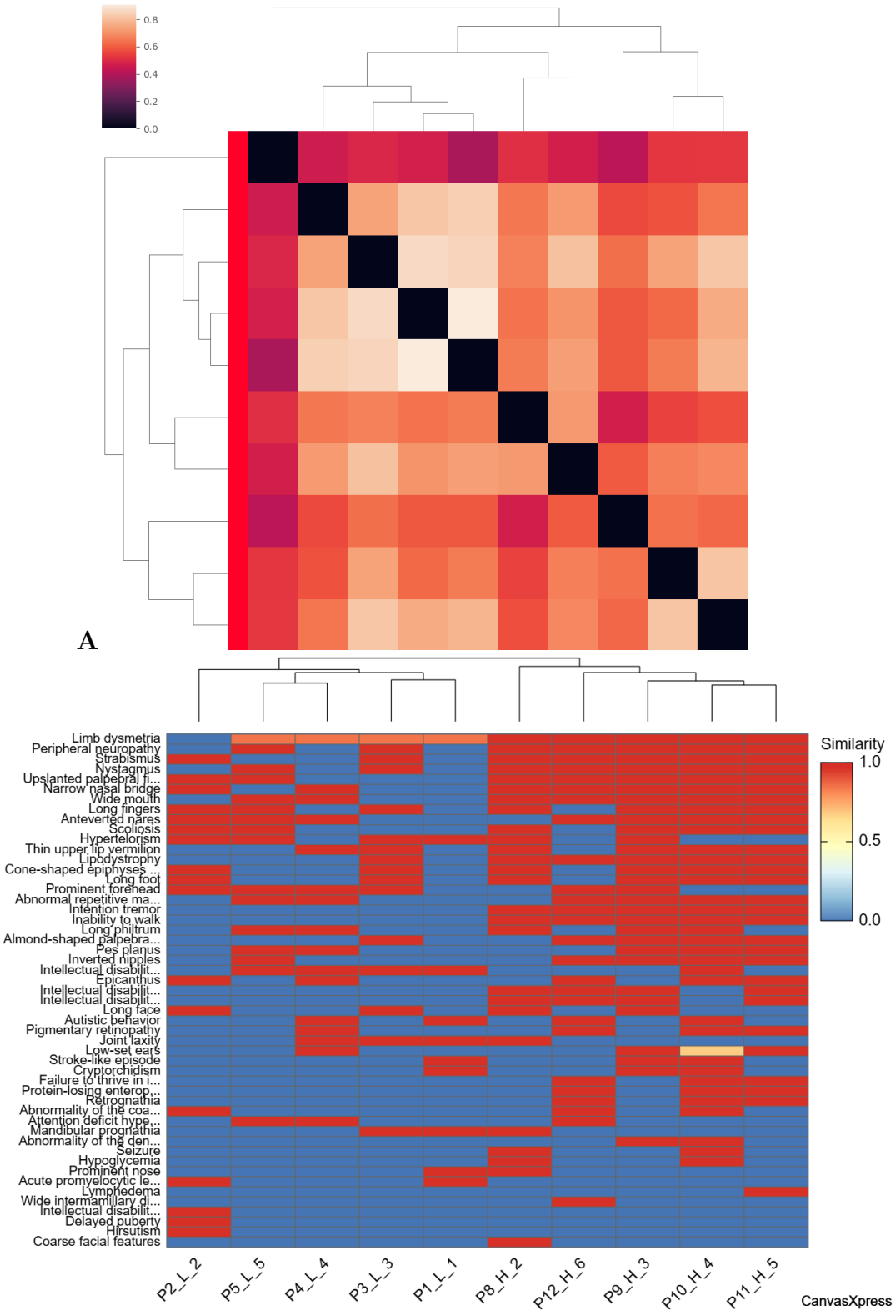


**Figura 11.4:** Similitud semántica entre los pacientes y los fenotipos. En **A** se representa la similitud semántica (color) y distancia (dendrograma) entre los pacientes distribuidos en cada fila y columna. Además, en el eje Y se muestran en colores las agrupaciones detectadas por *Cohort Analyzer*, que en este caso solo se encontró una (línea roja). En **B** se representa la similitud semántica entre los fenotipos de cada paciente del grupo rojo identificado en A (eje X) y el espectro fenotípico unificado de ese mismo grupo. El color muestra el valor de similitud semántica (Similarity) y el dendrograma muestra la distancia euclídea entre muestras según los valores de similitud. La alta (H) o baja (L) gravedad de la enfermedad está indicada en las etiquetas de los pacientes.

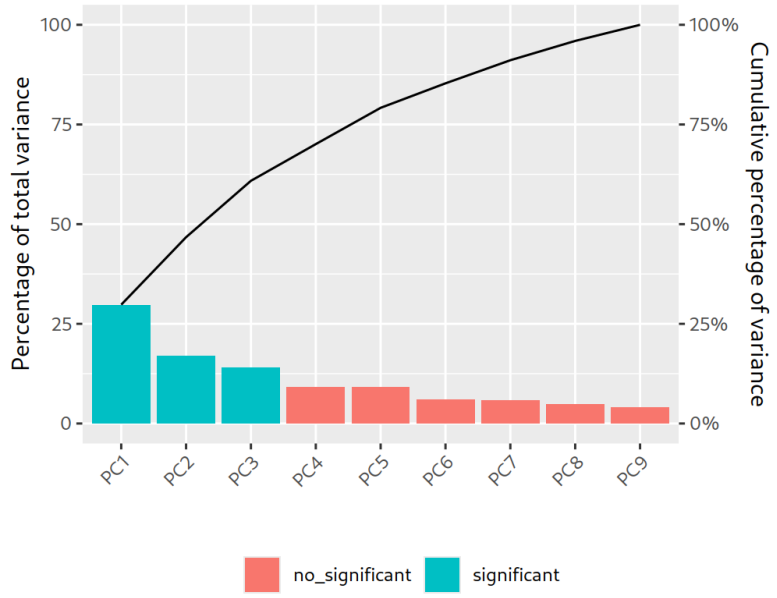


**Figura 11.5:** Valores propios o varianza de los componentes principales (CP). Se representan como barras los porcentajes de varianza resumida por los CP (**PC**), distinguiendo los relevantes (azul) de los no relevantes (rojo) y la varianza acumulada como una línea negra.

CP1 integrado del AFM están correlacionadas con R de Pearson mayor que $|0,95|$ (Figura 11.8). El CP1 de los genes correlaciona de forma positiva y el CP1 de los fenotipos lo hizo forma negativa (Figura 11.8). Además, el CP2 de los fenotipos correlaciona de manera negativa con el CP2 integrado. Ambos grupos de variables suplementarios que muestran la gravedad correlacionan de forma positiva con el CP1 integrado. Los CP individuales de las variantes, así como los CP2 de los genes y escalas de gravedad correlacionan con los CP integrados con un valor R de Pearson menor que $|0,6|$ (Figura 11.8).

En cuanto a la disposición de los individuos en el espacio de los CP integrados, se observó que los GED agrupan las muestras de manera más homogénea en el CP2, mientras que los fenotipos, por el contrario, tienden a separarlas. Sin embargo, las muestras P8_H_2 y P4_L_4 se separan más en este eje según las coordenadas integradas (media fenotipo-GED) (Figura 11.9).

Al estudiar la relación de las variables con los CP integrados se reveló que los 24 GED y las tres escalas de gravedad correlacionan de forma significativa con el CP1 integrado (Figura 11.10). En cuanto a las variables cualitativas, se observó que la condición de gravedad y 17 fenotipos están significativamente asociados al CP1 integrado, mientras que nueve fenotipos están significativamente asociados al

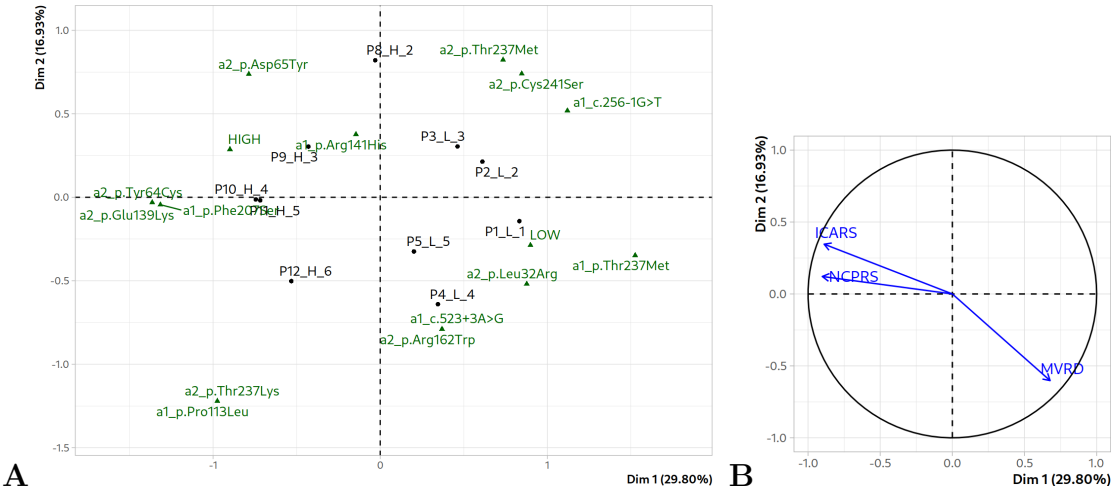


**Figura 11.6:** Análisis de Correspondencia Múltiple sobre los fenotipos de los pacientes y los factores suplementarios. En **A** se muestran los dos primeros componentes principales (**Dim 1** y **Dim 2**). Los pacientes están representados con un punto negro. Los triángulos y texto en verde representan el grupo de gravedad de los pacientes y las variantes patogénicas de la secuencia de ADN (**c.**) o proteína (**p.**) del gen *PMM2* que presentan los pacientes en ambos alelos (a1 y a2). Estas variantes están descritas por el nucleótido/aminoácido que presentaría un individuo sano, la posición y el nucleótido/aminoácido resultante del cambio. La alta (H) o baja (L) gravedad de la enfermedad está indicada en las etiquetas de los pacientes. En **B** se representan las correlaciones significativas (P ¡0,05) entre las escalas de gravedad (NCPRS, MVRD y ICARS) y los CP1 (**Dim 1**, eje horizontal) y CP2 (**Dim 2**, eje vertical) de los fenotipos.

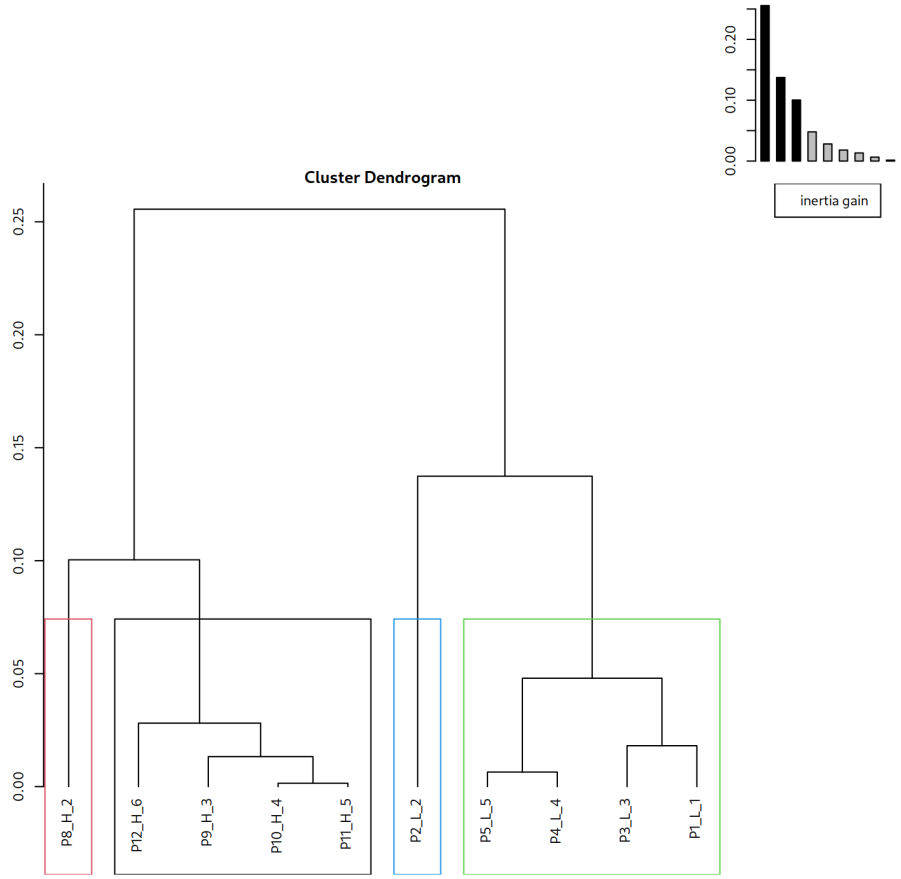


**Figura 11.7:** Agrupamiento jerárquico de los individuos basado en componentes principales (AJCP) del análisis de correspondencia múltiple aplicado a los fenotipos de los pacientes de PMM2-CDG. Se muestran el dendrograma y los grupos de individuos (cuadros de colores) según el AJCP y la contribución a la inercia de los distintos CP en el diagrama de barras distinguiendo los CP relevantes con color negro. La alta (H) o baja (L) gravedad de la enfermedad está indicada en las etiquetas de los pacientes

**Tabla 11.2:** Los fenotipos asociados con los componentes principales (CP) en el Análisis de Correspondencia Múltiple (ACM). En esta tabla se muestran los cinco fenotipos con sus términos en HPO que se asocian de forma significativa con los **CP** de ACM con mayor valor. El valor de asociación es el ajuste de $R^2$ asociado al valor P de un análisis de varianza. La significancia está indicada con etiquetas siendo *: 0,01 < P ≤ 0,05, **: 0,001 < P ≤ 0,01 y ***: P ≤ 0,001

| **Término HPO** | **CP** | **$R^2$** |
|---|---|---|
| Intention tremor | 1 | 0.808*** |
| Limb dysmetria | | 0,808*** |
| Inability to walk | | 0,808*** |
| Behavioral abnormality | | 0,662** |
| Inverted nipples | | 0,662** |
| Long face | 2 | 0.663** |
| Large hands | | 0,638** |
| Long foot | | 0,638** |
| Attention deficit hyperactivity disorder | | 0,605** |
| Long fingers | | 0,465* |
| Ataxia | 3 | 0.806*** |
| Intellectual disability, borderline | | 0,806*** |
| Delayed puberty | | 0,806*** |
| Hirsutism | | 0,806*** |
| Abnormality of the coagulation cascade | | 0,419* |

CP2 integrado (Tabla 11.3). En cuanto a las variantes patogénicas de los pacientes, no se encontró una relación significativa con los CP integrados.

Las muestras se dividen en cuatro grupos según el AJCP aplicado a los CP integrados y se observó una diferencia entre los pacientes con distinta gravedad según el dendrograma (Figure 11.11).

### 11.3.1. Discusión

El análisis de la expresión génica de los pacientes con PMM2-CDG secuenciados, realizado mediante ACP y descrito en el Capítulo 9, reveló que solo el CP3 del conjunto completo de genes permitía separar a los individuos según la gravedad de la enfermedad. En ese mismo capítulo, se observó que el CP1 de los genes con expresión diferencial era capaz de separar a los pacientes según la gravedad de la

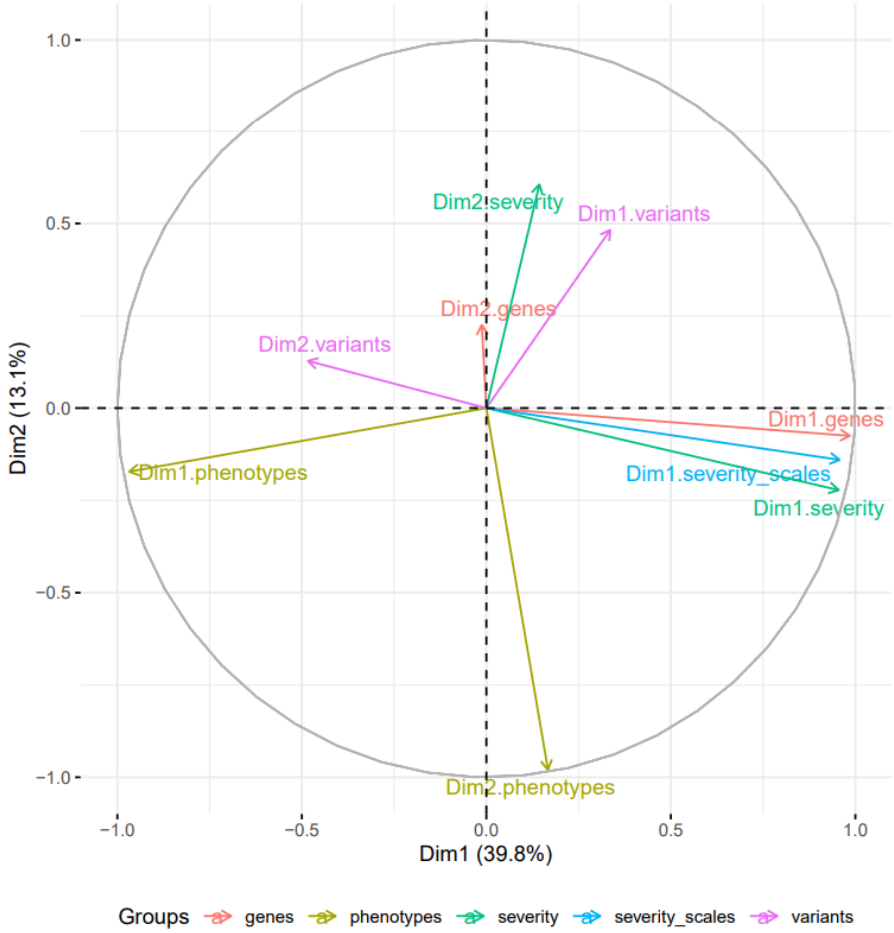


**Figura 11.8:** Correlación entre los componentes principales (CP) individuales de los grupos de variables de estudio y los CP integrados de AFM. En este gráfico se muestra la correlación de los distintos CP individuales de los grupos de variables con los CP1 (Dim1) y CP2 (Dim2) integrados, eje X e Y respectivamente. En cada eje se indica el porcentaje de varianza total resumida por cada CP. En cada color se muestra un grupo de variables siendo los GED (**genes**) y los fenotipos (**phenotypes**) las variables usadas para calcular los CP y las restantes, suplementarias.

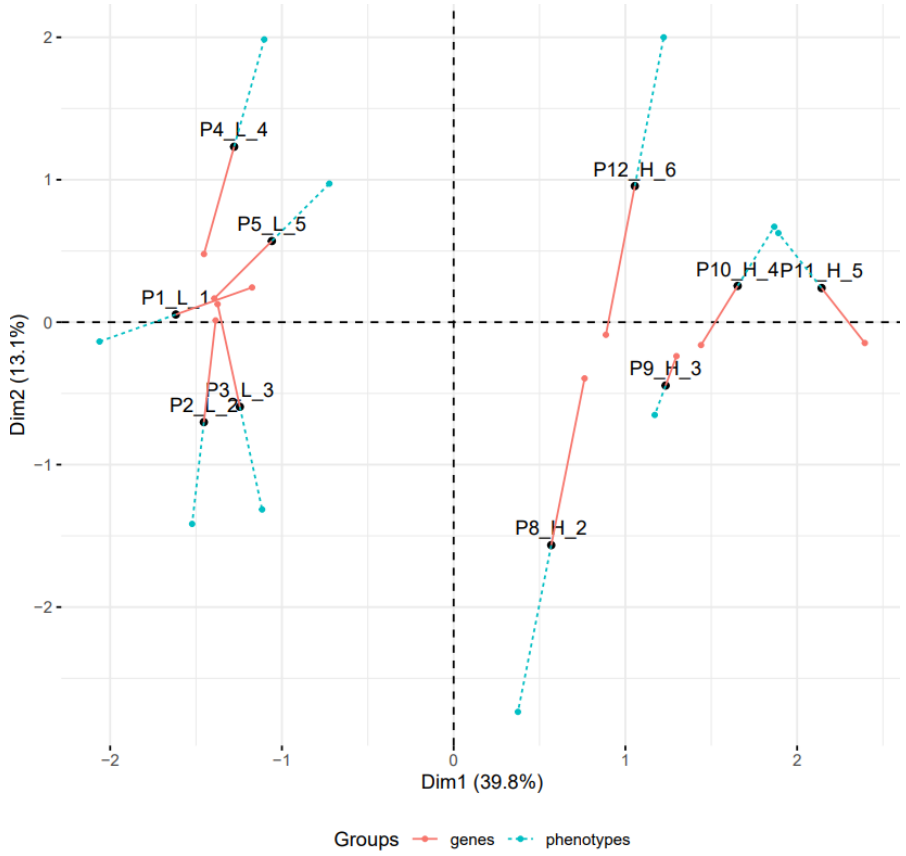


**Figura 11.9:** Distribución de los pacientes en el espacio de AFM. En este gráfico se representan la disposición de las muestras a lo largo de los dos CP integrados relevantes (Dim1 y Dim2) del análisis conjunto de fenotipos y genes. El porcentaje de la varianza total resumida por cada CP está indicado entre paréntesis. Las coordenadas en los CP integrados de los individuos están representadas por un punto negro. Los puntos de colores representan las coordenadas de los individuos según cada grupo de variables y las líneas representan la distancia de las coordenadas de los individuos en los CP integrados y las coordenadas según cada grupo de variables. La alta (H) o baja (L) gravedad de la enfermedad está indicada en las etiquetas de los pacientes.

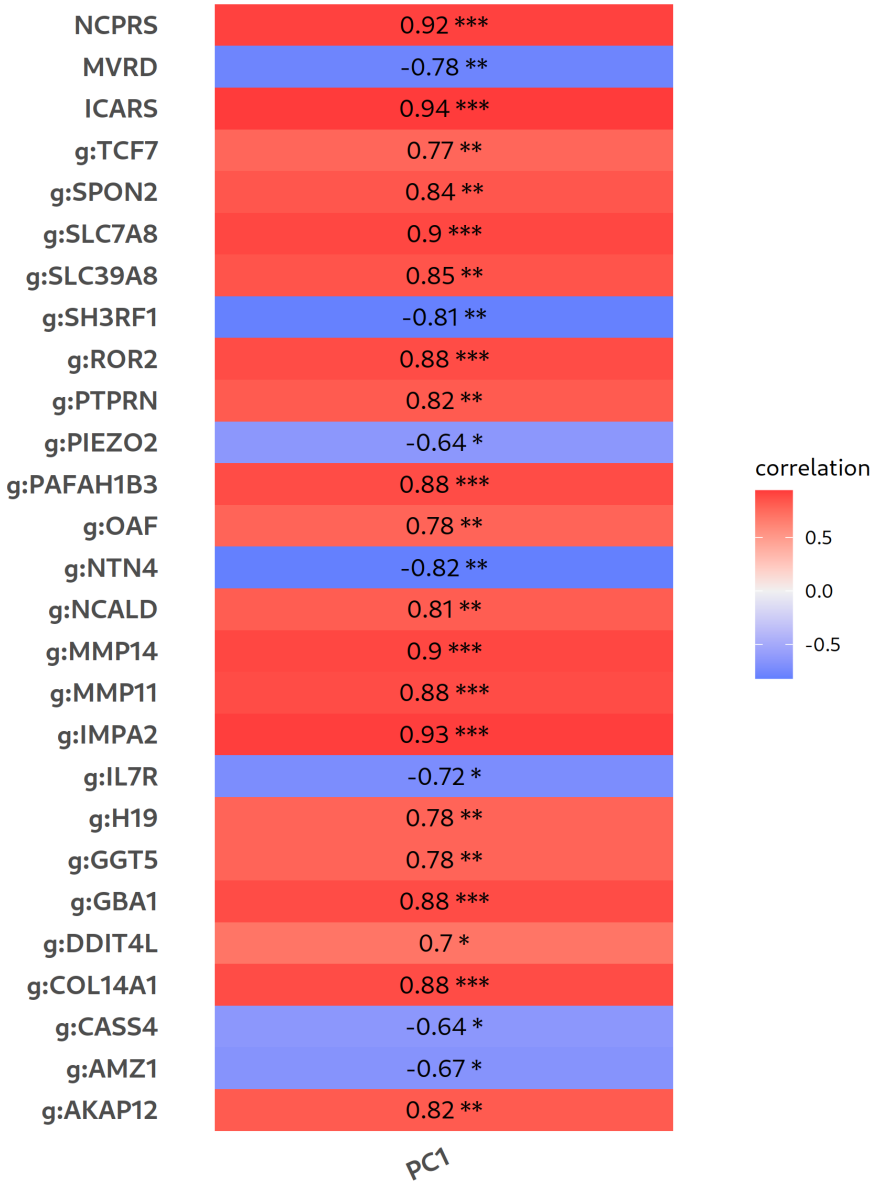


**Figura 11.10:** Relaciones significativas de las variables cuantitativas en el espacio de AFM. En esta figura se representan las relaciones significativas entre las variables cuantitativas y los CP integrados. Las variables que se muestran en la figura son los nombres los genes precedidos de la etiqueta **g:** y las escalas de gravedad. El valor de asociación es la correlación entre las variables cuantitativas y los CP integrados. La significancia está indicada con etiquetas siendo *: 0,01 $<$ P $\leq$ 0,05, **: 0,001 $<$ P $\leq$ 0,01 y ***: P $\leq$ 0,001

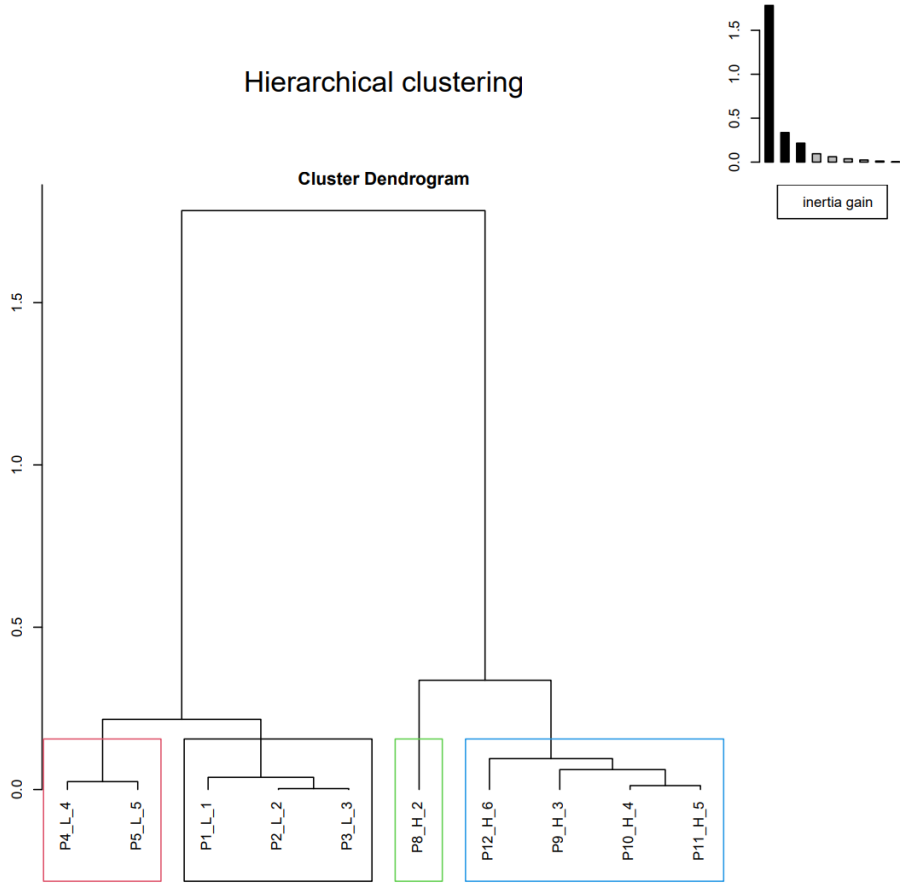


**Figura 11.11:** Agrupamiento jerárquico (AJCP) de los individuos basado en los componentes principales integrados del AFM aplicado a los fenotipos y genes con expresión diferencial de los pacientes de PMM2-CDG. Se muestran el dendrograma y los grupos de individuos (cuadros de colores) según el AJCP y la contribución a la inercia de los distintos CP en el diagrama de barras distinguiendo los CP relevantes con color negro. La alta (H) o baja (L) gravedad de la enfermedad está indicada en las etiquetas de los pacientes.

enfermedad y mostraba una correlación significativa con las escalas de gravedad NCPRS, ICARS y MVRD. Además, se observó que el AJCP de los CP relevantes de los GED agrupaba a los individuos según la gravedad de la patología, identificando incluso que las muestras P8_H_2 y P12_H_6 constituían un subgrupo de pacientes con alta gravedad sin relación aparente con ningún factor experimental. Esto condujo a profundizar en el estudio de los GED y su relación con los fenotipos.

El análisis de los fenotipos de estos mismos pacientes, basado en similitud semántica, reveló algunas diferencias entre los pacientes con distintos niveles de gravedad (Figura 11.4). Sin embargo, este enfoque no logró distinguir claramente entre ambos grupos ni permitir la inclusión de datos externos. Por esta razón, se optó por aplicar un análisis de reducción de dimensionalidad, como el ACM.

El análisis de los fenotipos de los pacientes mediante ACM confirmó la distinción entre los grupos de diferente gravedad, ya que estos se disponen en direcciones opuestas a lo largo del CP1 (Figura 11.6). Además, el CP1 muestra una correlación significativa con las distintas escalas de gravedad. Esta separación también es evidente en el CP2. Sin embargo, ningún factor adicional a los fenotipos permite interpretar la causa subyacente.

La aplicación del AJCP, que agrupa todas las muestras según sus coordenadas en los tres CP relevantes, dio lugar a cuatro grupos que permiten diferenciar claramente los individuos según la gravedad de su enfermedad. En conjunto, los resultados demuestran que el ACM ofrece una mayor resolución para definir grupos de individuos en comparación con el método de similitud semántica. Esto puede deberse a que el ACM, al calcular las posesiones relativas de las categorías (Figura 11.2B), da prioridad a las categorías más específicas [146], que en este caso corresponden con los fenotipos presentes en menos pacientes. En el contexto de una cohorte de pacientes con la misma enfermedad, donde muchos comparten la mayoría de los fenotipos, aquellos pacientes con fenotipos más específicos tienden a diferenciarse con mayor claridad en los CP.

Dado que los individuos con distinta gravedad pueden identificarse tanto por sus fenotipos como por los genes con expresión diferencial en un espacio de dimensiones reducidas, usando ACM y ACP respectivamente, se aplicó el AFM para estudiar a los pacientes, combinando ambos tipos de datos en un espacio de CP integrados. Se observó que los pacientes con distinta gravedad de la enfermedad se disponen a lados opuestos del CP1 integrado (Figura 11.9) y que los CP1 individuales —resultado de aplicar un análisis de reducción de la dimensionalidad a cada grupo por separado— de los fenotipos, genes, escalas de gravedad y la clasificación por gravedad, correlacionan con el CP1 integrado (Figura 11.8). De manera adicional, la asociación de las escalas de gravedad (Figura 11.10) y la clasificación por gravedad (Tabla 11.3) con el CP1 integrado es significativa. Estas pruebas en

conjunto demuestran que el CP1 integrado está íntimamente asociado a la gravedad de la enfermedad.

En cuanto a la contribución parcial de los genes y los fenotipos en las coordenadas de las muestras en el espacio integrado, es importante destacar que los genes contribuyen a mitigar el impacto del CP2 en la disposición de las muestras. Esto tiene un efecto relevante en la diferenciación de los grupos mediante los CP integrados al aplicar AJCP. Aunque se establecen cuatro grupos distintos de pacientes, este enfoque es capaz de diferenciar de forma más precisa los pacientes según su gravedad. Esto se deduce al comparar la inercia de los dendrogramas del AJCP aplicado a los CP integrados (Figura 11.11) y a los fenotipos (Figura 11.7) donde se observa que la inercia que separa los individuos por gravedad es mucho mayor en los CP integrados.

El CP1 integrado se asocia de manera significativa a los 24 GED y 17 de los fenotipos, de los cuales, los genes *SLC7A8*, *MMP14* y *IMPA2* y los fenotipos "*Inability to walk*" (HP:0002540), "*Limb dysmetria*" (HP:0002406) y "*Intention tremor*" (HP:0002080) presentaban una asociación más alta (Figura 11.10 y Tabla 11.3). Es interesante destacar que tanto el ACM como el AFM son capaces de asociar estos fenotipos a los respectivos CP1, y por tanto a la gravedad de la enfermedad (Tabla 11.2). Sin embargo, el ajuste $R^2$ de estos fenotipos con el CP1 integrado es superior, lo que sugiere que la integración de genes y fenotipos mediante el AFM constituye un método más eficaz para asociar fenotipos a la gravedad de la enfermedad. Aunque se observa una cierta separación de las variantes patogénicas en el espacio integrado, no se ha encontrado una asociación estadística significativa con ningún CP. Esto podría deberse a que las variantes de los pacientes analizados son altamente específicas. Para incluir las variantes patogénicas en un estudio futuro similar se debería escoger una cohorte con más pacientes con variantes menos heterogéneas.

Un dato interesante de destacar es que el AJCP, aplicado tanto a los CP de los genes y de los fenotipos por separado, ha sido capaz de identificar subgrupos de pacientes. Ambos agrupamientos, aunque parten de unos datos de distinta naturaleza, tienen como patrón común la divergencia entre el paciente P8_H_2 y los demás clasificados con alta gravedad. La coherencia entre los fenotipos y los genes con expresión diferencial hacen interesante explorar, en un futuro, el trasfondo genético de este paciente para explicar las diferencias de expresión génica y de sus fenotipos.

La capacidad del CP1 integrado para diferenciar a los pacientes según su gravedad, junto con el impacto reducido del CP2 integrado, demuestra que la integración de la expresión génica y los fenotipos de los pacientes en un espacio de dimensiones reducidas ofrece ventajas significativas para distinguir la gravedad de la enfermedad. Esto permite asociar con mayor precisión genes y fenotipos específicos a la

severidad de los pacientes.

**Tabla 11.3:** Relaciones significativas de las variables cualitativas en el espacio de AFM. En esta figura se representan las relaciones significativas entre las variables cualitativas y los CP integrados 1 y 2. En la primera sección de la tabla se muestran los fenotipos por sus términos *HPO* y en la segunda parte se muestra la variable "Condition" que describe ambas categorías de gravedad de los individuos "LOW" y "HIGH". El valor de asociación es el ajuste de $R^2$ asociado al valor P de un análisis de varianza. La significancia está indicada con etiquetas siendo *: $0{,}01 < P \leq 0{,}05$, **: $0{,}001 < P \leq 0{,}01$ y ***: $P \leq 0{,}001$

| **Término HPO** | **CP** | **$R^2$** |
|---|---|---|
| Intention tremor | 1 | $0{,}917^{***}$ |
| Limb dysmetria | 1 | $0{,}917^{***}$ |
| Inability to walk | 1 | $0{,}917^{***}$ |
| Lipodystrophy | 1 | $0{,}631^{**}$ |
| Failure to thrive in infancy | 1 | $0{,}58^{*}$ |
| Protein-losing enteropathy | 1 | $0{,}58^{*}$ |
| Retrognathia | 1 | $0{,}58^{*}$ |
| Intellectual disability, moderate | 1 | $0{,}539^{*}$ |
| Intellectual disability, severe | 1 | $0{,}539^{*}$ |
| Behavioral abnormality | 1 | $0{,}523^{*}$ |
| Inverted nipples | 1 | $0{,}523^{*}$ |
| Almond-shaped palpebral fissure | 1 | $0{,}485^{*}$ |
| Peripheral neuropathy | 1 | $0{,}466^{*}$ |
| Nystagmus | 1 | $0{,}466^{*}$ |
| Wide mouth | 1 | $0{,}459^{*}$ |
| Upslanted palpebral fissure | 1 | $0{,}423^{*}$ |
| Long face | 2 | $0{,}714^{**}$ |
| Attention deficit hyperactivity disorder | 2 | $0.568^{*}$ |
| Large hands | 2 | $0{,}517^{*}$ |
| Long foot | 2 | $0{,}517^{*}$ |
| Motor stereotypy | 2 | $0{,}515^{*}$ |
| Hypertelorism | 2 | $0{,}47^{*}$ |
| Pigmentary retinopathy | 2 | $0{,}47^{*}$ |
| Coarse facial features | 2 | $0{,}427^{*}$ |
| Autistic behavior | 2 | $0{,}407^{*}$ |
| Condition | 1 | $0{,}917^{***}$ |

# Parte V

# Discusión y conclusiones

# Capítulo 12

# Discusión

En el marco de esta tesis doctoral, se han desarrollado metodologías innovadoras para el análisis riguroso de datos fenotípicos y de expresión génica derivados de estudios de RNA-seq en enfermedades raras.

Tal y como se describió al inicio de esta memoria, en la Sección 2.2, se necesita escoger una metodología acorde a la cohorte de estudio [9] cuya principal distinción, en el caso de enfermedades raras, es su reducido tamaño. Por un lado, el análisis de cohortes de pacientes y conjuntos de enfermedades a nivel fenotípico requiere de un estudio de la calidad del fenotipado. Este análisis es fundamental para realizar estudios posteriores, ya que la caracterización completa, profunda y rigurosa de los perfiles de fenotipos es un requisito indispensable para estudiar con precisión las relaciones entre grupos de pacientes o de enfermedades. Es por ello que se ha participado en el desarrollo del programa *Cohort Analyzer* y se ha aplicado tanto al análisis de cohortes de pacientes (Capítulo 4) como al de un grupo de enfermedades raras relacionadas con el proceso de angiogénesis desregulada (Capítulo 5).

Este programa se empleó para evaluar la calidad del fenotipado de tres cohortes de pacientes diferentes (DECIPHER [79], ID/MCA [80] y pacientes diagnosticados con el síndrome PMM2-CDG). Los resultados demostraron que la cohorte PMM2-CDG, al tratarse un grupo específico de pacientes con un seguimiento clínico exhaustivo y diagnosticados con la misma enfermedad, tenía una mejor calidad del fenotipado que DECIPHER e ID/MCA (Capítulo 4, Figura 3). Este hecho benefició la identificación de subgrupos de pacientes para la cohorte PMM2-CDG. Por su parte, los resultados de *Cohort Analyzer* para el análisis de las cohortes DECIPHER e ID/MCA demostraron que ambas, sobretodo en el caso de ID/MCA, fueron pobremente caracterizadas a nivel del número de fenotipos por paciente y en profundidad de términos de la ontología *HPO* (Capítulo 4, Figura 3). Los resultados, una vez aplicados los filtros de selección por el número mínimo de fenotipos por paciente, en el caso de DECIPHER mejoraron ligeramente. Sin embargo, se pudo concluir que ambas cohortes (DECIPHER e ID/MCA) necesitaban un fe-

notipado más exhaustivo para relacionar su información con datos genómicos que permitiesen establecer asociaciones fenotipo-genotipo significativas, realizar la correcta estratificación de los pacientes o cualquier otro tipo de análisis posterior con estos datos (Capítulo 4, Figura 5, 6 y Figura suplementaria S3).

En el caso de confirmar que los datos fenotípicos de una cohorte tienen la información y calidad suficiente como para hacer análisis subsecuentes con ellos, una estrategia interesante consiste en la estratificación de los pacientes o enfermedades con características fenotípicas similares en subgrupos. Este agrupamiento puede ser el punto de partida para análisis de modelos de red con el fin de identificar específicamente y en profundidad las características, tanto a nivel fenotípico como genómico, de cada subgrupo de pacientes o enfermedades. Con esta idea, se utilizó *Cohort Analyzer* para estudiar todas aquellas enfermedades raras dependientes de angiogénesis caracterizadas a partir de una exhaustiva revisión bibliográfica, con el fin de caracterizar los procesos moleculares afectados e identificar genes nuevos implicados en esta desregulación (Capítulo 5). Siguiendo el protocolo publicado previamente por Rodríguez-Caso y colaboradores [82], se obtuvieron los perfiles fenotípicos en términos de la *HPO* de aquellas enfermedades raras que habían sido descritas con el término angiogénesis, y se analizaron y agruparon por similitud semántica con *Cohort Analyzer*.

Por un lado, se estudió la cohesión de los grupos de enfermedades al calcular el camino más corto entre los pares de genes conocidos asociados a cada grupo de enfermedad en la red de interacciones de proteínas de STRING [84] (Capítulo 5, Figura 1). Se observó que los genes asociados a cada grupo eran cercanos en la red de proteínas (Capítulo 5, Figura 3). Asimismo, empleando el programa *clusters_to_enrichments.R*, se determinó que los grupos eran coherentes a nivel funcional porque cada uno de ellos estaba enriquecido en términos *GO* diferentes (Capítulo 5, Figuras 4 y 5). Por otro lado, al expandir los genes asociados a los grupos de enfermedad, incorporando los genes presentes en el cálculo del camino más corto, se observó que el factor de coagulación *F7* tenía una variante asociada al proceso de angiogénesis que no había sido previamente asociado a ninguna de las enfermedades raras analizadas (Capítulo 5, Tabla 2). Por todo ello, se demuestra que este análisis se puede usar para predecir nuevos genes relacionados con un proceso biológico gracias al agrupamiento de las enfermedades conocidas.

En relación con el estudio de la expresión génica y el análisis de las funciones moleculares asociadas a enfermedades raras, se emplearon datos de RNA-seq proporcionados por grupos colaboradores del equipo de investigación donde se realizó esta tesis doctoral. En el Capítulo 6 se presentan los análisis de expresión realizados en un experimento con muestras de un modelo de estudio de la enfermedad de Lafora, así como un experimento piloto con cultivos de fibroblastos derivados de pacientes de la cohorte de PMM2-CDG, ambas clasificadas co-

mo enfermedades raras asociadas a desórdenes del metabolismo de los glúcidos [147, 148, 149]. Dado que los estudios de expresión génica realizados en esta tesis se caracterizan por un diseño experimental con un número limitado de muestras, se optó por integrar diversos algoritmos y estrategias analíticas con el objetivo de mejorar la precisión de los resultados. Este enfoque dio lugar al desarrollo del paquete *ExpHunterSuite*[81] que incluye los programas *degenes_hunter.R*, *functional_hunter.R*, *clusters_to_enrichments.R* y *multivar_mine.R*. En el caso de los análisis de expresión diferencial, como se detalla en el Capítulo 6, se optimizó el programa *degenes_hunter.R* [150] que aplica de forma conjunta los algoritmos de los paquetes *limma* [46], *DESeq2* [52], *edgeR* [53] y *NOISeq* [54] sobre los mismos conjuntos de datos. La aplicación de este programa a datos de RNA-seq de enfermedades raras mostró que los GEDs identificados mediante múltiples algoritmos eran menos propensos a generar falsos positivos (Capítulo 6, Figuras 3 y 4). Esto permitió reducir la arbitrariedad al escoger un único algoritmo de expresión diferencial y filtrar la cantidad de GEDs para una futura validación experimental o estudio funcional. Además de los análisis de expresión diferencial, se incluyó el análisis de redes de correlación de *WGCNA* [55] en *degenes_hunter.R*, que permite agrupar genes con un patrón de expresión similar en módulos de coexpresión. El calculo de expresión diferencial, de módulos de coexpresión y su posterior enriquecimiento funcional, realizado con el programa *functional_hunter.R*, ha permitido destacar funciones relevantes para el desarrollo de las enfermedades raras analizadas. Se ha confirmado la implicación de la comunicación entre la microglía y los astrocitos en la neurodegeneración que tiene lugar en la enfermedad de Lafora, descrita en estudios previos [151], y se localizaron módulos de genes relacionados con el proceso de inflamación para estudios posteriores (Capítulo 6, Figura 7). Asimismo, se identificaron procesos moleculares relacionados con la matriz extracelular y el colágeno, los cuales están asociados con las hemorragias intracerebrales y los episodios tipo ictus característicos de la enfermedad PMM2-CDG (Capítulo 6, Figura 6) [152, 153].

El programa *degenes_hunter.R* se utilizó también para el análisis de datos de expresión de miARN, llevando a cabo estudios de correlación con sus genes diana mediante la integración de datos de secuenciación tanto de miARN como de genes. Se emplearon tres conjuntos de datos en este trabajo: el primero correspondió a un estudio realizado en un modelo de la enfermedad de Lafora (el mismo conjunto analizado en el Capítulo 6); el segundo, a un experimento con cultivos de fibroblastos de pacientes pertenecientes a la cohorte de PMM2-CDG con distintos niveles de gravedad; y el tercero, a un experimento basado en un modelo de la enfermedad LMNA-DCM. En el caso de la enfermedad de Lafora, se confirmó el cambio de expresión de los miR-155 y miR-146a observado en estudios previos [41], y se propuso el miR-142a como candidato para su validación experimental. En los

casos de PMM2-CDG y LMNA-DCM, se obtuvieron 17 y 53 miARN expresados de forma diferencial, respectivamente. De ellos, no se encontró ninguna asociación previa con las enfermedades en las bases de datos empleadas (Capítulo 7, Tabla 1) a excepción del miR-155, el cual se ha descrito como regulador de procesos inflamatorios en la miocardiopatía dilatada [39]. De manera adicional, se calcularon los módulos de coexpresión de miARN mediante el análisis de redes de correlación y se encontraron 15 módulos para la enfermedad de Lafora, 55 para PMM2-CDG y siete para LMNA-DCM. Como se ha comentado previamente, de las tres enfermedades se disponía de datos de expresión génica obtenidos por secuenciación y también se analizó con *degenes_hunter.R*, los detalles pueden encontrarse en la Tabla 1 del Capítulo 7.

Tanto los datos de expresión diferencial como los módulos de coexpresión de genes y de miARN se han utilizado, además, para analizar las parejas de regulación miARN-gen diana de las enfermedades de Lafora, PMM2-CDG y LMNA-DCM. Para la detección de parejas miARN-gen diana se desarrolló el programa *coRmiT* (Capítulo 7), dentro del paquete *ExpHunterSuite*, que aplica todas las estrategias de correlación encontradas en la bibliografía que hacen uso de datos de expresión diferencial y módulos de coexpresión (Capítulo 7, Tabla suplementaria 1) [87, 154, 85, 86, 88]. Todas estas estrategias parten de la base que el mecanismo de regulación por miARN reprime la cantidad de ARN de sus genes diana, como se describió en la Sección 2.4, por lo que solo estudian las correlaciones negativas. Dado que se observó que las distintas estrategias encuentran parejas miARN-gen diana diferentes a partir de los mismos datos (Capítulo 7, Figura 2), se diseñó un protocolo de selección e integración de los resultados de todas las estrategias en base a su solapamiento con las bases de datos [90]. Se demostró que la estrategia óptima de correlación para cada miARN es constante, incluso ante cambios aleatorios en las bases de datos (Capítulo 7, Figura 4), evidenciando que el protocolo de selección e integración desarrollado es robusto, a pesar del incremento de información disponible en las bases de datos. Las dianas de los miARN detectadas por *coRmiT* fueron el objeto de un análisis de enriquecimiento funcional en *GO*, *KEGG* y *Reactome* para el que se usó el programa *clusters_to_enrichments.R*. En el caso de estudio de los pacientes de PMM2-CDG sólo se estudiaron las parejas miARN-gen diana implicadas en los pacientes con gravedad alta de la enfermedad, y se observó que el miR-let-7i afectaba a la expresión de genes de colágeno tipo IV y procesos moleculares asociados, lo que refuerza la importancia de las rutas implicadas en la regulación del colágeno y la matriz extracelular en la enfermedad descrita en el Capítulo 6. En el caso de estudio de LMNA-DCM, se confirmó el papel inflamatorio de miR-155 [39, 155, 156, 157, 158]. Además, se observó que el miR-196b está relacionado con procesos que mantienen la estructura de la matriz celular y la cohesión entre células, así como que el miR-135a es un posible

regulador de la vía de señalización de WNT y que el miR-182 afecta a la conducción del impulso cardíaco, tres procesos que están íntimamente relacionados con la enfermedad [39, 159, 43, 160]. Igualmente, en el Capítulo 8 se hizo un estudio en profundidad de los ratones $Lmna^{R249W}$ mediante las relaciones miARN-gen diana detectadas con *coRmiT* y que estaban previamente validadas por otros estudios. En este caso, se usó una versión de *coRmiT* anterior a la incorporación del método de selección e integración de estrategias, por lo que se seleccionaron las parejas encontradas por la estrategia con el mayor *Odds ratio* general. En dicho estudio, un grupo colaborador validó la expresión diferencial de los miR-133a, miR-139, miR-149, miR-155, miR-183, miR-196 y miR-324 mediante RTq-PCR. Gracias al enriquecimiento de los genes diana en términos *GO*, se confirmó el papel del miR-155 en la regulación de procesos inflamatorios y del sistema inmune [39, 155, 156, 157, 158], y se hipotetizó que el resto de los miARN validados tienen un papel relevante en procesos moleculares y fisiológicos fundamentales en LMNA-DCM. Gracias a estos resultados, se puede concluir que la cepa de ratones $Lmna^{R249W}$ es un buen modelo de estudio para la enfermedad LMNA-DCM.

En el caso de la enfermedad de Lafora se observó que las dianas de los miR-155 y miR-146a están relacionadas con la actividad de los receptores de glutamato. Sin embargo, se ha observado en otros estudios que ambos miARN están relacionados con procesos inflamatorios [155, 156, 157, 158] y que, además, los genes relacionados con la inflamación coexpresaban con estos miARN [151, 41]. Por consiguiente, se decidió emplear *coRmiT* para estudiar las parejas miARN-gen diana con correlaciones positivas. Se confirmó que el miR-155 correlacionaba de forma positiva con sus genes diana relacionados con procesos de inflamación, lo que contradice el mecanismo clásico de regulación de los miARN descrito [33] (Sección 2.4). Este fenómeno se ha descrito para el mismo miARN en enfermedades autoinmunes, como el lupus eritematoso [161], asociado a una infraexpresión de *DICER1*, que también ha sido observada en el caso de la enfermedad de Lafora (***Dicer1***: *log2FC* = -0,17; FDR = 5,33e-06)(Capítulo 7). Además, en la base de datos de parejas miARN-gen diana validadas por *TarBase*, se ha encontrado que *DICER1* es una diana de miR-155 [1]. Sin embargo, esta pareja no se ha detectado con coRmiT debido a los filtros de $|log_2FC|$ que se aplicaron. Esto sugiere que podría ser un mecanismo de regulación no canónico de miR-155 sobre el que sería interesante realizar estudios futuros. En conjunto, se puede afirmar que la aplicación de *coRmiT* ha permitido encontrar parejas miARN-gen diana que explican los mecanismos moleculares de las enfermedades raras analizadas.

De manera adicional, la detección de parejas miARN-gen diana con *coRmiT* se han usado de manera complementaria en un estudio de ARN competidores

[1] https://dianalab.e-ce.uth.gr/tarbasev9/interactions?gene=Dicer1%2C+DICER1&mirna=hsa-miR-155-5p

endógenos en pacientes de displasia arritmogénica (Capítulo 10). En dicho estudio, se aplicaron programas para identificar ARNcirc (*CirComPara2*, [134]) y sus uniones con los miARN (*circR*, [135]) que combinan, a su vez varios algoritmos. Para identificar las parejas ARNcirc-gen diana se integraron las parejas miARN-gen obtenidas con *coRmiT* y las parejas ARNcirc-miARN de *circR* ya que los ARNcirc y los genes compiten por su unión física a los miARN [44]. Los ARNcirc se priorizaron por la similitud semántica entre los fenotipos de sus genes diana con los fenotipos de la displasia arritmogénica y se seleccionaron los diez primeros (Capítulo 10, Tabla 10.1). Se observó que los siete ARNcirc con mayor similitud sólo tenían un gen (*ALG10B*) con fenotipos anotados en sus dianas. Este gen funciona como regulador del potencial de acción cardíaco en enfermedades como el síndrome de QT largo, una alteración estructural en los canales de potasio y sodio del corazón que aumenta la predisposición de sufrir arritmias cardíacas [139]. Los genes asociados a ARNcirc fueron analizados mediante enriquecimiento funcional en *GO* con el programa *clusters_to_enrichments.R* y se observó que el chr2:178793403-178793541, un ARNcirc predicho *de novo* en este estudio, podría ser una diana terapéutica debido a su asociación con las funciones de regulación del potencial de acción del músculo cardíaco (Capítulo 10, Figura 10.2).

Como cierre de esta tesis doctoral, se han implementado y aplicado algoritmos de reducción de la dimensionalidad derivados del ACP para estudiar la expresión de los genes y los fenotipos descritos en enfermedades raras. En primer lugar, se aplicó el ACP a los datos de expresión génica de pacientes de la cohorte PMM2-CDG (usado también en el Capítulo 7) y de un experimento de organoides obtenidos a partir de pacientes con síndrome de Schaaf-Yang cultivados a distintos tiempos (Capítulo 9). En este estudio, se observó como la aplicación del ACP sobre datos de expresión génica y la expresión de los GEDs va más allá de su uso habitual como herramienta de control de calidad de las muestras. También puede emplearse para proyectar factores experimentales en el espacio de CP o para realizar una estratificación de las muestras utilizando estos componentes. El análisis de los GED en el síndrome de Schaaf-Yang mediante ACP reveló que los organoides derivados de pacientes sanos mostraban una mejor diferenciación, en función del tiempo de cultivo, en comparación con los organoides de pacientes enfermos (Capítulo 9, Figura 9.3D). Esto sugiere que estos genes podrían influir en el desarrollo de los organoides. Por otro lado, la aplicación del ACP a los datos de expresión de los GEDs en pacientes con PMM2-CDG ha permitido identificar ciertos GEDs asociados a la gravedad de la enfermedad, los cuales muestran una expresión variable entre los pacientes de alta gravedad. Los resultados también han permitido determinar que tanto la medida de MVRD como la escala ICARS están más relacionadas con el patrón de expresión de los GED de pacientes que la escala NCPRS (Capítulo 9, Figura 9.2). Dado que ICARS es la escala usada para determinar la gravedad de la

enfermedad, se puede confirmar que la medida MVRD está muy relacionada con gravedad de la enfermedad, como se ha demostrado en otros estudios [162, 116].

Posteriormente, dado que se disponía de datos de expresión de genes y datos fenotípicos en *HPO* de los mismos pacientes de la enfermedad PMM2-CDG con diferente gravedad asociada, se realizó un estudio integrador de todos estos datos (Capítulo 11). Se observó que el método de estratificación por similitud semántica de los fenotipos que usa *Cohort Analyzer* no es capaz separar en grupos diferentes los pacientes con la misma gravedad, a pesar de que se observaron diferencias (Capítulo 11, Figura 11.4). Probablemente esto se deba al gran parecido entre los fenotipos de todos los pacientes. Por consiguiente, se decidió aplicar un análisis factorial múltiple sobre los dos grupos de variables, los niveles de expresión de los GEDs y los fenotipos presentados por los pacientes. Se determinó que los genes *SLC7A8*, *MMP14* y *IMPA2*, y los fenotipos "*Inability to walk*" (HP:0002540), "*Limb dysmetria*" (HP:0002406) y "*Intention tremor*" (HP:0002080) pueden ser indicadores de la gravedad de los pacientes. En cuanto a los genes, no se ha encontrado ninguna relación con la enfermedad ni con la gravedad de la misma en la bibliografía, pero sería interesante realizar un estudio en profundidad de los mismos. Además, en este estudio se observó cierta relación entre determinadas variantes del gen *PMM2* y la gravedad de la enfermedad, pero estas relaciones no fueron significativas. Por ello, debido a que la mayoría de pacientes estudiados tienen variantes únicas, se necesita de un estudio con más pacientes que reduzca el impacto específico de las variantes.

En resumen, en el contexto de esta tesis doctoral se han desarrollado y aplicado herramientas para el análisis de datos fenotípicos y de expresión génica en enfermedades raras. Estos avances han contribuido significativamente al aumento del conocimiento sobre dichas enfermedades y han abierto nuevas líneas de investigación que servirán como base para futuros estudios.

# Capítulo 13

# Conclusiones

A partir de la presente tesis doctoral, se exponen las siguientes conclusiones:

**Conclusión I** Los flujos de trabajo desarrollados para el análisis fenotípico de pacientes y enfermedades, basados en estrategias de agrupación, han permitido confirmar que la calidad del fenotipado de pacientes es esencial para caracterizar mecanismos moleculares comunes implicados en la enfermedad, y se han aplicado al estudio de enfermedades asociadas a procesos de desregulación de la angiogénesis.

**Conclusión II** El flujo de trabajo que se ha desarrollado para el estudio de datos de expresión de enfermedades raras, que integra diversos análisis de expresión diferencial y redes de coexpresión ha permitido detectar mecanismos moleculares relevantes para las enfermedades de Lafora, PMM2-CDG, LMNA-DCM y la displasia arritmogénica.

**Conclusión III** El estudio de las relaciones entre miARN y sus genes diana mediante estrategias de correlación con datos de expresión de ARN para encontrar biomarcadores o dianas terapéuticas ha sido de utilidad para confirmar la importancia de los miR-155 y miR-146a para la enfermedad de Lafora y para analizar la relevancia funcional del miR-let-7i en la enfermedad PMM2-CDG y de los miR-196b, miR-135a, miR-182 y miR-155 en la enfermedad LMNA-DCM.

**Conclusión IV** Se han relacionado ARNcirc, miARN y sus genes diana en la displasia arritmogénica mediante estrategias de correlación y se han priorizado los ARNcirc por su similaridad fenotípica con la enfermedad. De manera específica se han descrito siete ARNcirc que afectan al gen *ALG10B* y se ha propuesto un nuevo ARNcirc relevante.

**Conclusión V** El Análisis de Componentes Principales de la expresión génica ha permitido determinar que la medida MVRD está relacionada con la gravedad de los pacientes de PMM2-CDG y el patrón de expresión presentado por su modelos celulares de fibroblastos.

**Conclusión VI** Se han estudiado e integrado datos fenotípicos y de expresión génica de PMM2-CDG mediante la aplicación de análisis de reducción de la dimensionalidad para relacionarlos con las escalas de gravedad. Esto ha llevado a identificar que los genes *SLC7A8*, *MMP14* y *IMPA2* y los fenotipos "*Inability to walk*", "*Limb dysmetria*" y "*Intention tremor*" están asociados a la gravedad de la enfermedad.

# Parte VI

# Bibliografía

# Bibliografía

[1] Arrigo Schieppati, Jan Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041, 2008.

[2] Lea Eileen Brauner, Yao Yao, Lorenz Grigull, and Frank Klawonn. Patient-Oriented Questionnaires and Machine Learning for Rare Disease Diagnosis: A Systematic Review. *Journal of Clinical Medicine*, 13(17):5132, August 2024.

[3] Maggie P. Fu, Sarah M. Merrill, Mehul Sharma, William T. Gibson, Stuart E. Turvey, and Michael S. Kobor. Rare diseases of epigenetic origin: Challenges and opportunities. *Frontiers in Genetics*, 14:1113086, February 2023.

[4] Robin Z. Hayeems, Christine Michaels-Igbokwe, Viji Venkataramanan, Taila Hartley, Meryl Acker, Meredith Gillespie, Wendy J. Ungar, Roberto Mendoza-Londona, Francois P. Bernier, Kym M. Boycott, and Deborah A. Marshall. The complexity of diagnosing rare disease: An organizing framework for outcomes research and health economics based on real-world evidence. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 24(3):694–702, March 2022.

[5] Zornitza Stark and Richard H. Scott. Genomic newborn screening for rare diseases. *Nature Reviews. Genetics*, 24(11):755–766, November 2023.

[6] Alexa T. McCray, Kimberly LeBlanc, and Undiagnosed Diseases Network. Patients as Partners in Rare Disease Diagnosis and Research. *The Yale Journal of Biology and Medicine*, 94(4):687–692, December 2021.

[7] Takeya Adachi, Ayman W. El-Hattab, Ritu Jain, Katya A. Nogales Crespo, Camila I. Quirland Lazo, Maurizio Scarpa, Marshall Summar, and Duangrurdee Wattanasirichaigoon. Enhancing Equitable Access to Rare Disease Diagnosis and Treatment around the World: A Review of Evidence, Policies, and Challenges. *International Journal of Environmental Research and Public Health*, 20(6):4732, March 2023.

[8] Tom Melvin, Marc M. Dooms, Berthold Koletzko, Mark A. Turner, Damien Kenny, Alan G. Fraser, Marc Gewillig, and Anneliene Hechtelt Jonker. Orphan and paediatric medical devices in Europe: Recommendations to support their availability for on-label and off-label clinical indications. *Expert Review of Medical Devices*, October 2024.

[9] Sangita Mishra and M. P. Venkatesh. Rare disease clinical trials in the European Union: Navigating regulatory and clinical challenges. *Orphanet Journal of Rare Diseases*, 19(1):285, July 2024.

[10] Maninder Singh Setia. Methodology Series Module 1: Cohort Studies. *Indian Journal of Dermatology*, 61(1):21, 2016-01/2016-02.

[11] Moussa Baddour, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, and Thomas Labbe. Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, pages 1–8, August 2024.

[12] Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M. Bagley, Eduard Bakštein, James P. Balhoff, Gareth Baynam, Susan M. Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F. Bodenstein, Pablo Botas, Kaan Boztug, Jolana Čady, Tiffany J. Callahan, Rhiannon Cameron, Seth J. Carbon, Francisco Castellanos, J. Harry Caufield, Lauren E. Chan, Christopher G. Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon R. Davids, Maud de Dieuleveult, Vinicius de Souza, Bert B. A. de Vries, Esther de Vries, J. Raymond DePaulo, Beata Derfalvi, Ferdinand Dhombres, Claudia Diaz-Byrd, Alexander J. M. Dingemans, Bruno Donadille, Michael Duyzend, Reem Elfeky, Shahim Essaid, Carolina Fabrizzi, Giovanna Fico, Helen V. Firth, Yun Freudenberg-Hua, Janice M. Fullerton, Davera L. Gabriel, Kimberly Gilmour, Jessica Giordano, Fernando S. Goes, Rachel Gore Moses, Ian Green, Matthias Griese, Tudor Groza, Weihong Gu, Julia Guthrie, Benjamin Gyori, Ada Hamosh, Marc Hanauer, Kateřina Hanušová, Yongqun Oliver He, Harshad Hegde, Ingo Helbig, Kateřina Holasová, Charles Tapley Hoyt, Shangzhi Huang, Eric Hurwitz, Julius O. B. Jacobsen, Xiaofeng Jiang, Lisa Joseph, Kamyar Keramatian, Bryan King, Katrin Knoflach, David A. Koolen, Megan L. Kraus, Carlo Kroll, Maaike Kusters, Markus S. Ladewig, David Lagorce, Meng-Chuan Lai, Pablo Lapunzina, Bryan Laraway, David Lewis-Smith, Xiarong Li, Caterina Lucano, Marzieh Majd, Mary L. Marazita, Victor Martinez-Glez, Toby H. McHenry, Melvin G. McInnis, Julie A. McMurry, Michaela Mihulová, Caitlin E. Millett, Philip B. Mitchell, Veronika Moslerová, Kenji Narutomi, Shahrzad Nematollahi, Julian Nevado,

Andrew A. Nierenberg, Nikola Novák Čajbiková, John I. Nurnberger, Soichi Ogishima, Daniel Olson, Abigail Ortiz, Harry Pachajoa, Guiomar Perez de Nanclares, Amy Peters, Tim Putman, Christina K. Rapp, Ana Rath, Justin Reese, Lauren Rekerle, Angharad M. Roberts, Suzy Roy, Stephan J. Sanders, Catharina Schuetz, Eva C. Schulte, Thomas G. Schulze, Martin Schwarz, Katie Scott, Dominik Seelow, Berthold Seitz, Yiping Shen, Morgan N. Similuk, Eric S. Simon, Balwinder Singh, Damian Smedley, Cynthia L. Smith, Jake T. Smolinsky, Sarah Sperry, Elizabeth Stafford, Ray Stefancsik, Robin Steinhaus, Rebecca Strawbridge, Jagadish Chandrabose Sundaramurthi, Polina Talapova, Jair A. Tenorio Castano, Pavel Tesner, Rhys H. Thomas, Audrey Thurm, Marek Turnovec, Marielle E. van Gijn, Nicole A. Vasilevsky, Markéta Vlčková, Anita Walden, Kai Wang, Ron Wapner, James S. Ware, Addo A. Wiafe, Samuel A. Wiafe, Lisa D. Wiggins, Andrew E. Williams, Chen Wu, Margot J. Wyrwoll, Hui Xiong, Nefize Yalin, Yasunori Yamamoto, Lakshmi N. Yatham, Anastasia K. Yocum, Allan H. Young, Zafer Yüksel, Peter P. Zandi, Andreas Zankl, Ignacio Zarante, Miroslav Zvolský, Sabrina Toro, Leigh C. Carmody, Nomi L. Harris, Monica C. Munoz-Torres, Daniel Danis, Christopher J. Mungall, Sebastian Köhler, Melissa A. Haendel, and Peter N. Robinson. The Human Phenotype Ontology in 2024: Phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, January 2024.

[13] W. S. Robinson. A Method for Chronologically Ordering Archaeological Deposits. *American Antiquity*, 16(4):293–301, April 1951.

[14] Dekang Lin. An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[15] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Int. Res.*, 11(1):95–130, July 1999.

[16] Catia Pesquita, Daniel Faria, Hugo Bastos, António E.N. Ferreira, André O. Falcão, and Francisco M. Couto. Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics*, 9(SUPPL. 5):1–16, April 2008.

[17] Eugene Bragin, Eleni A. Chatzimichali, Caroline F. Wright, Matthew E. Hurles, Helen V. Firth, A. Paul Bevan, and G. Jawahar Swaminathan. DECIPHER: Database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic acids research*, 42(Database issue), January 2014.

[18] Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, January 2015.

[19] Melissa J. Landrum, Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O'Leary, George R. Riley, Wenyao Shi, George Zhou, Valerie Schneider, Donna Maglott, J. Bradley Holmes, and Brandi L. Kattman. ClinVar: Improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

[20] Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808, 2012.

[21] Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. The Comparative Toxicogenomics Database: Update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, January 2017.

[22] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, January 2014.

[23] Peter N. Robinson, Sebastian Köhler, Anika Oellrich, Sanger Mouse Genetics, Kai Wang, Christopher J. Mungall, Suzanna E. Lewis, Nicole Washington, Sebastian Bauer, Dominik Seelow, Peter Krawitz, Christian Gilissen, Melissa Haendel, and Damian Smedley. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 2014.

[24] Noha Alnazzawi, Paul Thompson, Riza Batista-Navarro, and Sophia Ananiadou. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Medical Informatics and Decision Making*, 15(2):S3, June 2015.

[25] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, December 1977.

[26] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, November 2021.

[27] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, June 2016.

[28] Heena Satam, Kandarp Joshi, Upasana Mangrolia, Sanober Waghoo, Gulnaz Zaidi, Shravani Rawool, Ritesh P. Thakare, Shahid Banday, Alok K. Mishra, Gautam Das, and Sunil K. Malonia. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, 12(7):997, July 2023.

[29] Francis Crick. *What Mad Pursuit: A Personal View of Scientific Discovery.* Penguin, 1990.

[30] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[31] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970.

[32] Xiaofeng Dai, Shuo Zhang, and Kathia Zaleta-Rivera. RNA: Interactions drive functionalities. *Molecular Biology Reports*, 47(2):1413, December 2019.

[33] Luca F. R. Gebert and Ian J. MacRae. Regulation of microRNA function in animals. *Nature Reviews. Molecular Cell Biology*, 20(1):21–37, January 2019.

[34] Amrutha Menon, Noraini Abd-Aziz, Kanwal Khalid, Chit Laa Poh, and Rakesh Naidu. miRNA: A Promising Therapeutic Target in Cancer. *International Journal of Molecular Sciences*, 23(19):11502, September 2022.

[35] Lobera Es, Varela Ma, Jimenez Rl, and Moreno Rb. miRNA as biomarker in lung cancer. *Molecular biology reports*, 50(11), November 2023.

[36] Han Yang, Yufang Liu, Longqing Chen, Juanjuan Zhao, Mengmeng Guo, Xu Zhao, Zhenke Wen, Zhixu He, Chao Chen, and Lin Xu. MiRNA-Based Therapies for Lung Cancer: Opportunities and Challenges? *Biomolecules*, 13(6):877, May 2023.

[37] Haobo Li, Margaret H. Hastings, James Rhee, Lena E. Trager, Jason D. Roh, and Anthony Rosenzweig. Targeting Age-Related Pathways in Heart Failure. *Circulation Research*, 126(4):533–551, February 2020.

[38] Amanda Shen-Yee Kong, Kok-Song Lai, Swee-Hua Erin Lim, Sivakumar Sivalingam, Jiun-Yan Loh, and Sathiya Maran. miRNA in Ischemic Heart Disease and Its Potential as Biomarkers: A Comprehensive Review. *International Journal of Molecular Sciences*, 23(16):9001, August 2022.

[39] Elena Alonso-Villa, Fernando Bonet, Francisco Hernandez-Torres, Óscar Campuzano, Georgia Sarquella-Brugada, Maribel Quezada-Feijoo, Mónica Ramos, Alipio Mangas, and Rocío Toro. The Role of MicroRNAs in Dilated Cardiomyopathy: New Insights for an Old Entity. *International Journal of Molecular Sciences*, 23(21):13573, November 2022.

[40] José Santiago Ibáñez-Cabellos, Federico V. Pallardó, José Luis García-Giménez, and Marta Seco-Cervera. Oxidative Stress and Epigenetics: miRNA Involvement in Rare Autoimmune Diseases. *Antioxidants (Basel, Switzerland)*, 12(4):800, March 2023.

[41] Carlos Romá-Mateo, Sheila Lorente-Pozo, Lucía Márquez-Thibaut, Mireia Moreno-Estellés, Concepción Garcés, Daymé González, Marcos Lahuerta, Carmen Aguado, José Luis García-Giménez, Pascual Sanz, and Federico V. Pallardó. Age-Related microRNA Overexpression in Lafora Disease Male Mice Provides Links between Neuroinflammation and Oxidative Stress. *International Journal of Molecular Sciences*, 24(2):1089, January 2023.

[42] Himanshu Goel and Amy Goel. MicroRNA and Rare Human Diseases. *Genes*, 15(10):1243, September 2024.

[43] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, September 2008.

[44] Chu-Xiao Liu and Ling-Ling Chen. Circular RNAs: Characterization, cellular roles, and applications. *Cell*, 185(12):2016–2034, June 2022.

[45] Kirk J Mantione, Richard M. Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M. Samuel, and George B. Stefano. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research*, 20:138–141, August 2014.

[46] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, January 2015.

[47] Muhammad Farooq Rai, Eric D. Tycksen, Linda J. Sandell, and Robert H. Brophy. Advantages of RNA-seq Compared to RNA Microarrays for Transcriptome Profiling of Anterior Cruciate Ligament Tears. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, 36(1):484–497, January 2018.

[48] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15, January 2013.

[49] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, 2012.

[50] Liyan Gao, Zhide Fang, Kui Zhang, Degui Zhi, and Xiangqin Cui. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, 27(5):662–669, March 2011.

[51] Huajuan Shi, Ying Zhou, Erteng Jia, Min Pan, Yunfei Bai, and Qinyu Ge. Bias in RNA-seq Library Preparation: Current Challenges and Solutions. *BioMed Research International*, 2021:6647597, 2021.

[52] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), December 2014.

[53] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, November 2009.

[54] Sonia Tarazona, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21), 2015.

[55] Peter Langfelder and Steve Horvath. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):1–13, December 2008.

[56] Albert-László Barabási and Márton Pósfai. *Network Science*. Cambridge University Press, Cambridge, 2016.

[57] Elena Díaz-Santiago, Fernando M. Jabato, Elena Rojano, Pedro Seoane, Florencio Pazos, James R. Perkins, and Juan A.G. Ranea. Phenotype-genotype

comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS Genetics*, 16(10), October 2020.

[58] Elena Díaz-Santiago, M. Gonzalo Claros, Raquel Yahyaoui, Yolanda de Diego-Otero, Rocío Calvo, Janet Hoenicka, Francesc Palau, Juan A.G. Ranea, and James R. Perkins. Decoding Neuromuscular Disorders Using Phenotypic Clusters Obtained From Co-Occurrence Networks. *Frontiers in molecular biosciences*, 8, April 2021.

[59] F. M. Jabato, Pedro Seoane, James R. Perkins, Elena Rojano, Adrián García Moreno, M. Chagoyen, Florencio Pazos, and Juan A.G. Ranea. Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations. *Human Genetics*, 140(3):457–475, March 2021.

[60] Elena Rojano, Fernando M. Jabato, James R. Perkins, José Córdoba-Caballero, Federico García-Criado, Ian Sillitoe, Christine Orengo, Juan A.G. Ranea, and Pedro Seoane-Zonjic. Assigning protein function from domain-function associations using DomFun. *BMC bioinformatics*, 23(1), December 2022.

[61] Avi Ma'ayan. Introduction to Network Analysis in Systems Biology. *Science signaling*, 4(190):tr5, September 2011.

[62] Theodosia Charitou, Kenneth Bryan, and David J. Lynn. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, 48(1):27, March 2016.

[63] Pietro Hiram Guzzi and Swarup Roy. 4 - Complex network models. In Pietro Hiram Guzzi and Swarup Roy, editors, *Biological Network Analysis*, pages 53–75. Academic Press, January 2020.

[64] Yishu Liu, Xue Li, Chao Chen, Nan Ding, Shiyu Ma, and Ming Yang. Exploration of compatibility rules and discovery of active ingredients in TCM formulas by network pharmacology. *Chinese Herbal Medicines*, 16(4):572–588, October 2024.

[65] Yuanfang Ren, Ahmet Ay, and Tamer Kahveci. Shortest path counting in probabilistic biological networks. *BMC Bioinformatics*, 19(1):465, December 2018.

[66] Mikhail Tuzhilin. Relations between average shortest path length and another centralities in graphs, December 2024.



[67] Juan I.Fuxman Bass, Alos Diallo, Justin Nelson, Juan M. Soto, Chad L. Myers, and Albertha J.M. Walhout. Using networks to measure similarity between genes: Association index selection. *Nature methods*, 10(12):1169, December 2013.

[68] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Chris-

tina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023.

[69] Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, January 2024.

[70] Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, Tooba Abbassi-Daloii, Elisson N Lopes, Aishwarya Iyer, Javier Millán Acosta, Lars G Willighagen, Kozo Nishida, Anders Riutta, Helena Basaric, Chris T Evelo, Egon L Willighagen, Martina Kutmon, and Alexander R Pico. WikiPathways 2024: Next generation pathway database. *Nucleic Acids Research*, 52(D1):D679–D689, January 2024.

[71] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 27(1):29–34, January 1999.

[72] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, August 2021.

[73] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. PGC-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, July 2003.

[74] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[75] Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, February 2015.

[76] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, July 2006.

[77] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[78] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62, 1936.

[79] Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533, 2009.

[80] Anneke T. Vulto-van Silfhout, Jayne Y. Hehir-Kwa, Bregje W M van Bon, Janneke H M Schuurs-Hoeijmakers, Stephen Meader, Claudia J M Hellebrekers, Ilse J M Thoonen, Arjan P M de Brouwer, Han G. Brunner, Caleb Webber, Rolph Pfundt, Nicole de Leeuw, and Bert B A De Vries. Clinical Significance of De Novo and Inherited Copy-Number Variation. *Human Mutation*, 34(12):1679–1687, 2013.

[81] James R. Perkins, Pedro Seoane, Fernando M. Jabato, José Cordoba-Cabllero, Elena Rojano, Rocío Bautista, M. Gonzalo Claros, Isabel Gonzalez, and Juan A. G. Ranea. ExpHunterSuite: Package For The Comprehensive Analysis Of Transcriptomic Data, 2024.

[82] Luis Rodríguez-Caso, Armando Reyes-Palomares, Francisca Sánchez-Jiménez, Ana R. Quesada, and Miguel Ángel Medina. What is known on angiogenesis-related rare diseases? A systematic review of literature. *Journal of Cellular and Molecular Medicine*, 16(12):2872, December 2012.

[83] Tim E Putman, Kevin Schaper, Nicolas Matentzoglu, Vincent P Rubinetti, Faisal S Alquaddoomi, Corey Cox, J Harry Caufield, Glass Elsarboukh, Sarah Gehrke, Harshad Hegde, Justin T Reese, Ian Braun, Richard M Bruskiewich, Luca Cappelletti, Seth Carbon, Anita R Caron, Lauren E Chan, Christopher G Chute, Katherina G Cortes, Vinícius De Souza, Tommaso Fontana, Nomi L Harris, Emily L Hartley, Eric Hurwitz, Julius O B Jacobsen, Madan Krishnamurthy, Bryan J Laraway, James A McLaughlin, Julie A McMurry, Sierra A T Moxon, Kathleen R Mullen, Shawn T O'Neil, Kent A Shefchek, Ray Stefancsik, Sabrina Toro, Nicole A Vasilevsky, Ramona L Walls, Patricia L Whetzel, David Osumi-Sutherland, Damian Smedley, Peter N Robinson, Christopher J Mungall, Melissa A Haendel, and Monica C Munoz-Torres. The Monarch Initiative in 2024: An analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52(D1):D938–D949, January 2024.

[84] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023.

[85] Rodrigo Coutinho De Almeida, Yolande F.M. Ramos, Ahmed Mahfouz, Wouter Den Hollander, Nico Lakenberg, Evelyn Houtman, Marcella Van Hoolwerff, H. Eka D. Suchiman, Alejandro Rodríguez Ruiz, P. Eline Slagboom, Hailiang Mei, Szymon M. Kielbasa, Rob G.H.H. Nelissen, Marcel Reinders, and Ingrid Meulenbelt. RNA sequencing data integration reveals an miRNA interactome of osteoarthritis cartilage. *Annals of the Rheumatic Diseases*, 78(2):270–277, February 2019.

[86] Duy N. Do, Pier Luc Dudemaine, Bridget E. Fomenky, and Eveline M. Ibeagha-Awemu. Integration of miRNA weighted gene co-expression network and miRNA-mRNA co-expression analyses reveals potential regulatory functions of miRNAs in calf rumen development. *Genomics*, 111(4):849–859, July 2019.

[87] Edoardo Missiaglia, Chris J. Shepherd, Ewa Aladowicz, David Olmos, Joanna Selfe, Gaëlle Pierron, Olivier Delattre, Zoe Walters, and Janet Shipley. MicroRNA and gene co-expression networks characterize biological and clinical behavior of rhabdomyosarcomas. *Cancer Letters*, 385:251–260, January 2017.

[88] Priscila S.N. De Oliveira, Luiz L. Coutinho, Aline S.M. Cesar, Wellison J. Da Silva Diniz, Marcela M. De Souza, Bruno G. Andrade, James E. Koltes, Gerson B. Mourão, Adhemar Zerlotini, James M. Reecy, and Luciana C.A. Regitano. Co-expression networks reveal potential regulatory roles of miRNAs in fatty acid composition of Nelore cattle. *Frontiers in Genetics*, 10(JUL), 2019.

[89] Fernando M. Jabato, José Córdoba-Caballero, Elena Rojano, Carlos Román-Mateo, Pascual Sanz, Belén Pérez, Diana Gallego, Pedro Seoane, Juan A.G. Ranea, and James R. Perkins. Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite. *Scientific Reports 2021 11:1*, 11(1):1–12, July 2021.

[90] Yuanbin Ru, Katerina J. Kechris, Boris Tabakoff, Paula Hoffman, Richard A. Radcliffe, Russell Bowler, Spencer Mahaffey, Simona Rossi, George A. Calin, Lynne Bemis, and Dan Theodorescu. The multiMiR R package and database: Integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Research*, 42(17):e133–e133, September 2014.

[91] Keunhong Son, Sungryul Yu, Wonseok Shin, Kyudong Han, and Keunsoo Kang. A Simple Guideline to Assess the Characteristics of RNA-Seq Data. *BioMed Research International*, 2018(1):2906292, 2018.

[92] Koki Tsuyuzaki, Hiroyuki Sato, Kenta Sato, and Itoshi Nikaido. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, 21(1):9, January 2020.

[93] Xiaoying Chen, Bo Zhang, Ting Wang, Azad Bonni, and Guoyan Zhao. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinformatics*, 21(1):269, June 2020.

[94] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, January 2016.

[95] Farnoosh Abbas-Aghababazadeh, Qian Li, and Brooke L. Fridley. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS ONE*, 13(10):e0206312, October 2018.

[96] Samuel Marguerat and Jürg Bähler. RNA-seq: From technology to biology. *Cellular and Molecular Life Sciences*, 67(4):569–579, February 2010.

[97] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25:1–18, March 2008.

[98] Michael Greenacre, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, December 2022.

[99] Oskar Bruning, Wendy Rodenburg, Paul F. K. Wackers, Conny van Oostrom, Martijs J. Jonker, Rob J. Dekker, Han Rauwerda, Wim A. Ensink, Annemieke de Vries, and Timo M. Breit. Confounding Factors in the Transcriptome Analysis of an In-Vivo Exposure Experiment. *PLoS ONE*, 11(1), 2016.

[100] Irina Chadaeva, Rimma Kozhemyakina, Svetlana Shikhevich, Anton Bogomolov, Ekaterina Kondratyuk, Dmitry Oshchepkov, Yuriy L. Orlov, and Arcady L. Markel. A Principal Components Analysis and Functional Annotation of Differentially Expressed Genes in Brain Regions of Gray Rats Selected for Tame or Aggressive Behavior. *International Journal of Molecular Sciences*, 25(9):4613, April 2024.

[101] Hongxu Liu, Maojin Yao, and Jiaoyan Ren. Codonopsis pilosula-derived glycopeptide dCP1 promotes the polarization of tumor-associated macrophage from M2-like to M1 phenotype. *Cancer Immunology, Immunotherapy : CII*, 73(7):128, May 2024.

[102] Peter Nambala, Harry Noyes, Joyce Namulondo, Oscar Nyangiri, Vincent Pius Alibu, Barbara Nerima, Annette MacLeod, Enock Matovu, Janelisa Musaya, and Julius Mulindwa. Transcriptome profiles of Trypanosoma brucei rhodesiense in Malawi reveal focus specific gene expression profiles associated with pathology. *PLOS Neglected Tropical Diseases*, 18(5):e0011516, May 2024.

[103] Kokilavani Sivaraman, Bin Liu, Beatriz Martinez-Delgado, Julia Held, Manuela Büttner, Thomas Illig, Sonja Volland, Gema Gomez-Mariano, Nils Jedicke, Tetyana Yevsa, Tobias Welte, David S. DeLuca, Sabine Wrenger, Beata Olejnicka, and Sabina Janciauskiene. Human Bronchial Epithelial Cell Transcriptome Changes in Response to Serum from Patients with Different Status of Inflammation. *Lung*, 202(2):157–170, 2024.

[104] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics (Oxford, England)*, 17(9):763–774, September 2001.

[105] Federico Marini and Harald Binder. pcaExplorer: An R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*, 20:331, June 2019.

[106] Kevin Blighe and Aaron Lun. PCAtools: PCAtools: Everything principal components analysis, 2023.

[107] Laure Frésard, Craig Smail, Nicole M. Ferraro, Nicole A. Teran, Xin Li, Kevin S. Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, Brunilda Balliu, Joe R. Davis, Boxiang Liu, Cameron J. Prybol, Jennefer N. Kohler, Diane B. Zastrow, Chloe M. Reuter, Dianna G. Fisk, Megan E. Grove, Jean M. Davidson, Taila Hartley, Ruchi Joshi, Benjamin J. Strober, Sowmithri Utiramerur, David R. Adams, Aaron Aday, Mercedes E. Alejandro, Patrick Allard, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Eva Baker, Ashok Balasubramanyam, Hayk Barseghyan, Gabriel F. Batzli, Alan H. Beggs, Babak Behnam, Hugo J. Bellen, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, David P. Bick, Camille L. Birch, Devon Bonner, Braden E. Boone, Bret L. Bostwick, Lauren C. Briere, Elly Brokamp, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Shan Chen, Gary D. Clark, Terra R. Coakley, Joy D. Cogan, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Precilla D'Souza, Mariska Davids, Jean M. Davidson, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Katrina M. Dipple, Laurel A. Donnell-Fink, Naghmeh Dorrani, Daniel C. Dorset, Emilie D. Douine, David D. Draper, Annika M. Dries, Laura Duncan, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Gregory M. Enns, Ascia Eskin, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Noah D. Friedman, William A. Gahl, Emily Glanton, Rena A. Godfrey, Alica M. Goldman, David B. Goldstein, Sarah E. Gould, Jean Philippe F. Gourdine, Catherine A. Groden, Andrea L. Gropman, Melissa Haendel, Rizwan Hamid, Neil A. Hanchard, Frances High, Ingrid A. Holm, Jason Hom, Ellen M. Howerton, Yong Huang, Fariha Jamal, Yong hui Jiang, Jean M. Johnston, Angela L. Jones, Lefkothea Karaviti, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Donna M. Krasnewich, Susan Korrick, Mary Koziura, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, C. Christopher Lau, Jozef Lazar, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Shawn E. Levy, Richard A. Lewis, Sharyn A. Lincoln, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A.

MacRae, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Thomas C. Markello, Ronit Marom, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Thomas May, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Jason D. Merker, Thomas O. Metz, Matthew Might, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Jennifer L. Murphy, Donna M. Muzny, Michele E. Nehrebecky, Stan F. Nelson, J. Scott Newberry, John H. Newman, Sarah K. Nicholas, Donna Novacic, Jordan S. Orange, James P. Orengo, J. Carl Pallais, Christina Gs Palmer, Jeanette C. Papp, Neil H. Parker, Loren Dm Pena, John A. Phillips, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Jacinda B. Sampson, Susan L. Samson, Kelly Schoch, Daryl A. Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Rebecca Signer, Edwin K. Silverman, Janet S. Sinsheimer, Kevin S. Smith, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A. Sullivan, David A. Sweetser, Queenie K.G. Tan, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Tiina K. Urv, Eric Vilain, Tiphanie P. Vogel, Daryl M. Waggott, Colleen E. Wahl, Nicole M. Walley, Chris A. Walsh, Melissa Walker, Jijun Wan, Michael F. Wangler, Patricia A. Ward, Katrina M. Waters, Bobbie Jo M. Webb-Robertson, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Elizabeth A. Worthey, Shinya Yamamoto, John Yang, Yaping Yang, Amanda J. Yoon, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, Allison Zheng, Kym Boycott, Alex MacKenzie, Jacek Majewski, Michael Brudno, Dennis Bulman, David Dyment, Lars Lind, Erik Ingelsson, Alexis Battle, Gill Bejerano, Jonathan A. Bernstein, Euan A. Ashley, Kym M. Boycott, Jason D. Merker, Matthew T. Wheeler, and Stephen B. Montgomery. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine 2019 25:6*, 25(6):911–919, June 2019.

[108] Diana Gallego, Mercedes Serrano, Jose Cordoba-Caballero, Alejandra Gámez, Pedro Seoane, James R. Perkins, Juan A. G. Ranea, and Belén Pérez. Transcriptomic analysis identifies dysregulated pathways and therapeutic targets in PMM2-CDG. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1870(5):167163, June 2024.

[109] Laura Castilla-Vallmanya, Mónica Centeno-Pla, Mercedes Serrano, Héctor Franco-Valls, Raúl Martínez-Cabrera, Aina Prat-Planas, Elena Rojano, Juan A.G. Ranea, Pedro Seoane, Clara Oliva, Abraham J. Paredes-Fuentes, Gemma Marfany, Rafael Artuch, Daniel Grinberg, Raquel Rabionet, Susanna

Balcells, and Roser Urreizti. Advancing in Schaaf-Yang syndrome pathophysiology: From bedside to subcellular analyses of truncated MAGEL2. *Journal of medical genetics*, 2023.

[110] Michael D. Fountain and Christian P. Schaaf. Prader-Willi Syndrome and Schaaf-Yang Syndrome: Neurodevelopmental Diseases Intersecting at the MAGEL2 Gene. *Diseases*, 4(1):2, March 2016.

[111] Simon Thuleau and Francois Husson. FactoInvestigate: Automatic Description of Factorial Analysis, 2024.

[112] Joe H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

[113] François Husson, Julie Josse, and Jérôme Pagès. Principal Component Methods Hierarchical Clustering Partitional Clustering: Why Would We Need to Choose for Visualizing Data? *Unpublished Data*, September 2010.

[114] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, December 1953.

[115] M. Argüelles, C. Benavides, and I. Fernández. A new approach to the identification of regional clusters: Hierarchical clustering on principal components. *Applied Economics*, 46(21):2511–2519, July 2014.

[116] De Diego Almarza and Víctor José. *Evaluación Cuantitativa Clínica y Radiológica en Pacientes PMM2-CDG a través de una Red Nacional de Profesionales*. PhD thesis, Universitat de Barcelona, October 2020.

[117] Jing Liu, Jeffrey S. Barrett, Efthimia T. Leonardi, Lucy Lee, Satrajit Roychoudhury, Yong Chen, and Panayiota Trifillis. Natural History and Real-World Data in Rare Diseases: Applications, Limitations, and Future Perspectives. *Journal of Clinical Pharmacology*, 62(Suppl 2):S38–S55, December 2022.

[118] Francesca Graziano, Alessandro Zorzi, Simone Ungaro, Barbara Bauce, Ilaria Rigato, Alberto Cipriani, Martina Perazzolo Marra, Kalliopi Pilichou, Cristina Basso, and Domenico Corrado. The 2023 European Task Force Criteria for Diagnosis of Arrhythmogenic Cardiomyopathy: Historical Background and Review of Main Changes. *Reviews in Cardiovascular Medicine*, 25(9):348, September 2024.

[119] Mireia Alcalde, Rocío Toro, Fernando Bonet, José Córdoba-Caballero, Estefanía Martínez-Barrios, Juan Antonio Ranea, Marta Vallverdú-Prats, Ramon Brugada, Viviana Meraviglia, Milena Bellin, Georgia Sarquella-Brugada, and Oscar Campuzano. Role of microRNAs in arrhythmogenic cardiomyopathy: Translation as biomarkers into clinical practice. *Translational Research: The Journal of Laboratory and Clinical Medicine*, 259:72–82, September 2023.

[120] Liang Chen, Changliang Wang, Huiyan Sun, Juexin Wang, Yanchun Liang, Yan Wang, and Garry Wong. The bioinformatics toolbox for circRNA discovery and analysis. *Briefings in Bioinformatics*, 22(2):1706–1728, March 2021.

[121] Alessia Buratin, Maddalena Paganin, Enrico Gaffo, Anna Dal Molin, Juliette Roels, Giuseppe Germano, Maria Teresa Siddi, Valentina Serafin, Matthias De Decker, Stéphanie Gachet, Kaat Durinck, Frank Speleman, Tom Taghon, Geertruij Te Kronnie, Pieter Van Vlierberghe, and Stefania Bortoluzzi. Large-scale circular RNA deregulation in T-ALL: Unlocking unique ectopic expression of molecular subtypes. *Blood Advances*, 4(23):5902–5914, December 2020.

[122] Lasse S. Kristensen, Theresa Jakobsen, Henrik Hager, and Jørgen Kjems. The emerging roles of circRNAs in cancer and oncology. *Nature Reviews. Clinical Oncology*, 19(3):188–206, March 2022.

[123] Xiaolong Qi, Da-Hong Zhang, Nan Wu, Jun-Hua Xiao, Xiang Wang, and Wang Ma. ceRNA in cancer: Possible functions and clinical implications. *Journal of Medical Genetics*, 52(10):710–718, October 2015.

[124] Jakub O. Westholm, Pedro Miura, Sara Olson, Sol Shenker, Brian Joseph, Piero Sanfilippo, Susan E. Celniker, Brenton R. Graveley, and Eric C. Lai. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Reports*, 9(5):1966–1980, December 2014.

[125] Jun Cheng, Franziska Metge, and Christoph Dieterich. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics (Oxford, England)*, 32(7):1094–1096, April 2016.

[126] Xiao-Ou Zhang, Rui Dong, Yang Zhang, Jia-Lin Zhang, Zheng Luo, Jun Zhang, Ling-Ling Chen, and Li Yang. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Research*, 26(9):1277–1287, January 2016.

[127] Yuan Gao, Jinyang Zhang, and Fangqing Zhao. Circular RNA identification based on multiple seed matching. *Briefings in Bioinformatics*, 19(5):803–810, September 2018.

[128] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D. Mackowiak, Lea H. Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–338, March 2013.

[129] Daehwan Kim and Steven L. Salzberg. TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):1–15, August 2011.

[130] Steve Hoffmann, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, Lesca M. Holdt, Daniel Teupser, Jörg Hackermüller, and Peter F. Stadler. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biology*, 15(2):R34, February 2014.

[131] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, May 2013.

[132] Alessia Buratin, Stefania Bortoluzzi, and Enrico Gaffo. Systematic benchmarking of statistical methods to assess differential expression of circular RNAs. *Briefings in Bioinformatics*, 24(1):bbac612, January 2023.

[133] José Córdoba-Caballero, Elena Rojano, James R. Perkins, and Pedro Seoane. DEG_workflow: Workflow to perform Differential Expression Gene analysis from raw fastq files., 2024.

[134] Enrico Gaffo, Alessia Buratin, Anna Dal Molin, and Stefania Bortoluzzi. Sensitive, reliable and robust circRNA detection from RNA-seq with CircComPara2. *Briefings in Bioinformatics*, 23(1):bbab418, January 2022.

[135] Martina Dori, Jimmy Caroli, and Mattia Forcato. Circr, a Computational Tool to Identify miRNA:circRNA Associations. *Frontiers in Bioinformatics*, 2:852834, 2022.

[136] Mulin Jun Li, Jian Zhang, Qian Liang, Chenghao Xuan, Jiexing Wu, Peng Jiang, Wei Li, Yun Zhu, Panwen Wang, Daniel Fernandez, Yujun Shen, Yiwen Chen, Jean-Pierre A. Kocher, Ying Yu, Pak Chung Sham, Junwen Wang, Jun S. Liu, and X. Shirley Liu. Exploring genetic associations with ceRNA

regulation in the human genome. *Nucleic Acids Research*, 45(10):5653, May 2017.

[137] Binghua Kan, Guiru Yan, Yuan Shao, Ziliang Zhang, and Hui Xue. CircRNA RNF10 inhibits tumorigenicity by targeting miR-942-5p/GOLIM4 axis in breast cancer. *Environmental and Molecular Mutagenesis*, 63(7):362–372, August 2022.

[138] Fei Liu, Yang Sang, Yang Zheng, Lina Gu, Lingjiao Meng, Ziyi Li, Yuyang Dong, Zishuan Wei, Cuizhi Geng, and Meixiang Sang. circRNF10 Regulates Tumorigenic Properties and Natural Killer Cell-Mediated Cytotoxicity against Breast Cancer through the miR-934/PTEN/PI3k-Akt Axis. *Cancers*, 14(23):5862, November 2022.

[139] Wei Zhou, Dan Ye, David J. Tester, Sahej Bains, John R. Giudicessi, Carla M. Haglund-Turnquist, Kate M. Orland, Craig T. January, Lee L. Eckhardt, Kathleen R. Maginot, and Michael J. Ackerman. Elucidation of ALG10B as a Novel Long-QT Syndrome-Susceptibility Gene. *Circulation. Genomic and Precision Medicine*, 16(2):e003726, April 2023.

[140] Emily R. Holzinger, Scott M. Dudek, Alex T. Frase, Ronald M. Krauss, Marisa W. Medina, and Marylyn D. Ritchie. ATHENA: A tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 385–396, 2013.

[141] Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics (Oxford, England)*, 20(16):2626–2635, November 2004.

[142] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C. Causton, Panisa Pochanard, Eyal Mozes, Levi A. Garraway, and Dana Pe'er. An Integrated Approach to Uncover Drivers of Cancer. *Cell*, 143(6):1005–1017, December 2010.

[143] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, November 2017.

[144] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, December 2012.

[145] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14:1177932219899051, January 2020.

[146] Jérôme Pagès. *Multiple Factor Analysis by Example Using R*. Chapman and Hall/CRC, New York, November 2014.

[147] Patricia Yuste-Checa, Alejandra Gámez, Sandra Brasil, Lourdes R. Desviat, Magdalena Ugarte, Celia Pérez-Cerdá, and Belén Pérez. The Effects of PMM2-CDG-Causing Mutations on the Folding, Activity, and Stability of the PMM2 Protein. *Human Mutation*, 36(9):851–860, September 2015.

[148] Alejandra Gámez, Mercedes Serrano, Diana Gallego, Alicia Vilas, and Belén Pérez. New and potential strategies for the treatment of PMM2-CDG. *Biochimica Et Biophysica Acta. General Subjects*, 1864(11):129686, November 2020.

[149] Maria Adelaida García-Gimeno, Erwin Knecht, and Pascual Sanz. Lafora Disease: A Ubiquitination-Related Pathology. *Cells*, 7(8):87, August 2018.

[150] Isabel Gonzalez Gayte, Rocío Bautista Moreno, Pedro Seoane Zonjic, and M Gonzalo Claros. DEgenes Hunter - A Flexible R Pipeline for Automated RNA-seq Studies in Organisms without Reference Genome. *GENOMICS AND COMPUTATIONAL BIOLOGY*, 3, 2017.

[151] Marcos Lahuerta, Daymé Gonzalez, Carmen Aguado, Alihamze Fathinajafabadi, José Luis García-Giménez, Mireia Moreno-Estellés, Carlos Romá-Mateo, Erwin Knecht, Federico V. Pallardó, and Pascual Sanz. Reactive Glia-Derived Neuroinflammation: A Novel Hallmark in Lafora Progressive Myoclonus Epilepsy That Progresses with Age. *Molecular Neurobiology*, 57(3):1607–1621, March 2020.

[152] Douglas B. Gould, F. Campbell Phalan, Saskia E. van Mil, John P. Sundberg, Katayoun Vahedi, Pascale Massin, Marie Germaine Bousser, Peter Heutink, Jeffrey H. Miner, Elisabeth Tournier-Lasserve, and Simon W. M. John. Role of COL4A1 in small-vessel disease and hemorrhagic stroke. *The New England Journal of Medicine*, 354(14):1489–1496, April 2006.

[153] Marion Jeanne, Cassandre Labelle-Dumais, Jeff Jorgensen, W. Berkeley Kauffman, Grazia M. Mancini, Jack Favor, Valerie Valant, Steven M. Greenberg, Jonathan Rosand, and Douglas B. Gould. COL4A2 mutations impair COL4A1 and COL4A2 secretion and cause hemorrhagic stroke. *American Journal of Human Genetics*, 90(1):91–101, January 2012.

[154] Gabriella B. Oliveira, Luciana C.A. Regitano, Aline S.M. Cesar, James M. Reecy, Karina Y. Degaki, Mirele D. Poleti, Andrezza M. Felício, James E. Koltes, and Luiz L. Coutinho. Integrative analysis of microRNAs and mRNAs revealed regulation of composition and metabolism in Nelore cattle. *BMC Genomics*, 19(1):126, February 2018.

[155] Jan Krejci, Dalibor Mlejnek, Dana Sochorova, and Petr Nemec. Inflammatory Cardiomyopathy: A Current View on the Pathophysiology, Diagnosis, and Treatment. *BioMed Research International*, 2016:4087632, 2016.

[156] Anagha A. Divekar, Shweta Dubey, Pallavi R. Gangalum, and Ram Raj Singh. Dicer insufficiency and microRNA-155 overexpression in lupus regulatory T cells: An apparent paradox in the setting of an inflammatory milieu. *Journal of Immunology (Baltimore, Md.: 1950)*, 186(2):924–930, January 2011.

[157] Irene M. Pedersen, Guofeng Cheng, Stefan Wieland, Stefano Volinia, Carlo M. Croce, Francis V. Chisari, and Michael David. Interferon modulation of cellular microRNAs as an antiviral mechanism. *Nature*, 449(7164):919–922, October 2007.

[158] Christian Besler, Daniel Urban, Stefan Watzka, David Lang, Karl-Philipp Rommel, Reinhard Kandolf, Karin Klingel, Holger Thiele, Axel Linke, Gerhard Schuler, Volker Adams, and Philipp Lurz. Endomyocardial miR-133a levels correlate with myocardial inflammation, improved left ventricular function, and clinical outcome in patients with inflammatory cardiomyopathy. *European Journal of Heart Failure*, 18(12):1442–1451, December 2016.

[159] Xiaolin Zhang, Xiuli Shao, Ruijia Zhang, Rongli Zhu, and Rui Feng. Integrated analysis reveals the alterations that LMNA interacts with euchromatin in LMNA mutation-associated dilated cardiomyopathy. *Clinical Epigenetics*, 13(1):3, January 2021.

[160] Suet Nee Chen, Orfeo Sbaizero, Matthew R. G. Taylor, and Luisa Mestroni. Lamin A/C Cardiomyopathy: Implications for Treatment. *Current Cardiology Reports*, 21(12):160, November 2019.

[161] Gang Wang, Lai-Shan Tam, Edmund Kwok-Ming Li, Bonnie Ching-Ha Kwan, Kai-Ming Chow, Cathy Choi-Wan Luk, Philip Kam-Tao Li, and Cheuk-Chun Szeto. Serum and urinary cell-free MiR-146a and MiR-155 in patients with systemic lupus erythematosus. *The Journal of Rheumatology*, 37(12):2516–2522, December 2010.

[162] N. L. Serrano, D. Cuadras, V. de Diego, A. F. Martínez-Monseny, R. Velazquez-Fragua, L. Lopez, A. Felipe, M. C. Miranda, F. Carratala, M. L. Couce, L. G. Gutierrez-Solana, A. Macaya, J. Muchart, R. Montero, R. Artuch, C. Perez-Cerda, B. Perez, D. Itzep, B. Perez-Duenas, and M. Serrano. Quantitative assessment of the evolution of cerebellar syndrome in children with phosphomannomutase deficiency (PMM2-CDG). *European Journal of Paediatric Neurology*, 21:e2–e3, June 2017.