Métodos de análisis de enriquecimiento funcional en estudios genómicos

Francisco García García

Departamento de Bioquímica y Biología Molecular

Programa de Doctorado de Biomedicina y Biotecnología

Dirigida por:

Dr. David Montaner González

Dr. Joaquín Dopazo Blázquez

Julio 2016

Tesis tutorizada por Dr. José Enrique Pérez Ortín

Resumen

Los métodos computacionales tienen un papel fundamental en la resolución de problemas clínicos y biológicos. La generación de grandes cantidades de datos procedentes de tecnologías de alto rendimiento y el incremento de información accesible en bases de datos biológicos, han potenciado la demanda de nuevas metodologías capaces de relacionar ambos elementos.

Para hacer frente a esta necesidad de la comunidad de investigación biomédica, en este trabajo presentamos una serie de contribuciones en el campo de la Genómica Funcional que ayudarán a los investigadores en la interpretación funcional de sus análisis de datos genómicos.

Durante la primera parte de esta tesis, se describe el desarrollo y aplicación de un método de análisis de enriquecimiento orientado a los estudios de microARNs (miARNs). Estas pequeñas moléculas de ARN han sido implicadas en numerosas enfermedades. Sin embargo, en gran medida sigue siendo desconocido, su impacto específico en las vías biológicas y fenotipos celulares, además de su preciso papel en la regulación de los genes. Para mejorar y facilitar la investigación de las funciones y la regulación miARN-gen, hemos generado una metodología de enriquecimiento funcional para estudios de miARNs que permite interpretar funcionalmente este tipo de resultados.

La flexibilidad de esta propuesta posibilita la inclusión de otras variables relevantes en el estudio (ponderaciones de los elementos biológicos, relaciones entre miARN-gen…), así como un abordaje multidimensional integrando diferentes tipos de datos genómicos.

El análisis de enriquecimiento funcional de datos genómicos ofrece resultados que se enmarcan en el experimento valorado. Sin embargo, el reducido tamaño muestral de la mayoría de los experimentos y su acotación a un escenario específico, suponen factores limitantes en la evaluación de este tipo de estudios. Por ello, para mejorar la integración de diversos experimentos en el contexto funcional y proporcionar mayor claridad en su interpretación, durante la segunda parte de esta tesis, presentamos un método de metaanálisis funcional que detecta resultados de interés global, reduciendo el efecto del experimento específico. Además comprueba la consistencia de los experimentos y genera un estimador del efecto que presenta un poder estadístico mayor que el obtenido en cada experimento por separado.

Ambos enfoques proporcionan una mejor interpretación de los análisis de datos genómicos en el marco de la Biología de Sistemas.

Abstract

Computational methods play a key role for the resolution of clinical and biological problems. Generation of large amounts of data from high-throughput technologies and the increase of accessible information from biological databases have boosted the demand for new methodologies able to link both elements.

To address this need for the biomedical research community, this work presents a serie of contributions in the field of Functional Genomics to help to researchers in the functional interpretation of genomic data analysis.

In the first part of this thesis, we describe the development and implementation of an enrichment analysis oriented to miRNAs studies. These small RNA molecules have been implicated in numerous diseases. However, it remains largely unknown, its specific impact on biological pathways and cellular phenotypes in addition to its precise role in the regulation of genes. To improve and facilitate the investigation of the functions and miRNA-gene regulation, we generated a functional enrichment method for miRNAs studies to interpret such results. The flexibility of this approach allows us the inclusion of other relevant variables in the study (weights of biological elements, relationships between miRNA-gene …) and a multidimensional approach integrating

different genomic data.

Functional enrichment analysis of genomic data provides results that are part of the assessed experiment. However, the small sample size of most of the experiments and its marking to a specific scenario, represent limiting factors when evaluating such studies. Therefore, to improve the integration of various experiments in the functional context and provide clarity in the interpretation, in the second part of this thesis we present a meta-analysis method to detect functional results of global interest, reducing the effect of specific experiment. This methodology also checks the consistency of the experiments and generates an estimate of the effect having a greater statistical power than that obtained for each experiment separately.

Both approaches allow us a better interpretation of the analysis of genomic data in the context of Systems Biology.

Agradecimientos

"Sólo un exceso es recomendable en el mundo: el exceso de gratitud". Jean de La Bruyère.

La realización de esta tesis ha sido un trabajo especial para mí. No habría sido posible sin la participación directa o indirecta de un grupo de personas a las cuales me gustaría agradecer su personal colaboración.

Quería expresar mi gratitud a David y Ximo por su dedicación y por compartir conmigo su experiencia y conocimiento científico.

Gracias a todos los compañeros del Departamento por su interacción y generosidad, y por propiciar el excelente clima de trabajo que he disfrutado durante todos estos años. También me gustaría agradecer a los compañeros que han trabajado con nosotros en otros periodos de tiempo y que me han proporcionado positivas experiencias profesionales y personales.

Quería dar las gracias a mis amigos y a mi familia por su soporte con cariño y confianza.

Y finalmente quería dedicar esta tesis a José Luis por su apoyo incondicional, como siempre.

Índice general

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

El creciente volumen de datos disponibles de origen biológico y la necesidad de abordar problemas biológicos y clínicos con procedimientos computacionales, han consolidado la Bioinformática y la Biología Computacional como disciplinas esenciales para el desarrollo de la Genómica y de la Biología Molecular.

Estas técnicas nos permiten abordar simultáneamente la actividad de los genes, mutaciones, microARNs,… y mejorar el conocimiento y la comprensión de su funcionamiento. Sus aplicaciones se presentan en diferentes áreas como la farmacogenómica (respuesta a fármacos, diagnóstico predictivo), el diagnóstico molecular (identificación de patógenos, diagnóstico y pronóstico de patologías) y el estudio de la evolución de los organismos.

El desarrollo de métodos que aproximen la interpretación funcional en los resultados genómicos, supone un avance importante en la práctica investigadora y clínica, potenciando el valor traslacional de estos estudios.

1.1. Fundamentos de Biología Molecular

La célula es la unidad estructural y funcional de todos los seres vivos. En los organismos, las células organizan y distribuyen las funciones, configurando su actividad.

La Figura 1.1 muestra los elementos que se distinguen en la estructura básica celular en organismos eucariotas. En su núcleo se localizan los *cromosomas* formados por *ácido desoxirribonucleico* (ADN), que contiene la información genética.

Los componentes básicos de las moléculas de ADN son unas bases nitrogenadas llamadas *nucleótidos*. La importancia especial de los nucleótidos se debe a su capacidad de almacenamiento de la información biológica. Hay cinco tipos básicos de nucleótidos: adenina (A), guanina (G), citosina (C), timina (T) y uracilo (U). Además del ácido desoxirribonucleico, que contiene las bases A, G, C y T, también existe el ácido ribonucleico (ARN), que incorpora las bases A, G , C y U.

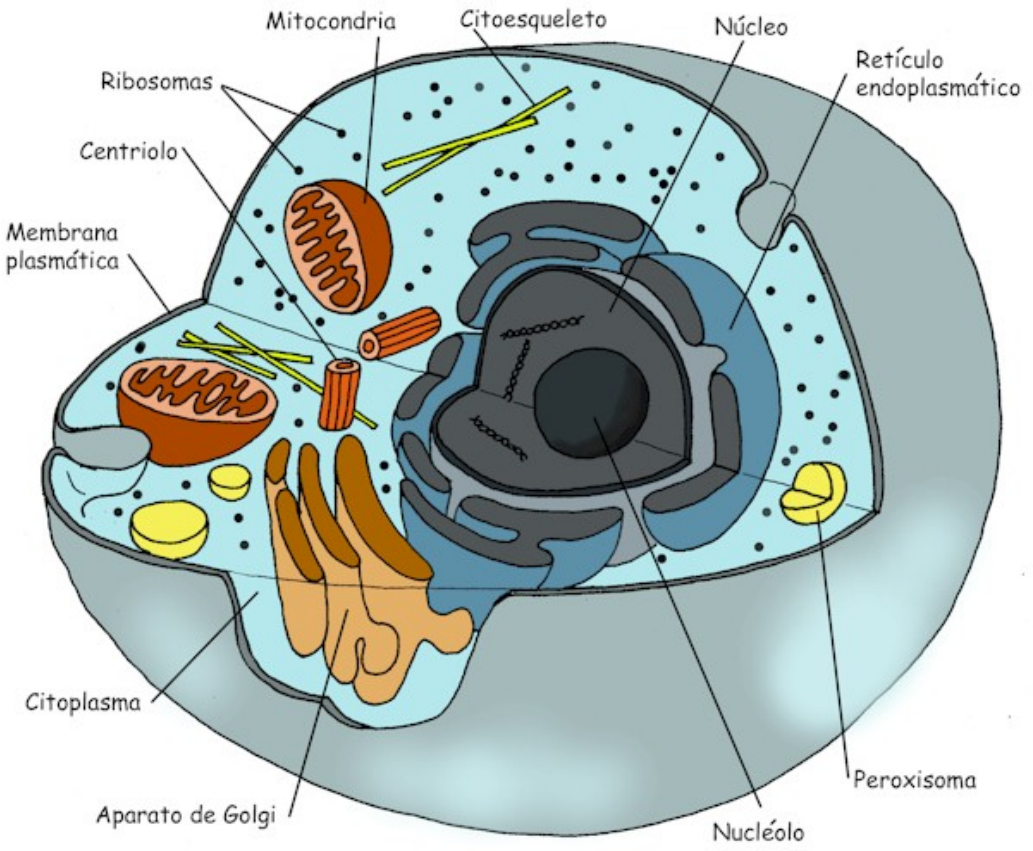

Figura 1.1: Estructura básica de la célula.

En la Figura 1.2 aparecen los diferentes niveles de estructura hasta llegar al ADN. La imagen inicial presenta el núcleo de la célula donde se encuentran los cromosomas formados por ADN. A continuación se amplía un fragmento

de ADN, donde se distinguen las 2 hebras que lo configuran, incluyendo las bases que se complementan por parejas. Por último se visualiza como un conjunto de estas bases constituye un gen.

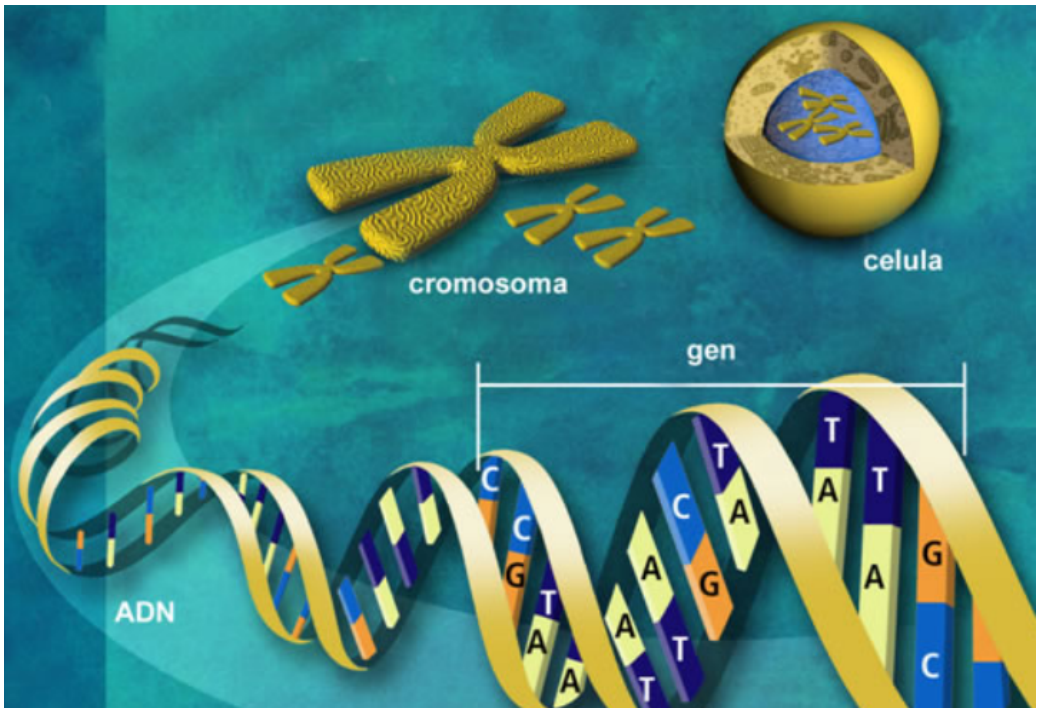

Figura 1.2: ADN, de la célula al gen.

El *dogma central de la Biología Molecular*, propuesto por Francis Crick en 1970, nos indica el proceso mediante el cual, a partir del ADN se produce la síntesis de *proteínas* (Crick, 1970), que son los elementos biológicos que marcan las pautas del comportamiento de la célula. Un esquema general de este *dogma* se muestra en la Figura 1.3.

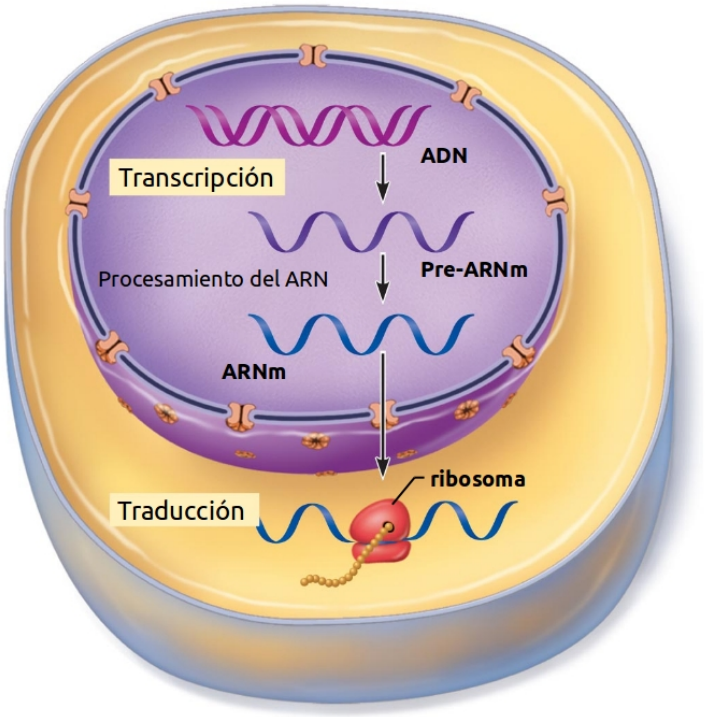

Figura 1.3: Esquema del dogma central de la Biología Molecular.

Este proceso consta básicamente de tres pasos:

- *Transcripción del ADN.* Las secuencias de ADN son copiadas a ARN mediante una enzima llamada ARN polimerasa que sintetiza un ARN mensajero (ARNm), el cual mantiene la información de la secuencia del ADN.
- *Maduración del ARNm.* Este ARN contiene tanto secuencias codificantes (exones) como no codificantes (intrones). En este proceso, el ARNm resultante de la transcripción del ADN elimina a través de unas enzimas todas las regiones no codificantes, y una vez se ha obtenido el ARNm maduro (conteniendo únicamente las zonas codificantes), éste se desplaza hasta el citoplasma (constituyendo ahora una molécula de ARNm) para dirigir la síntesis de una molécula de una proteína

determinada.

- *Traducción y síntesis de proteínas.* El ARNm transporta información genética desde los cromosomas hasta los *ribosomas* (complejos macromoleculares que sintetizan proteínas). Esta información será *leída* por la maquinaria traductora en una secuencia de tripletes de nucleótidos (codones). Cada uno de estos tripletes codifica un aminoácido específico, produciendo proteínas.

La aparición de avances científicos ha implicado la revisión y ampliación de este *dogma* (Griffiths et al., 2000).

El *nivel de expresión de un gen* es la cantidad de ARN que se transcribe, determinando la cantidad de proteínas que serán producidas por el gen. En todas las células, la expresión de cada gen individual está *regulada*, es decir, no se producen todas las proteínas posibles continuamente, sino que la célula ajusta el ritmo de la transcripción y traducción de cada gen independientemente, de acuerdo con sus necesidades puntuales. Las células tienen propiedades muy distintas como resultado de las diferencias en la abundancia, distribución y estado de sus proteínas. Estos cambios están determinados por cambios en los niveles de ARNm. Por lo tanto existe una conexión biológica entre la composición de ARNm de una célula y su estado.

Como la expresión de un gen está directamente relacionada con la cantidad de proteína que se produce, estudiando los niveles de expresión génica medimos cuánto está activado un gen en particular durante el proceso de la transcripción del ADN en ARN y la traducción de éste en proteínas.

En el proceso de la regulación génica, los *miARNs* presentan un papel im-

portante, participando en la inhibición del ARNm. Son moléculas de ARN no codificantes, que fueron identificadas por primera vez en *Caenorhabditis elegans* (Lee et al. 1993). Desde entonces, su importancia se ha puesto de manifiesto por su actividad en el control de la regulación post-transcripcional en procesos fisiológicos y patológicos. Se ha descrito su implicación en el desarrollo de tejidos (Wang et al. 2011), diferenciación celular (Stefani & Slack 2008), apoptosis (Spizzo et al. 2010), metabolismo de grasas y lípidos (Poy et al. 2004), exocitosis, división y diferenciación de células madre (Shcherbata et al. 2006) y en el desarrollo de diferentes enfermedades, entre ellas el cáncer (Spizzo et al. 2009; Rodríguez & Asfar 2013).

Los estudios de los perfiles de expresión de miARNs proporcionan una decisiva información diagnóstica, permitiendo la detección de nuevos biomarcadores. Aunque hay asociaciones descritas entre la presencia de miARNs y determinadas funciones en grupos de enfermedades, todavía hay un gran desconocimiento sobre la interpretación funcional de los estudios de los miARNs. El desarollo de metodologías que cubran este área de la Genómica Funcional mejoraría la comprensión del proceso de regulación y potenciaría sus aplicaciones biológicas y clínicas.

1.2. Tecnologías de alto rendimiento

El avance de la tecnología ha incidido en la mayoría de las áreas del conocimiento. En Biología estos cambios tecnológicos han estado motivados fundamentalmente por la aparición de un nuevo paradigma, en el que se ha dejado de estudiar de forma individualizada el gen para abordar conjun-

tamente las relaciones entre todos los genes de un organismo, es decir, su genoma. Este nuevo escenario ha potenciado el desarrollo de las *tecnologías de alto rendimiento*, que reciben este nombre porque evalúan simultáneamente la actividad de miles de genes u otras unidades biológicas. Durante las últimas dos décadas, estas tecnologías de medición a escala genómica han producido un gran cambio en la investigación biológica y en la práctica clínica.

Los *microarrays y la secuenciación masiva* constituyen las tecnologías de alto rendimiento con mayor importancia en Genómica.

1.2.1. Microarrays

La aparición de los microarrays de ADN (Lockhart et al. 1996; Schena 1996) proporcionó un primer abordaje global en estudios genómicos. El array es una superficie sólida, generalmente de vidrio, silicio o plástico, cuyo funcionamiento consiste en la medición del nivel de hibridación entre una sonda específica y la molécula diana.

Plataformas comerciales como Affymetrix, Agilent o Illumina presentaron dispositivos para el estudio de la expresión génica, la variación del número de copias o polimorfismos de un solo nucleótido (Figura 1.4). Además de la investigación biológica básica, los microarrays se han utilizado con éxito en el diagnóstico clínico (Buyse et al. 2006) y se ha demostrado que los modelos predictivos, caracterizados por su precisión y reproducibilidad, pueden ser construidos sobre la base de mediciones genómicas que proporcionan (M. Consortium & others 2010).

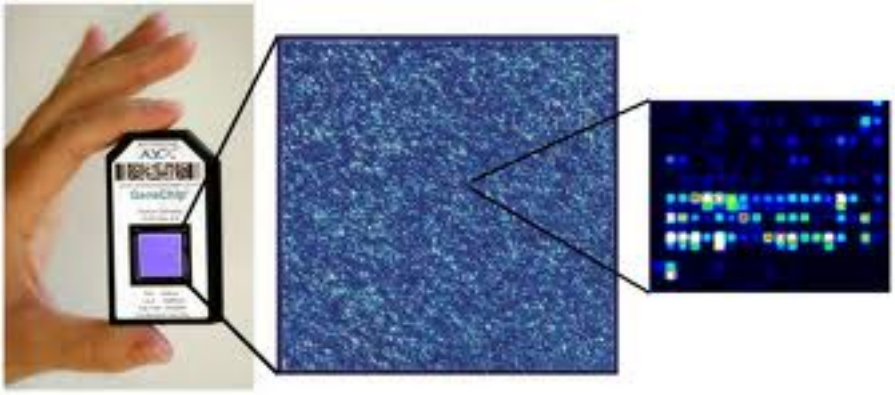

Figura 1.4: Chip de expresión génica de Affymetrix.

1.2.2. Secuenciación masiva

Más recientemente, con la aparición de la nueva generación de tecnologías de secuenciación (Church 2006; Hall 2007) se han abierto otros escenarios en el descubrimiento de nuevos genes o *isoformas* (distintas formas de una misma proteína), variantes genómicas no detectadas anteriormente y el conocimiento de miARNs noveles.

La *secuenciación de ADN* es un conjunto de métodos y técnicas bioquímicas cuya finalidad es la determinación del orden de los nucleótidos (A, C, G y T) en un fragmento de ADN. El desarrollo de la secuenciación del ADN ha acelerado significativamente la investigación y los descubrimientos en biología.

La tecnología ha fomentado el desarrollo de plataformas de secuenciación de alto rendimiento que a diferencia de los sistemas de secuenciación tradicionales, son capaces de generar paralelamente y de forma masiva, millones de fragmentos de ADN en un único proceso de secuenciación con un coste

económico y de tiempo reducidos.

La secuenciación masiva abarca diversas tecnologías:

- *Secuenciación exómica y genómica.* El *genoma* es el conjunto de todos los genes de un organismo. El *exoma* es la parte del genoma formado por los exones, es decir, las partes de los genes con capacidad de codificación y que darán lugar a las proteínas. El exoma constituye la parte funcional más importante del genoma y la que contribuye en mayor medida al *fenotipo* final de un organismo (características físicas y de comportamiento).

  La información producida por la secuenciación del exoma y el genoma facilita el descubrimiento de las bases genéticas vinculadas a determinadas enfermedades. A partir de estos resultados, los investigadores conocerán mejor los mecanismos asociados a estas patologías y podrán diseñar más eficientemente fármacos que se ajusten a las características de cada individuo desarrollando la denominada *medicina personalizada* (1. G. P. Consortium & others 2010).

- *Secuenciación de ARN.* Proporciona la medición del nivel de expresión de un gen entre diferentes grupos experimentales y la detección de nuevos genes no incluidos hasta el momento en el genoma de referencia (Carninci et al. 2005; Nagalakshmi et al. 2008). Esta tecnología permite abordar estudios con diferentes tipos de ARN como miARN, ARN de transferencia y ARN ribosómico.
- *Metilación.* Las técnicas de secuenciación masiva también son aplicables al estudio de los *patrones de metilación* (marcadores de procesos celulares) en la investigación sobre la regulación de procesos como la

diferenciación celular, el desarrollo o la aparición de enfermedades (Li et al. 2010).

- *ChIP-Seq.* Esta técnica combina la *inmunoprecipitación de la cromatina* (método bioquímico usado principalmente para determinar la localización en el genoma de determinados elementos biológicos) con la secuenciación masiva, identificando las zonas de interacción entre la proteína y el ADN (Park 2009).

A diferencia de los microarrays, con estas nuevas tecnologías no aparece un contexto cerrado o pre-definido de elementos biológicos, de modo que presentan nuevas opciones para la obtención de resultados de interés (Marioni et al. 2008; Zhao et al. 2014).

1.3. Análisis de datos genómicos

1.3.1. Estrategias de análisis

Existen diferentes estrategias de análisis en función del tipo de dato genómico que dispongamos y los objetivos establecidos en el estudio.

En los *estudios transcriptómicos*, tras el pre-procesamiento específico de los datos según el tipo de tecnología de alto rendimiento utilizada (microarrays o secuenciación masiva), dispondremos de una matriz que incluye la cuantificación de la expresión para cada unidad biológica (gen, tránscrito, miARN…) y para cada una de las muestras pertenecientes al estudio. Este es el punto de partida común para las estrategias de análisis habituales que describimos a continuación:

- *Análisis de la expresión diferencial.* Detecta el grupo de elementos biológicos (genes, miARNs…) que muestran un nivel de expresión diferenciado (activación/represión) entre las condiciones experimentales del estudio. El diseño experimental determinará qué métodos estadísticos son adecuados: comparación de dos clases, más de dos grupos, series temporales, correlación, análisis de supervivencia,… (Pan 2002; Cui et al. 2003; Smyth 2005; Huang et al. 2015). Este abordaje es utilizado en diversas áreas de estudio como la detección de biomarcadores (Conesa-Zamora, García-Solano, **García-García**, et al. 2013), la caracterización transcriptómica para la mejora del conocimiento de enfermedades (Iglesias, Leis, Perez, Gumuzio, **García-García**, et al. 2014; Gil-Ibáñez, **García-García**, Dopazo, et al. 2015) o en el desarrollo de plantas (Gutiérrez, González-Pérez, **García-García**, et al. 2014; Hillung, **García-García**, Dopazo, et al. 2016). También son extensibles estas estrategias a estudios metabolómicos o proteómicos (Moschen, Bengoa, Di Rienzo, Caro, Tohge, Watanabe, Hollmann, González, Rivarola, **García-García**, et al. 2015; Puchades-Carrasco, Jantus-Lewintre, Pérez-Rambla, **García-García**, et al. 2016).

- *Predicción de clases.* Genera un predictor supervisado de clases a partir del entrenamiento de las muestras iniciales, permitiendo la clasificación de un nuevo individuo (Lee et al. 2005).

- *Análisis clúster.* Agrupa las muestras o unidades biológicas que presentan un patrón de expresión común. En el primer caso estamos interesados en clasificar de forma no supervisada las muestras de un estudio, con el objetivo de caracterizarlas transcriptómicamente (Hillung, **García-García**, Dopazo, et al. 2016). En el segundo, el interés se cen-

tra en la búsqueda de unidades con patrones similares de expresión, que probablemente puedan ser explicados en un contexto funcional (Dubitzky et al. 2007; D'haeseleer 2005; Herrero et al. 2001; Si et al. 2013). Las técnicas de *biclustering* son capaces de tratar simultáneamente la agrupación de muestras y unidades biológicas (Madeira & Oliveira 2004).

En los estudios de *variación genómica* mediante arrays de genotipado o bien secuenciación masiva de genoma, exoma o paneles de genes, queremos conocer y priorizar las variantes genómicas asociadas a un fenotipo (Lupo, **García-García**, Sancho, et al. 2015; Urreizti, Roca, Trepat, **García-García**, et al. 2015; Romero, **García-García**, López-Perolio, et al. 2015). También desde estas tecnologías es posible detectar variaciones estructurales mediante metodologías desarrolladas específicamente para este fin.

Las *herramientas web* de análisis de datos genómicos presentan un papel fundamental en la aproximación de estas metodologías de análisis de datos genómicos a los investigadores experimentales que no están familiarizados con los procedimientos computacionales pero sí con las tecnologías de alto rendimiento en el marco de sus proyectos. *Babelomics* (Alonso, Salavert, **García-García**, et al. 2015), *BiERapp* (Alemán, **García-García**, Salavert, et al. 2014), *TEAM* (Alemán, **García-García**, Medina, et al. 2014), *HiPathia* (http://hipathia.babelomics.org) y *PathAct* (Salavert et al. 2016) son algunos de estos recursos web que facilitan el análisis de datos genómicos a este amplio grupo de usuarios.

1.3.2. Repositorios de datos

Con el desarrollo de las tecnologías de alto rendimiento y su progresiva reducción de costes, se ha generado una gran cantidad de datos procedentes de numerosos experimentos. La comunidad científica ha demandado el almacenamiento sistematizado y el acceso libre a estos datos, mediante el uso de diferentes recursos.

GEO (*Gene Expression Omnibus*) es un repositorio público del NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov) que fue generado en el año 2000. GEO almacena y difunde libremente datos procedentes de distintas tecnologías de alto rendimiento y que han sido proporcionados por los investigadores.

El proyecto *The Cancer Genome Atlas* http://cancergenome.nih.gov (McLendon & others 2008) es una colaboración entre las organizaciones NCI (*National Cancer Institute*, http://www.cancer.gov) y NHGRI (*National Human Genome Research Institute*, https://www.genome.gov) que tiene por objetivo la generación de información relevante referente a los cambios genómicos en la mayoría de tipos de cáncer. El proyecto se inició en el año 2006, recogiendo datos de más de 11000 pacientes que han posibilitado la caracterización completa de más de 30 tipos de cáncer. Los datos fueron obtenidos desde distintos tipos de tecnologías de alto rendimiento y están disponibles para la comunidad científica.

En este trabajo, hemos utilizado datos procedentes de ambos repositorios para evaluar las metodologías de análisis que presentamos.

Las diferentes estrategias de análisis descritas en el apartado anterior, identi-

fican grupos de elementos biológicos (genes, miARNs, variantes genómicas…) o bien proporcionan una ordenación de todas estas unidades siguiendo criterios estadísticos, biológicos o clínicos (tránscritos o miARNs ordenados por su nivel de expresión diferencial, variantes genómicas clasificadas según su nivel de patogenicidad u otra ordenación de interés en el estudio). Estos resultados ofrecen una respuesta inicial a la pregunta de investigación. Sin embargo, la caracterización funcional de estos resultados puede ofrecer al investigador una mejor interpretación biológica en el marco del experimento realizado.

1.4. Caracterización funcional

Las distintas estrategias de análisis descritas en el apartado anterior identifican listas de elementos biológicos de interés o bien proporcionan una ordenación de todas estas unidades que aparecen en el estudio, siguiendo criterios estadísticos, biológicos o clínicos con relevancia para los investigadores.

Por otra parte, disponemos de una gran cantidad de información biológica en bases de datos que pueden proporcionar una interpretación interesante de los resultados anteriores. En ellas se registran las *anotaciones funcionales* (funciones asociadas a unidades biológicas y que están incluidas en un vocabulario estructurado). Algunos ejemplos:

- *Gene Ontology* (Ashburner & others 2000) que contiene anotaciones como los procesos biológicos, funciones moleculares o componentes celulares que caracterizan a grupos de genes.
- *KEGG* (Kanehisa & Goto 2000) presenta redes de interacciones mole-

culares.

- Bases de datos que contienen información relativa sobre miARNs: *miRBase* (Kozomara & Griffiths-Jones 2013), *TargetScan Predicted* y *Conserved Targets* (Friedman et al. 2009).
- Y otras en las que encontramos anotación de interés funcional en variantes genómicas como la predicción del tipo de consecuencia funcional (Flicek et al. 2013), indicadores de patogenicidad (Ramensky et al. 2002; Kumar et al. 2009), fenotipos asociados (Stenson et al. 2012; Landrum et al. 2014; Consortium & others 2014), mínima frecuencia alélica en poblaciones de referencia generales (1. G. P. Consortium & others 2010; Fu et al. 2013) y locales (Dopazo, Amadoz, Bleda, García-Alonso, Alemán, **García-García**, et al. 2016).

La información contenida en estas bases de datos ha debido superar un proceso de revisión, de modo que presenta un nivel aceptable de fiabilidad, aunque no significa que sea una información completa.

Para combinar los dos elementos anteriores (resultados genómicos procedentes de los experimentos y la información biológica de las bases de datos), se necesitan metodologías para la *caracterización funcional* de los resultados de estos análisis, haciendo uso de la información suministrada por las bases de datos.

El caso más sencillo de caracterización funcional sería una *descripción de las funciones asociadas a un grupo de genes u otras unidades biológicas.* Así por ejemplo, tras consultar alguna base de datos biológicos de interés, resultaría útil conocer las funcionalidades de un grupo de genes que están diferencialmente expresados, sin embargo esta descripción puede ser coincidente con la

totalidad de genes que integran el genoma. En definitiva, necesitamos otros procedimientos que sean capaces de afinar en la caracterización funcional de resultados genómicos y que incorporen criterios robustos que garanticen su interpretación funcional.

Este tipo de corrección se presenta en los métodos de *análisis de enriquecimiento* o *análisis de sobrerrepresentación (AS)* (Al-Shahrour et al. 2004; Al-Shahrour et al. 2007). Incluyen dos etapas: en la primera se selecciona un grupo de genes (u otras unidades biológicas), considerando algún criterio de interés para el investigador (nivel de expresión diferencial de miARNs, genes que incluyan mutaciones asociadas a una enfermedad, genes donde se han detectado regiones metiladas,…). En la segunda, las funciones anotadas a estos genes se comparan contra las funciones anotadas a un grupo de genes de referencia (por ejemplo el genoma exceptuando los genes iniciales). Para cada función, se evaluará estadísticamente (prueba chi-cuadrado, test exacto de Fisher,…) la proporción de genes con la anotación a esta función en cada unas de las dos listas. De este modo será posible detectar una sobrerrepresentación significativa en alguna de las listas, es decir, ese grupo de genes estará *enriquecido* con la función valorada.

Una de las limitaciones que presentan los AS es la elección de un punto de corte a partir del cual seleccionamos un grupo de genes con diferencia de expresión, número de copias alteradas, tasa de evolución determinada…. sobre los que evaluaremos el enriquecimiento funcional. La variación de este umbral motivará que los resultados funcionales pudieran ser diferentes (Pan et al. 2005). En definitiva, hay una dependencia entre la decisión de un criterio de selección y los resultados finales obtenidos. Los análisis de grupos

de genes resuelven este problema ya que consideran todos los genes que han sido evaluados en el estudio, pero ordenados con un criterio relevante para el investigador. De modo que este abordaje, busca la detección de grupos de genes que presenten un nivel común en la característica que los ordena (nivel de expresión diferencial, perfiles de metilación,…) y que al mismo tiempo participen en las mismas funciones.

El primero de estos procedimientos sobre análisis de grupos de genes, fue presentado por Mootha et al. (2003) y posteriormente desarrollado por Subramanian et al. (2005) con el nombre de *Gene Set Enrichment Analysis (GSEA)*. Este método no selecciona únicamente genes con valores altos (o bajos) de expresión, sino que utiliza la lista completa de todos los genes y los clasifica según su nivel de expresión diferencial. El objetivo del GSEA es determinar si los genes pertenecientes a un grupo, tienden a presentarse en la parte superior (o inferior) de la lista ordenada de todos los genes por su nivel de expresión diferencial, en cuyo caso se correlaciona el conjunto de genes con la distinta clase fenotípica.

Posteriormente han aparecido diferentes aproximaciones metodológicas sobre análisis de grupos de genes. Sartor & others (2009) y Montaner & Dopazo (2010) incluyen la regresión logística en sus desarrollos metodológicos funcionales. Estos modelos estudian la dependencia entre una variable binaria y otra variable continua, por ello proporcionan una directa aplicación en el análisis funcional clásico. Además, los modelos logísticos posibilitan la inclusión de otras variables, ofreciendo un abordaje multidimensional genómico. De modo que se podría plantear un experimento donde evaluáramos simultáneamente transcriptómica y metilación. En ese caso, la variable binaria

representa la anotación o no anotación de un función a un gen (0 representaría los genes que no tienen anotada esa función y 1 los genes que sí tienen anotada esa función) y cada variable continua sería un índice que resume el nivel de expresión génica, nivel de metilación… La regresión logística nos permitirá comprobar si la anotación funcional de los genes está relacionada con la expresión diferencial, con la metilación, o con ambas.

También con la *caracterización funcional* de resultados de análisis con datos genómicos, las herramientas web juegan un papel importante en la difusión y aproximación de estas metodologías de interpretación funcional, entre el extenso grupo de usuarios que son los investigadores experimentales. Con Babelomics (Alonso et al. 2015) es posible utilizar tanto métodos de sobrerrepresentación (Al-Shahrour et al. 2004; Al-Shahrour et al. 2007) como análisis de grupos de genes (Montaner & Dopazo 2010) utilizando diferentes bases de datos y con la opción de abordar un enriquecimiento funcional con redes de interacción proteína-proteína (Minguez et al. 2009; García-Alonso et al. 2012). El uso de estas metodologías se ha aplicado con éxito en diversas áreas, mejorando el conocimiento de los mecanismos moleculares implicados en los estudios genómicos (Iglesias, Beloqui, **García-García**. et al. 2013; Puig-Butille, Escámez, **García-García**, et al. 2014; Alvarez-Mora, Rodriguez-Revenga, Madrigal, **García-García**. et al. 2015).

Con esta introducción proporcionamos un marco general con los conceptos más importantes que serán necesarios para la comprensión de este trabajo.

Capítulo 2

Motivación, objetivos y contribuciones

2.1. Motivación

La creciente importancia de la Genómica, la mayor facilidad para generar datos genómicos y la accesibilidad a las bases de datos, tanto biológicos como clínicos, han potenciado la demanda de métodos para la caracterización funcional de los resultados en los estudios genómicos. Para hacer frente a esta necesidad de la comunidad de investigación biomédica, hemos desarrollado nuevos abordajes de interpretación funcional en estudios de miARNs y en metaanálisis de estudios genómicos.

Los miARNs han sido implicados en diversos procesos fisiológicos y patológicos, incluyendo el desarrollo de diferentes enfermedades como el cáncer (Leidinger et al. 2013; Schratt et al. 2006; Xie et al. 2014). Sin embargo, sigue siendo desconocido su específico impacto en las vías biológicas y fenotipos celulares, además de su preciso papel en la regulación de los genes. Para mejorar y facilitar la investigación de las funciones y la regulación miARN-gen, hemos generado un método de enriquecimiento funcional para estudios de miARNs que permite interpretar funcionalmente este tipo de resultados.

El análisis de enriquecimiento funcional de datos procedentes de *tecnologías de alto rendimiento* ofrece resultados que se enmarcan en el experimento estudiado. Sin embargo, el reducido tamaño muestral de la mayoría de los experimentos y su acotación a un escenario específico, suponen factores limitantes en la evaluación de estos estudios genómicos (Shen & Tseng 2010; Chen et al. 2013). Por ello, para mejorar la integración de diversos experimentos en el contexto funcional y proporcionar mayor claridad en la interpretación funcional, en la segunda parte de esta tesis, presentamos un método de metaanálisis funcional capaz de detectar resultados de interés

global, reduciendo el efecto del experimento específico. Además comprueba la consistencia de los experimentos y genera un estimador del efecto que presenta un poder estadístico mayor que el obtenido en cada experimento por separado.

2.2. Objetivos

El objetivo principal de esta tesis es la generación de nuevos métodos de caracterización funcional de resultados de análisis de datos genómicos en diferentes tipos estudios, que permitan una mejor interpretación en el marco de la Biología de Sistemas.

2.2.1. Interpretación funcional para los resultados de estudios genómicos con miARNs

El primer objetivo específico fue el desarrollo de un procedimiento para la interpretación funcional de los resultados en estudios genómicos con miARNs. Se incluyeron las siguientes tareas:

1. Revisión de métodos de interpretación funcional de estudios con miARNs.
2. Desarrollo del método de análisis de grupos de miARNs.
3. Aplicación del procedimiento generado en diferentes grupos de estudios.
4. Comparación de resultados entre metodologías.

2.2.2. Métodos de metaanálisis funcional

El segundo objetivo específico se centró en la generación de un método de metaanálisis orientado a los resultados funcionales de un conjunto de estudios genómicos. A continuación se describe la secuencia de tareas desarrolladas:

1. Revisión de métodos de metaanálisis en estudios genómicos.
2. Desarrollo del método de metaanálisis para datos genómicos.
3. Aplicación del procedimiento generado en diferentes grupos estudios.
4. Evaluación de resultados obtenidos.

2.3. Principales contribuciones

La configuración y realización de esta tesis comenzó en el año 2013. Inicialmente se revisaron y abordaron diferentes estrategias de análisis de datos genómicos procedentes de distintas tecnologías de alto rendimiento. Posteriormente se seleccionaron las metodologías de análisis de enriquecimiento funcional de mayor interés para su desarrollo.

Durante el periodo comprendido entre 2013 y 2016 se han generado las siguientes contribuciones científicas que han favorecido la detección de puntos de interés en la interpretación funcional de estudios genómicos y el desarrollo de nuevos procedimientos metodológicos:

2.3.1. Publicaciones

1. González-Tendero, A., Torre, I., Garcia-Canadilla, P., Crispi, F., **García-García, F**, Dopazo, J., Bijnens, B., Gratacos, E. (**2013**). Intrauterine growth restriction is associated with cardiac ultrastructural and gene expression changes related to the energetic metabolism in a rabbit model. *American Journal of Physiology-Heart and Circulatory Physiology*, 305(12), H1752-H1760.

2. Iglesias, J. M., Beloqui, I., **García-García, F**., Leis, O., Vazquez-Martin, A., Eguiara, A., Cufi, S., Pavon, A., Menendez, J. A., Dopazo, J., Martin, A. G. (**2013**). Mammosphere formation in breast carcinoma cell lines depends upon expression of E-cadherin. *PloS one*, 8(10), e77281.

3. Puig-Butillé, J. A., Malvehy, J., Potrony, M., Trullas, C., **García-García, F.**, Dopazo, J., Puig, S. **(2013)**. Role of CPI-17 in restoring skin homoeostasis in cutaneous field of cancerization: effects of topical application of a film-forming medical device containing photolyase and UV filters. *Experimental Dermatology*, 22(7), 494-496.

4. Sánchez-Tena, S., Lizárraga, D., Miranda, A., Vinardell, M. P., **García-García, F**., Dopazo, J., Torres, J. L., Saura-Calixto, F., Capella, G., Cascante, M. (**2013**). Grape antioxidant dietary fiber inhibits intestinal polyposis in ApcMin/+ mice: relation to cell cycle and immune response. *Carcinogenesis*, 34(8), 1881-1888.

5. Silbiger, V. N., Luchessi, A. D., Hirata, R. D., Lima-Neto, L. G., Cavichioli, D., Carracedo, A., Brión, M., Dopazo, J., **García-García, F**.,

Dos Santos, E., Ramos, R. F., Sampaio, M. F., Armaganijan, D., Sousa, A., Hirata, M. H. (**2013**). Novel genes detected by transcriptional profiling from whole-blood cells in patients with early onset of acute coronary syndrome. *Clinica Chimica Acta*, 421, 184-190.

6. Sánchez-Tena, S., Reyes-Zurita, F. J., Díaz-Moralli, S., Vinardell, M. P., Reed, M., **García-García, F**., Dopazo, J., Lupiáñez, J. A., Günther, U., Cascante, M. (**2013**), Maslinic Acid-Enriched Diet Decreases Intestinal Tumorigenesis in ApcMin/+ Mice through Transcriptomic and Metabolomic Reprogramming. *PLoS One*. 2013;8(3):e59392. doi: 10.1371/journal.pone.0059392.

7. **García-García, F**., Montaner, D. **(2013)**. Web applications in teaching statistics and analysis of genomic data. Teaching statistics. *Experiences of innovation. ISBN: 978-84-9858-872-9*. Add Editorial. Pages: 201-209.

8. Conesa-Zamora, P., García-Solano, J., **García-García, F**., Del Carmen, T. M., Trujillo-Santos, J., Torres-Moreno, D., Oviedo-Ramírez, I., Carbonell-Muñoz, R., Muñoz-Delgado, E., Rodriguez-Braun, E., Conesa, A., Pérez-Guillermo, M. **(2013)**. Expression profiling shows differential molecular pathways and provides potential new diagnostic biomarkers for colorectal serrated adenocarcinoma. *Int J Cancer*,doi: 10.1002/ijc.27674.

9. Torre, I., González-Tendero, A., García-Cañadilla, P., Crispi, F., **García-García, F**., Bijnens, B., Iruretagoyena, I., Dopazo, J., Amat-Roldán, I., Gratacós, E. **(2014)**. Permanent Cardiac Sarcomere Changes in a Rabbit Model of Intrauterine Growth Restriction. *PLoS*

One 9(11): e113067. doi:10.1371/journal.pone.0113067.

10. Iglesias, J. M., Leis, O., Perez, E., Gumuzio, J., **García-García, F**., Aduriz, A., Beloqui, I., Hernández, S., Lopez-Mato, M. P., Dopazo, J., Pandiella, A., Menéndez, J., Garcia-Martin, A. **(2014)**. The activation of the Sox2 RR2 pluripotency transcriptional reporter in human breast cancer cell lines is dynamic and labels cells with higher tumorigenic potential. *Front. Oncol.*, doi: 10.3389/fonc.2014.00308.

11. Alemán, A., **García-García, F**., Medina, I., Dopazo, J. **(2014)**. A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. *Nucleic Acids Res*, pii: gku472.

12. Alemán, A., **García-García, F**., Salavert, F., Medina, I., Dopazo, J. **(2014)**. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole exome sequencing studies. *Nucleic Acids Research*, PMID: 24803668.

13. Gutiérrez, J., González-Pérez, S., **García-García, F**., Daly, C. T., Lorenzo, O., Revuelta, J. L., McCabe, P. F., Arellano, J. B. **(2014)**. Programmed cell death activated by Rose Bengal in Arabidopsis thaliana cell suspension cultures requires functional chloroplasts. *Journal of Experimental Botany*, doi:10.1093/jxb/eru151.

14. Puig-Butille, J. A., Escámez, M. J., **García-García, F**., Tell-Marti, G., Fabra, A., Martínez-Santamaría, L., Badenas, C., Aguilera, P., Pevida, M., Dopazo, D., del Río, M., Puig, S. **(2014)**. Capturing the biological impact of CDKN2A and MC1R genes as an early predisposing event in melanoma and non melanoma skin cancer. *Oncotarget*, Vol 5, No 6.

15. Lupo, V., **García-García, F**., Sancho, P., Tello, C, García-Romero, M., Villarreal, L., Alberti, M. A., Sivera, R., Dopazo, J., Pascual-Pascual, S., Márquez-Infante, C., Casasnovas, C., Sevilla, T., Espinós, C. **(2015)**. Assessment of targeted next generation sequencing as a tool for the diagnosis of Charcot-Marie-Tooth disease and hereditary motor neuropathy. *The Journal of Molecular Diagnostics* DOI: http://dx.doi.org/10.1016/j.jmoldx.2015.10.005.

16. Urreizti, R., Roca, N., Trepat, J., **García-García, F**., Alemán, A., Orteschi, D, Marangi, G., Neri, G., Opitz, J. M., Dopazo, J., Cormand, B., Vilageliu, L., Balcells, S., Grinberg, D. **(2015)**. Screening of CD96 and ASXL1 in 11 patients with Opitz C or Bohring-Opitz syndromes. *Am J Med Genet*. DOI: 10.1002/ajmg.a.37418.

17. Gil-Ibañez, P., **García-García, F**., Dopazo, J., Bernal, J., Morte, B. **(2015)**. Global transcriptome analysis of primary cerebrocortical cells: Identification of genes regulated by triiodothyronine in specific cell types. *Cereb Cortex*. PMID: 26534908.

18. Carretero, M., Guerrero-Aspizua, S., Illera, N., Galvez, V., Navarro, M., **García-García, F**., Dopazo, J., Jorcano, J. L., Larcher, F., del Rio, M. **(2015)**. Differential features between chronic skin inflammatory diseases revealed in new psoriasis and atopic dermatitis skin-humanized mouse models. *J Invest Dermatol*. PMID: 26398345.

19. Moschen, S., Bengoa, S., Di Rienzo, J., Caro, M., Tohge, T., Watanabe, M., Hollmann, J., González, S., Rivarola, M., **García-García, F**., Dopazo, J., Hopp, H., Hoefgen, R., Fernie, A., Paniego, N., Fernández, P., Heinz, R. **(2015)**. Integrating transcriptomic and metabolomic

analysis to understand natural leaf senescence in sunflower. *Plant Biotechnology Journal*, doi: 10.1111/pbi.12422.

20. Alvarez-Mora, M. I., Rodriguez-Revenga, L., Madrigal, I., **García-García, F**., Duran, M., Dopazo, J., Estivill, X., Milà, M. **(2015)**. Deregulation of key signaling pathways involved in oocyte maturation in FMR1 premutation carriers with Fragile X-associated primary ovarian insufficiency. *Gene*, pii: S0378-1119(15)00748-9. doi: 10.1016/j.gene.2015.06.039.

21. Alonso, R., Salavert, F., **García-García, F**., Carbonell, J., Bleda, M., Garcia-Alonso, L., Sanchis-Juan, A., Pérez-Gil, D., Marin-Garcia, P., Sanchez, R., Cubuk, C., Hidalgo, M., Amadoz, A., Hernansaiz-Ballesteros, R., Alemán, A., Tarraga, J., Montaner, D., Medina, I., Dopazo, J. **(2015)**. Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucl. Acids Res.*, doi: 10.1093/nar/gkv384.

22. Avila-Fernandez, A., Perez-Carro, R., Cortón, M., Lopez-Molina, M. I., Campello, L., Garanto, A., Fernadez-Sanchez, L., Duijkers, L., Lopez-Martinez, M. A., Riveiro-Alvarez, R., Rodrigues, L., Sanchez-Alcudia, R., Martin-Garrido, E., Reyes, N., **García-García, F**., Dopazo, J., Garcia-Sandoval, B., Collin, R. W., Cuenca, N., Ayuso, C. **(2015)**. Whole Exome Sequencing Reveals ZNF408 as a New Gene Associated With Autosomal Recessive Retinitis Pigmentosa with Vitreal Alterations. *Human Molecular Genetics*, doi: 10.1093/hmg/ddv140.

23. Romero, A., **García-García, F**., López-Perolio, I., Ruiz, G., García-Sáenz, J. A., Garre, P., Ayllón, P., Benito, E., Dopazo, J., Díaz-Rubio, E., Caldés, T., Hoya, M. **(2015)**. BRCA1 alternative splicing landsca-

pe in breast tissue samples. _BMC Cancer. DOI: 10.1186/s12885-015-1145-9.

24. Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Alemán, A., **García-García, F**., Rodríguez, J. A., Daub, J. T., Muntané, G., Rueda, A., Vela-Boza, A., López-Domingo, F. J., Florido, J. P., Arce, P., Navarro, A., Borrego, S., Santoyo-López, J., Antiñolo, G. **(2016)**. 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. *Mol Biol Evol.*, doi: 10.1093/molbev/msw005.

25. Puchades-Carrasco, L., Jantus-Lewintre, E., Pérez-Rambla, C., **García-García, F**., Lucas, R., Calabuig, S., Blasco, A., Dopazo, J., Camps, C., Pineda-Lucena, A. **(2016)**. Serum metabolomic profiling facilitates the non-invasive identification of metabolic biomarkers associated with the onset and progression of non-small cell lung cancer. *Oncotarget*, doi: 10.18632/oncotarget.7354.

26. Hillung, J., **García-García, F**., Dopazo, J., Cuevas, J., Elena, S. **(2016)**. The transcriptomics of an experimentally evolved plant-virus interaction. *Sci Rep.* 6:24901. doi: 10.1038/srep24901.

27. **García-García, F**., Panadero, J., Dopazo, J., Montaner, D. **(2016)**. Integrated Gene Set Analysis for microRNA Studies. *Bioinformatics*, pii: btw334.

28. Puig-Butillé, J. A., Tell, G., Visconti, A., Escámez, M. J., Jimenez, P., **García-García F**., Nsengimana, J., Falchi, M., Newton-Bishop, J., Dopazo, J., del Río, Marcela., Bataille, V., Puig S. Genomic expression differences between cutaneous cells from red hair color individuals and

black hair color individuals based on bioinformatic analysis. *Oncotarget.* (En revisión).

29. Cortón M., Avila-Fernández, A., Campello, L., Sánchez M., Benavides-Mori, B, López-Molina M. I., Fernández-Sánchez, L., Sánchez Alcudia, R., Da Silva, L., Martín-Garrido, E., Zurita, O., Fernández-San José, P., Pérez-Carro, R., **García-García, F**., Dopazo, J., García-Sandoval, B., Cuenca, N. Identification of the Photoreceptor Transcriptional Co-Repressor SAMD11 as Novel Cause of Autosomal Recessive Retinitis Pigmentosa. *Sci Rep.* (En revisión).

30. Prieto J., Leon M., Ponsoda X., **García-García F**., Bort R., Serna E., Barneo-Muñoz M., Palau F., Dopazo J., Lopez C., Torres J. Dysfunctional mitochondrial fission impairs cell reprogramming. *Sci Rep.* (En revisión).

2.3.2. Herramientas web para el análisis de datos genómicos

1. *Babelomics 5*

   http://babelomics.bioinfo.cipf.es/

   Alonso, R., Salavert, F., **García-García, F**., Carbonell, J., Bleda, M., Garcia-Alonso, L., Sanchis-Juan, A., Pérez-Gil, D., Marin-Garcia, P., Sanchez, R., Cubuk, C., Hidalgo, M., Amadoz, A., Hernansaiz-Ballesteros, R., Alemán, A., Tarraga, J., Montaner, D., Medina, I., Dopazo, J.

2. *BiERapp*

http://bierapp.babelomics.org/

Alemán, A., **García-García, F**., Salavert, F., Medina, I., Dopazo, J.

3. *TEAM* (Targeted Enrichment Analysis and Management)

   http://team.babelomics.org/

   Alemán, A., **García-García, F**., Medina, I., Dopazo, J.

4. *CSVS* (CIBERER Spanish Variant Server)

   http://csvs.babelomics.org/

   Alemán, A., **García-García, F**., Salavert, F., Medina, M., Medina, I., Dopazo, J.

5. *PanelMaps*

   Juanes, J. M., **García-García, F**., Dopazo, J., Arnau, V.

2.3.3. Comunicaciones en congresos

1. Lupo, V., Tello, C., **García-García, F**., García-Romero, C., Espinós, C. (**2014**). Panel of genes for the diagnosis of Charcot-Marie-Tooth and Distal Spinal Atrophy. *Reunión anual de la Sociedad Española de Neurología*, Mallorca, España.

2. **García-García, F**., Alemán, A., Dopazo, J. (**2014**). BIER platform: analyzing and understanding genomic and biomedical data. *XII Symposium on Bioinformatics*. Sevilla, España.

3. **García-García, F**., Panadero, J., Montaner, D., Dopazo, J. (**2014**). Integrated Gene Set Analysis for microRNA Studies. *XII Symposium on Bioinformatics*. Sevilla, España.

4. Alemán, A., **García-García, F**., Salavert, F., Medina, I., Dopazo, J. **(2014)**. TEAM: A web tool for the design, analysis and management of panels of genes for clinical applications. I Jornadas de Investigación Integral en Ciencias Omicas y Estilo de Vida. Valencia, España.

5. Alemán, A., **García-García, F**., Salavert, F., Medina, I., Dopazo, J. **(2014)**. BiERapp: A web-based interactive framework for the prioritization of disease candidate genes in whole exome sequencing studies. I Jornadas de Investigación Integral en Ciencias Omicas y Estilo de Vida. Valencia, España.

6. Turpin, M. C., Carbonell-Muñoz, R., Garcia-Solano, J., Torres-Moreno, D., **García-García, F**., Conesa, A., Perez-Guillermo, M., Conesa-Zamora, P. **(2014)**. Differentially expressed functions and genes between serrated adenocarcinoma and sporadic colorectal carcinoma showing histological and molecular features of high level of microsatellite instability. *23rd Biennial Congress of the European Association for Cancer Research*, Munich, Germany. European Journal of Cancer 2014, 50, Supplement 5, S219. 10.1016/S0959-8049(14)50796-4.

7. Lupo, V., **García-García, F**., Sancho, P., Tello, C, García-Romero, M., Villarreal, L., Alberti, M. A., Sivera, R., Dopazo, J., Pascual-Pascual, S., Márquez-Infante, C., Casasnovas, C., Sevilla, T., Espinós, C. **(2015)**. Gene Panel for the Diagnosis of COP Charcot-Marie Tooth and Distal Spinal Atrophy. Congreso Nacional de Genética Humana. Palma de Mallorca.

8. Bogliolo, M., Pujol, R., Casado, J. A., **García-García, F**., Dopazo, J., Bueren, J., Surrales, J. **(2015)**. Molecular characterizacion of 44

Fanconi anemia patients by whole exome sequencing. *VIII Congreso CIBERER*. Madrid.

9. Pérez-Rambla, C.,Puchades-Carrasco, L., Jantus-Lewintre, E., **Garcia-Garcia, F.**, Lucas, R., Calabuig, S., Blasco, A., Dopazo, J., Camps, C., Pineda-Lucena, A. **(2015)**. Metabolomics by NMR Facilitates the Non-Invasive Diagnosis and Staging of NSCLC. Journal of Thoracic Oncology. Vol:10, 9, S489-S489.

10. **García-García F**., Alemán A., Salavert F., Dopazo J. (**2016**). BiER: Plataforma Bioinformática para las Enfermedades Raras. *I Congreso de Investigación Traslacional de Enfermedades Raras*. Valencia.

11. Alemán, A., Salavert, F., **García-García, F**., Dopazo, J. (**2016**). Soluciones bioinformáticas para diagnóstico mediante paneles y descubrimiento de nuevas variantes de enfermedad. *IX Congreso Anual CIBERER*, Castelldefels, Barcelona.

12. Alemán, A., Salavert, F., **García-García, F**., Dopazo, J. (**2016**). El CIBERER Spanish Variant Server y la importancia de la variación local en la investigación en enfermedades raras. *IX Congreso Anual CIBERER*, Castelldefels, Barcelona.

13. **García-García F**., Alemán A., Salavert F., Dopazo J. (**2016**). Actividad colaborativa de la plataforma BiER desde la U715. *IX Congreso Anual CIBERER*, Castelldefels, Barcelona.

14. Amadoz, A., Bleda, M., Garcia-Alonso, L., Alemán, A., **García-García, F**., Rodríguez, J. A., Daub, J. T., Muntané, G., Rueda, A., Vela-Boza, A., López-Domingo, F. J., Florido, J. P., Arce, P., Navarro,

A., Borrego, S., Santoyo-López, J., Antiñolo, G., Dopazo, J. (**2016**). Spanish Population-Specific Differences in Disease-Related Genetic Variation. *XIII Symposium on Bioinformatics*, Valencia.

15. Juanes, J. M., **García-García, F**., Dopazo, J., Arnau, V. (**2016**). PanelMaps: a web tool for detection and visualization of altered regions for targeted sequencing. *XIII Symposium on Bioinformatics*, Valencia.

16. **García-García, F**., Ansari, I., Escribano, C., Dopazo, J., Montaner, D. (**2016**). Functional Meta-Analysis for Genomic Studies. *XIII Symposium on Bioinformatics*, Valencia.

17. **García-García, F**., Alemán, A., Salavert, F., Medina, M., Mendieta, J., Dopazo, J. (**2016**). BiER collaborative projects. XIII Symposium on Bioinformatics. *XIII Symposium on Bioinformatics*, Valencia.

18. Alemán, A., Salavert, F., Medina, M., **García-García, F**., Carbonell, J., Hidalgo, M., Amadoz, A., Çubuk, C., Gallego, A., Dopazo, J. (**2016**). Web tools for the analysis of genomic data and the discovery new disease genes. *XIII Symposium on Bioinformatics*, Valencia.

19. Morin, M., Lewis, M., **García-García, F.**, Borreguero, L., Barca, V., Ajenjo, M., Dopazo, J., Steel, K. P., Moreno-Pelayo, M. A. (**2016**). RNA-seq data analysis in the miR96 mutant mouse Diminuendo reveals the nasal epithelium as a target tissue to explore drug-based therapeutic approaches. 10th Molecular Biology of Hearing and Deafness conference, Cambridge, UK, 17-20 May 2016.

2.3.4. Actividades docentes universitarias y profesionales

1. Colaborador docente en formación universitaria de postgrado:

- Máster en Bioinformática de la Universidad de Valencia. Coordinador de las asignaturas: "Estudios in silico en Biomedicina" y "Nociones básicas de Bioinformática y Genómica".
- Máster en Biotecnología Biomédica de la Universidad Politécnica de Valencia.
- Máster en Gestión y Desarrollo de Tecnologías Biomédicas de la Universidad Carlos III de Madrid.
- Máster en Bioinformática y Biología Computacional de la Escuela Nacional de Salud y el Instituto de Salud Carlos III de Madrid.

2. Profesor asociado de Estadística en la Universidad Politécnica de Valencia.

3. Colaborador docente en las unidades de Medicina del Trabajo y Enfermería Obstétrico-Ginecológica de la Escuela Valenciana de Estudios para la Salud de la Conselleria de Sanitat de Valencia.

4. Docencia en análisis bioinformático de datos genómicos:

- Herramientas bioinformáticas para el análisis de datos genómicos. Hospital La Fe, Valencia. 2013.
- Bionformatics tools to analyze Genomic Data. National Supercomputing Center IT4Innovations, Ostrava, Czech Republic. 2014.

- Medicina Personalizada. Aplicaciones clínicas y tecnologías ómicas. Escuela Valenciana de Estudios para la Salud de la Conselleria de Sanitat de Valencia. Valencia. 2015.
- Exome-Seq Data Analysis. Workshop Applied Genetics and Genomics. El Escorial, Madrid. 2015.
- NGS Data Analysis: RNA-Seq and Resequencing. Fundación Jiménez Díaz, Madrid. 2015.
- International Course of Massive Data Analysis. Centro de Investigación Príncipe Felipe. Valencia. Ediciones: 2013 y 2016.
- NGS course: from reads to candidate genes. Centro de Investigación Príncipe Felipe. Valencia. Ediciones: 2013, 2014, 2015 y 2016.
- Genomic and Transcriptomic Data Analysis. Instituto Gulbenkian de Ciència. Oeiras, Portugal. 2016.
- Herramientas bionformáticas para el análisis de datos de paneles de genes. Hospital Sant Pau. Barcelona. 2016.

5. Actividad docente en análisis estadístico de datos en Biomedicina:

- Análisis multivariante con Stata. Métodos de Investigación en Ciencias de la Salud. Escuela Valenciana de Estudios para la Salud. Generalitat Valenciana. Valencia. 2013.
- Inferencia Estadística. CEFIRE. Generalitat Valenciana. Valencia. 2014.
- Métodos estadísticos aplicados al análisis de datos con R en Vigilancia de la Salud. Escuela Valenciana de Estudios para la Salud. Generalitat Valenciana. Valencia. Ediciones: 2014 y 2016.

- Descriptive statistics, samples and populations. Research Development Programme. Centro de Investigación Príncipe Felipe. Valencia. 2015.
- Epidemiología y Estadística en Salud Reproductiva. Escuela Valenciana de Estudios para la Salud. Generalitat Valenciana. Valencia. Ediciones: 2013, 2014 y 2015.
- Statistical Graphs. Research Development Programme. Centro de Investigación Príncipe Felipe. Valencia. Ediciones: 2014, 2015 y 2016.
- Developing Statistical Intuition. Research Development Programme. Centro de Investigación Príncipe Felipe. Valencia. 2016.

Capítulo 3

Análisis de enriquecimiento para grupos de genes en estudios de microARNs

3.1. Introducción

Los microARNs (miARNs) son moléculas pequeñas de ARN no codificantes que participan en la regulación génica post-transcripcional (He & Hannon 2004). Se adhieren por complementariedad antisentido a los ARN mensajeros de otros genes constituyendo una doble hebra. Así impiden la traducción de ese gen, ya que el complejo traductor ribosómico no puede acoplarse para elaborar la proteína correspondiente. Los miARNs constituyen un método celular natural y propio de autorregulación de la expresión génica.

Desde que en 1993 se identificaran los primeros miARNs (Lee et al. 1993), se ha producido un incremento de los estudios genómicos donde se evalúan sus perfiles de expresión para la obtención de información diagnóstica con la búsqueda de nuevos biomarcadores.

Actualmente se están desarrollando un gran número de experimentos genómicos cuyo objetivo es la evaluación de la relación entre los niveles de miARN y fenotipo. Estos experimentos generalmente utilizan tecnologías de alto rendimiento (microarrays o secuenciación masiva) para determinar la expresión de los miARNs entre diferentes condiciones biológicas, seguido de un análisis de expresión diferencial para evaluar la asociación de cada miARN al fenotipo.

En este tipo de análisis, en primer lugar se suelen seleccionar los miARNs significativamente diferentes entre los grupos experimentales comparados y a continuación se exploran sus *genes diana* (conjunto de genes regulados por el miARN) para inferir posibles consecuencias funcionales de la desregulación de estos miARNs. Existen diversas bases de datos que describen

la funcionalidad de los genes, como la *Gene Ontology (GO)* (Ashburner & others 2000), *KEGG* (Kanehisa & Goto 2000) o *Reactome* (Joshi-Tope & others 2005), que son de uso común en esta segunda etapa. Algunos autores prefieren anotar primero los miARNs en las funciones de sus genes diana y luego realizar la interpretación funcional a nivel de miARN (Godard & Eyll 2015; Bleazard et al. 2015). A pesar de ser menos intuitivo, con este enfoque se ha demostrado que se reduce el efecto del sesgo de información de una base de datos. Este paradigma de dos pasos, conocido como análisis de sobrerrepresentación (AS), se ha utilizado ampliamente en experimentos de expresión de genes y actualmente suele ser el método utilizado para la determinación del perfil funcional de los genes regulados por los miARNs.

Incluso en el contexto de la expresión génica, los enfoques de los AS han sido criticados, siendo descritos algunos inconvenientes importantes (Dopazo 2009; Khatri et al. 2012).

La mayoría de estas desventajas se basan en la pérdida de información causada por el uso de este grupo seleccionado de genes y el tratamiento igualitario que se les aplica, un problema que se plantea de nuevo en el escenario de miARNs.

Así por ejemplo, en los análisis de la expresión diferencial de genes, los AS sólo consideran los genes que muestran grandes diferencias de expresión y sin embargo los cambios sutiles, en conjuntos de genes funcionalmente relacionados, pueden ser más relevantes en el marco biológico subyacente. Sesgos similares ocurren cuando se analizan los datos de expresión de miARN.

Por un lado, algunos genes pueden ser regulados por un gran cambio en un solo miARN. Si esto ocurre en un experimento, el miARN será identificado

como diferencialmente expresado y por lo tanto los AS pueden ser utilizados con las limitaciones anteriormente mencionadas.

Por otro lado, algunas otras desregulaciones de genes menos robustas, pueden pasar desapercibidas porque los miARNs que las causan, no aparecen entre los candidatos más diferencialmente expresados, por lo tanto, en estos casos el efecto conjunto de genes combinados se habrá perdido.

Además, los genes también pueden ser inhibidos por el efecto aditivo de varios cambios pequeños de miARNs (Doxakis 2010; Papapetrou 2010). Este escenario es común, pero por lo general se descuida en los AS porque los miARNs causales son poco probables que sean seleccionados en el enfoque de dos etapas.

Finalmente, un gen puede ser regulado por varios miARNs con patrones de expresión opuestos (Bleazard et al. 2015). Esto puede inducir efectos compensatorios que presumiblemente no son considerados por los enfoques de los AS. Un simple ejemplo de esta situación sería un gen modulado por dos miARNs, uno de ellos regulado en el grupo de los casos y el otro regulado en los controles. El gen aparece regulado o inhibido en ambas condiciones y por tanto, es irrelevante para la comparación de casos y controles. A pesar de esto, los algoritmos de los AS son propensos a identificar tales genes como relevantes en la comparación, ya que se han seleccionado sus miARNs reguladores en el paso de la expresión diferencial del análisis (Godard & Eyll 2015).

Por lo tanto, la aplicación de metodologías de los AS implica intrínsecamente una comprensión relativamente simplista de la Biología. En el contexto de la expresión génica, las limitaciones de los AS ya se han superado con los

métodos de análisis para el enriquecimiento de grupos de genes (*GSA, Gene Set Analysis*) (Mootha & others 2003).

Los enfoques de los GSA pueden modelizar con éxito incluso la importancia del gen más débil, reforzando por lo tanto la interpretación funcional de datos genómicos. Los métodos de GSA han estado disponibles desde hace mucho tiempo para los experimentos basados en genes. Sin embargo, las metodologías GSA no son habituales para la evaluación del perfil funcional en estudios de miARN. Esta ausencia de aplicaciones con el enfoque de GSA en datos miARN, no es realmente sorprendente por dos razones:

- En primer lugar, la anotación funcional depende normalmente de los genes, por lo tanto, con el fin de interpretar los datos de miARN (por ejemplo, en términos de GO o KEGG), los científicos deben primero definir cómo deben vincularse los miARNs y la información de una base de datos. Para este propósito, la transferencia de la evidencia experimental de miARN a gen es implícitamente necesaria.
- En segundo lugar, la mayoría de los algoritmos de los GSA presentan un análisis a nivel de gen y los pasos de enriquecimiento son muy interdependientes para ser fácilmente separados. Esta falta de flexibilidad de la mayoría de los algoritmos de los GSA dificulta su reimplementación y uso en el contexto de los miARNs.

Por ejemplo, en el clásico algoritmo GSEA (Subramanian et al. 2005), la significación estadística del enriquecimiento se evalúa usando un análisis de permutación basado en el fenotipo, aplicado a la matriz de datos de expresión génica. Por lo tanto, la etapa de la expresión diferencial se lleva a cabo dentro del esquema de remuestreo, y no se puede cambiar sin reescribir el algoritmo.

En este trabajo se propone una nueva metodología de tipo GSA para interpretar funcionalmente los datos de expresión de miARN. Aprovechando el efecto inhibidor aditivo que los miARNs pueden tener en los genes, en primer lugar, proponemos un procedimiento para la transferencia de evidencia de la expresión diferencial en miARN a nivel de gen, mediante un indicador que representa la inhibición diferencial. A continuación utilizamos modelos de regresión logística (Montaner et al. 2009; Sartor & others 2009; Montaner & Dopazo 2010) para interpretar esta información de la inhibición de los genes en términos de conjuntos de genes.

Para ilustrar la aplicación de nuestro método, hemos analizado 20 grupos diferentes de datos reales procedentes del proyecto *The Cancer Genome Atlas* (McLendon & others 2008). Las muestras tumorales se comparan con el tejido normal en un análisis de la expresión diferencial de miARN, a continuación el perfil funcional en términos de GO se lleva a cabo para cada uno de ellos. Varios términos GO, ya conocidos por estar relacionados con cáncer, aparecen como desregulados en los diferentes tipos de cáncer, validando la idoneidad de nuestro enfoque. Esperamos que nuestro algoritmo, implementado en el paquete *mdgsa* (Montaner & Dopazo 2010) de *Bioconductor* (Gentleman & others 2004), sea útil para los analistas de datos, así como los detallados materiales complementarios de este documento.

3.2. Datos

En el momento de escribir esta tesis, 32 estudios estaban disponibles en el proyecto *The Cancer Genome Atlas*. Se descargaron y analizaron 20 de

estos estudios, habiendo seleccionado aquellos que incluían información de expresión de miARN medida con tecnología Illumina HiSeq (Bentley & others 2008), y que contienen tanto muestras tumorales como sanas. La Tabla 3.1 muestra los datos descargados de los diferentes estudios y el número de muestras incluidas en cada uno de ellos.

ID	total	casos	controles	casos-controles pareados	descripción
BLCA	271	252	19	19	Bladder Urothelial Carcinoma
BRCA	807	720	87	86	Breast invasive carcinoma
CESC	218	215	3	3	Cervical squamous cell carcinoma
COAD	243	235	8	0	Colon adenocarcinoma
ESCA	113	102	11	11	Esophageal carcinoma
HNSC	519	475	44	43	Head and Neck squamous cell carcinoma
KICH	91	66	25	25	Kidney Chromophobe
KIRC	311	240	71	68	Kidney renal clear cell carcinoma
KIRP	245	211	34	34	Kidney renal papillary cell carcinoma
LIHC	283	233	50	49	Liver hepatocellular carcinoma
LUAD	474	428	46	39	Lung adenocarcinoma
LUSC	376	331	45	45	Lung squamous cell carcinoma
PAAD	100	96	4	4	Pancreatic adenocarcinoma
PCPG	182	179	3	3	Pheochromocytoma and Paraganglioma
PRAD	117	100	17	17	Prostate adenocarcinoma
READ	93	90	3	0	Rectum adenocarcinoma
SKCM	75	74	1	0	Skin Cutaneous Melanoma
STAD	345	306	39	39	Stomach adenocarcinoma
THCA	558	499	59	59	Thyroid carcinoma
UCEC	418	386	32	19	Uterine Corpus Endometrial Carcinoma

Tabla 3.1: Estudios de tumores analizados de TCGA. Las columnas indican: identificador de enfermedad en TCGA, número total de muestras en el análisis, número de muestras tumorales, número de muestras control (tejido normal sólido), número de muestras pareadas disponibles en el grupo de datos y tipo de cáncer.

3.3. Métodos

Las matrices con los conteos de la expresión de miARN fueron descargados del portal de datos *The Cancer Genome Atlas* https://tcga-data.nci.nih.gov/tcga. Con el análisis de expresión diferencial se compararon las muestras de tumores primarios respecto el tejido normal sólido, utilizando un enfoque no pareado para los 20 conjuntos de datos. Además, también

se realizó un análisis pareado para 17 de ellos: los estudios que contenían muestras tumorales y normales de la misma persona. Estos análisis a nivel de miARN fueron realizados utilizando el paquete *edgeR* (Robinson & others 2010) de *Bioconductor*.

Para cada comparación se determinaron los correspondientes contrastes estadísticos y los valores p a nivel de miARN. El valor p representa la fuerza de la expresión miARN diferencial entre los casos y controles, mientras que el signo del estadísico del contraste, indica el sentido o "dirección" de esa diferencia; en nuestro caso, valores estadísticos positivos indican sobreexpresión en los casos comparados con los controles y los valores estadísticos negativos muestran infraexpresión. Para cada miARN, esta doble información se puede combinar en un índice único que representa simultáneamente la fuerza y el sentido de la expresión diferencial usando la siguiente transformación:

$$r = -signo(\text{estadístico}) \cdot log(valor\ p) \tag{3.1}$$

Los valores calculados son comparables entre diferentes miARNs ya que representan los valores originales de p. Además, también se conserva el signo del contraste estadístico, preservando la información acerca de la "dirección" de la sobreexpresión. Por lo tanto, es un índice que clasifica los miARNs de acuerdo con su nivel de expresión diferencial: desde los miARNs más sobreexpresados en los casos (altos valores positivos) hasta los que están más infraexpresados en los casos, (índices que son más negativos). De acuerdo con la definición, miARNs con un valor de índice cercano a cero, son aquellos con los niveles de expresión similares en ambos casos y controles, es decir, los que no están expresados diferencialmente. En este caso, se han obtenido

estos valores utilizando los procedimientos implementados en *edgeR*, aunque cualquier otro método estadístico e incluso los clásicos fold-changes, podrían ser aplicables para obtener un índice de ordenación.

3.3.1. Efecto aditivo de los miARNs sobre los genes

Las moléculas de miARN regulan la expresión de genes a través de la complementariedad de las bases emparejadas (Bartel 2004), por lo tanto, la inhibición de cierto gen debe ser proporcional a la cantidad de moléculas de miARN dirigidas a él. Además, muchos miARNs diferentes pueden interceptar el mismo gen, teniendo así un efecto aditivo sobre sus niveles de expresión (Gusev 2009; Lim & others 2005). Por lo tanto, la interferencia de un gen debe estar directamente relacionada con la suma de los niveles de expresión de los miARNs que lo regulan. Cuando se comparan las muestras biológicas, las diferencias en la expresión de miARN entre las condiciones experimentales pueden reflejarse en diferentes patrones de genes de inhibición. Esta inhibición diferencial de cada gen es considerada proporcional a la suma de las diferencias de expresión de sus miARNs de unión, por lo que estas relaciones podrían expresarse utilizando la siguiente ecuación:

$$t_i = \sum_{j \in G_i} r_j \tag{3.2}$$

donde t_i representa el incremento en la inhibición del gen i, r_j indica la expresión diferencial de cada miARN j y G_i es el conjunto de microARNs que regulan a un determinado gen. La utilidad de marcadores resumen del efecto de varios miARNs en un determinado gen i, han sido descrita anteriormente

(Morin & others 2008; Lee et al. 2012).

Utilizando la ecuación 3.2 se puede "transferir" la información relevante en nuestro experimento desde el nivel de miARN a nivel de gen, es decir, desde los valores de la expresión diferencial del miARN hasta la estimación de la diferencia de inhibición en un gen.

Determinando el cálculo para todos los genes en los datos experimentales de un estudio determinado, podemos derivar un nuevo índice transferido que ordena los genes de acuerdo a su diferencia de inhibición, causada por la actividad del miARN entre las condiciones biológicas. Los genes que muestran un mayor índice de inhibición diferencial serían los más propensos a ser interceptados en los casos, mientras que aquellos que muestran los índices más bajos corresponderían a los genes que están más inhibidos en los controles en comparación con los casos. Los genes con un índice de inhibición diferencial próximo a cero, son aquellos que no muestran diferencias significativas referentes a su regulación por miARNs. La Figura 3.1 muestra un resumen de la interpretación de la expresión diferencial de miARN y los resultados a nivel de gen.

Expresión diferencial media	Estadístico dif. exp. (r_j)	Nivel interpretación miARN	Índice de transferencia $t_i = \sum r_j$	Interpretación a nivel de gen
+	+	miARN j está sobreexpresado en casos	+	Gen i más inhibido en casos
−	−	miARN j sobreexpresado en controles	−	Gen i más inhibido en controles (*desregulado* en casos)

Figura 3.1: Interpretación del estadístico de la expresión diferencial a nivel de miARN y el *índice de transferencia* a nivel de gen.

Habría que señalar que un fuerte patrón de inhibición diferencial de un gen puede ser debido a una gran expresión diferencial en sólo uno de los miARNs

que lo estén regulando. Pero también es probable que algunos de estos grandes efectos, sean causados por el efecto aditivo de un gen determinado que está siendo diana de muchos miARNs diferentes, cada uno con débiles patrones de expresión diferencial entre las condiciones.

También merece la pena destacar que genes que no presentan una inhibición diferencial, puedan ser aquellos para los que ninguno de sus microARNs reguladores se expresan diferencialmente, o bien aquellos para los que los patrones de expresión diferencial de sus miARNs vinculantes se anulan entre sí, sumando 0.

Por ejemplo, en un estudio con un diseño experimental caso-control, el primer escenario sería que ninguno de los miARNs, que regulan a un gen determinado, estén diferencialmente expresados. En cuyo caso, todos los valores en la ecuación serían igual a cero y también su suma, es decir, el valor del parámetro sería cero.

El segundo escenario se produciría cuando un subconjunto de microARNs reguladores de un gen sobreexpresado, incrementa la inhibición génica en los casos, pero otro subconjunto de miARNs que están infraexpresados, aumentan la inhibición en los controles. En este caso, ambos efectos de inhibición se anulan entre sí, no produciendo diferencias regulatorias entre los casos y controles para ese gen. En este segundo caso, algunos valores serán positivos y algunos serán negativos, pero su suma resultará un valor próximo a cero.

Obviamente, para implementar la ecuación capaz de "transferir" la información de los miARNs a sus genes diana, la relación entre miARNs y genes diana debe definirse previamente. En este estudio se obtuvo esta información de las bases de datos *TargetScan Predicted* y *Conserved Targets* (Friedman et

al. 2009) pero cualquier otra fuente de información similar, podría ser utilizada con nuestro software. En la actualidad, la mayor parte de la información disponible en relación con las dianas de miARNs, se predice por métodos computacionales, los cuales tienen una precisión limitada (Selbach & others 2008) incorporando sesgos funcionales (Bleazard et al. 2015). Por lo tanto, se debe ser cauteloso en la interpretación o validación de los resultados. En cualquier caso, nuestro método y su implementación computacional mantendrán su validez, pudiendo ser utilizados con las bases de datos actuales y con las próximas que incluyan un nivel de precisión mayor.

Otra característica interesante de esta metodología consiste en la flexibilidad que ofrece la ecuación indicada, a la hora de modificarla para la inclusión de ponderaciones según la calidad de la información miARN-diana. Por otra parte, además de esta característica descrita, la ponderación puede también utilizarse para mejorar la modelización de los datos mediante la inclusión de información biológica adicional, como el número de dianas que representa cada gen en relación con los miARNs o los niveles de expresión génica, siempre que esté disponible.

3.3.2. Análisis de grupos de genes

En la sección anterior hemos descrito cómo la información de la expresión diferencial medida a nivel de miARN puede ser coherentemente "transferida" a nivel del gen, mediante el cálculo de nuestro índice de inhibición de genes. Este índice transferido implica la clasificación de los genes de forma que la regulación de genes a través del miARN es fácilmente interpretable.

Esta ordenación de los genes es informativa en sí misma, pero también tiene la ventaja de ser sencilla de interpretar en términos de grupos de genes tales como los descritos por las bases de datos GO (Ashburner & others 2000), KEGG (Kanehisa & Goto 2000) o Reactome (Joshi-Tope & others 2005), si se aplica el adecuado método de análisis de grupos de genes.

Los modelos de regresión logística han sido utilizados con éxito para el análisis de grupos de genes, a partir de una ordenación estadística. Sartor & others (2009) describen cómo este modelo puede ser utilizado para interpretar funcionalmente estudios de expresión génica diferencial y Montaner et al. (2009) introdujeron su uso en un esquema que pondera la importancia del gen. Más tarde, Montaner & Dopazo (2010) desarrollaron este abordaje en el contexto de múltiples dimensiones genómicas y se analizaron otras características genómicas distintas de la clásica expresión génica. Más recientemente, Mi & others (2012) adaptaron esta estrategia para tratar los sesgos producidos por la longitud de los genes en los estudios de secuenciación masiva de ARN.

Dada la ordenación estadística t para los genes, para cada clase funcional F que se está estudiando, este enfoque basado en la regresión logística modeliza la dependencia entre la pertenencia del gen g_i a la clase F y el valor asignado al gen de la manera siguiente:

$$\log \frac{P(g_i \in F)}{P(g_i \notin F)} = \kappa + \alpha \, t_i \tag{3.3}$$

Cuando la estimación del parámetro α que representa la pendiente es significativamente positiva, entonces indicamos que hay un enriquecimiento de

la función dada, sobre los valores altos de la ordenación t. Si la estimación de la pendiente α es negativa, entonces decimos que el enriquecimiento se produce en los valores más bajos de la clasificación.

Al interpretar nuestro índice de transferencia, un valor positivo en la clasificación representa un cierto grado de inhibición génica en los casos respecto los controles. Por lo tanto, una estimación positiva en la ecuación indica que los genes inhibidos en los casos, están enriquecidos en la función. Por el contrario, un valor negativo, corresponde a un enriquecimiento de la función en los genes que están más inhibidos en los controles que en los casos. Una estimación que no es significativamente distinta de cero indicaría que no hay un patrón de enriquecimiento conjunto de genes en relación con la ordenación que disponemos. La Figura 3.2 muestra un resumen de esta interpretación.

Signo de la pendiente de regresión (α)	Interpretación modelo logístico	Interpretación a nivel de gen	Interpretación a nivel de GO
+	La mayoría de los t_i son positivos para $g_i \in F$	La mayoría de los genes en el GO pueden ser interceptados en casos	La función está más interceptada o inhibida en casos
−	La mayoría de los t_i son negativos para $g_i \in F$	La mayoría de los genes en el GO pueden ser interceptados en controles	La función está más interceptada o inhibida en controles (*desregulada* en casos)

Figura 3.2: Interpretación del parámetros de la regresión logística en términos del gen y los *grupos de genes*.

La ecuación 2 dará como resultado $t_i = 0$, para los genes no regulados por ningún miARN y estos ceros no tendrán ningún efecto importante en la ecuación. Por lo tanto, efectivamente, nuestro análisis de grupos de genes está orientado a genes que son diana de al menos un miARN. En las estrategias de

los AS, el uso específico de genes regulados ha sido descrito como ventajoso en comparación con otros enfoques que utilizan todos los genes anotados en su evaluación (Bleazard et al. 2015; Godard & Eyll 2015).

En nuestro estudio hemos utilizado los términos GO (Ashburner & others 2000) para definir nuestros grupos de genes. La anotación de los genes fue descargada de la web de Ensembl. Hemos analizado tres ontologías: procesos biológicos, componentes celulares y funciones moleculares para obtener una estimación de la sobrerrepresentación y el correspondiente valor de p en cada uno de los términos GO examinados. Realizamos una corrección de los valores p en el contexto de las comparaciones múltiples y de esta forma se controló la tasa de falsos positivos aplicando el método de Benjamini & Yekutieli (2001).

Un diagrama de la secuencia de pasos del análisis es mostrado en la Figura 3.3. Aquí se presentan los resultados para un término GO específico: "neurofilament cytoskeleton", en el estudio del cáncer de mama (BRCA) con muestras emparejadas, que se utilizó para describir el uso del algoritmo propuesto.

3.4. Resultados

3.4.1. Nivel de microARN

El análisis de la expresión diferencial se llevó a cabo para cada tipo de cáncer utilizando *edgeR* y seguido de una corrección del valor p para controlar la tasa de falsos positivos (Benjamini & Hochberg 1995). La Tabla 3.2 muestra

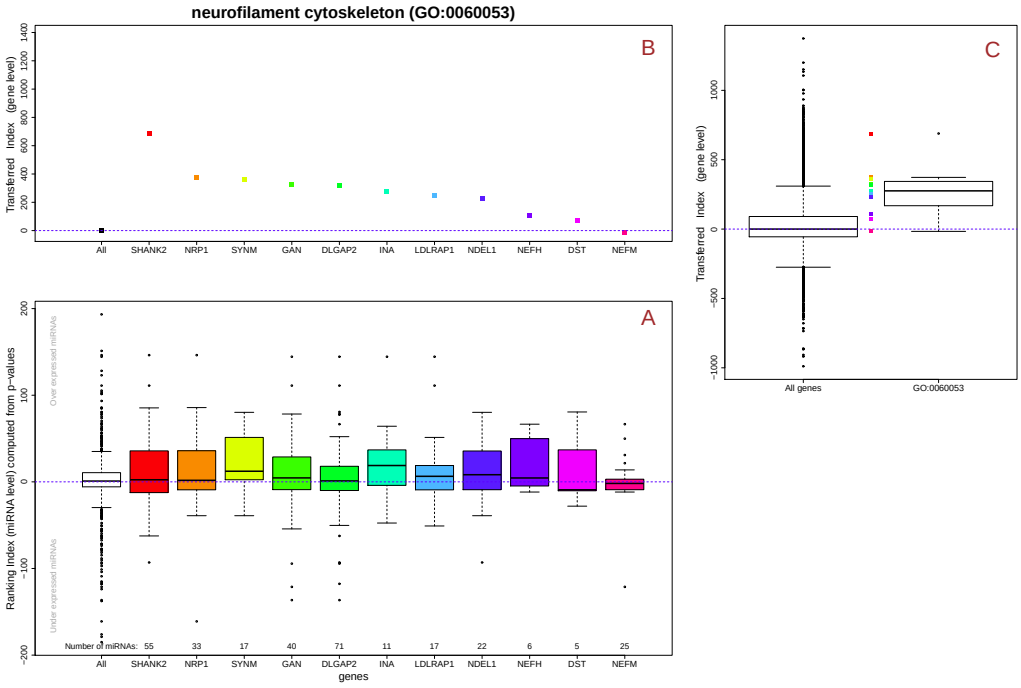

Figura 3.3: Fases del análisis para la función *neurofilament cytoskeleton* (GO:0060053). El gráfico **(A)** representa la distribución del índice de ordenación calculado según la ecuación 3.1. La caja blanca muestra la distribución de todos los miARNs en el estudio. En nuestro caso, los valores positivos pertenecen a aquellos miARNs más expresados en tumores mientras que los valores negativos se refieren a miARNs más expresados en controles. Cada una de las cajas coloreadas representa el mismo índice, pero sólo para el subconjunto de miARNs que regulan un gen en el término GO. El gráfico **(B)** representa *el índice de transferencia* del gen, introducido en la ecuación 3.2. Para cada uno de los genes en el término GO todos los índices a nivel de miARN son sumados en un único valor. El gráfico **(C)** representa la distribución del *índice de transferencia* para el genoma completo (caja izquierda) y para los genes dentro del término GO *neurofilament cytoskeleton* (puntos y caja derecha). Aquí podemos apreciar como la distribución conjunta de los genes en el término GO es mayor que la distribución basal de todos los genes. El modelo de regresión logística detecta este patrón e informa que el término GO está más enriquecido en muestras tumor, por lo que el componente celular *neurofilament cytoskeleton* está más interceptado por la acción de miARNs en casos que en controles.

el número de miARNs sobre e infrarregulados en cada uno de los tipos de cáncer en análisis pareados y no pareados. Destacamos el gran número de miARNs diferencialmente expresados, incluso después de la corrección de los valores de p por comparaciones múltiples. Esto es debido a las grandes diferencias que existen entre las muestras tumor y control, pero también destaca el gran número de miARNs que regulan los genes expresados en

un solo tejido. Como consecuencia, la interpretación de los resultados estadísticos para extraer conclusiones biológicamente significativas puede ser una tarea desalentadora.

	No pareado			Pareado		
ID	Infra	noDif	Sobre	Infra	noDif	Sobre
BLCA	128	337	353	127	343	219
BRCA	200	244	396	202	215	269
CESC	92	621	73	29	537	65
COAD	174	291	262			
ESCA	98	443	152	62	464	133
HNSC	204	285	360	164	305	222
KICH	166	297	199	217	252	169
KIRC	169	191	323	213	180	215
KIRP	221	262	295	223	242	237
LIHC	120	278	407	200	283	213
LUAD	152	292	405	130	264	259
LUSC	169	215	462	180	313	244
PAAD	23	607	11	8	606	14
PCPG	70	608	43	40	507	55
PRAD	76	429	104	38	513	31
READ	136	307	204			
SKCM	46	680	6			
STAD	152	308	356	138	307	206
THCA	218	351	257	226	347	145
UCEC	243	284	347	211	272	229

Tabla 3.2: Número de miARNs sobre, infra y no diferencialmente regulados en cada tipo de cáncer.

Las dificultades en la interpretación sobre qué funciones biológicas están desreguladas por miARNs en cáncer, se hacen más evidentes si exploramos los genes que son dianas de estos miARNs diferencialmente expresados. La Tabla 3.3 muestra el número de genes que son diana de los miARNs sobre e infrarregulados en cada tipo de cáncer. Algunos efectos de saturación pueden ser causados por la gran cantidad de miARNs expresados diferencialmente y también por el número aún mayor de los genes diana conocidos para cada miARN. En promedio, 8000 genes son diana de miARNs sobre e

infrarregulados, y por otra parte el número de genes comunes en las dianas de miARNs es muy alto, alrededor de 6000 (Tabla 3.3). En algunos casos extremos, más teórico que práctico, la mayoría de los genes en el genoma podrían ser diana simultáneamente por miARNs sobre o infrarregulados, pero a diferencia de anteriores enfoques de los AS, nuestra metodología sigue siendo una propuesta válida en estos casos.

	No pareado			Pareado		
ID	Infra	Común	Sobre	Infra	Común	Sobre
BLCA	8345	6763	8599	8087	5955	7528
BRCA	8968	7700	9465	9305	7724	9001
CESC	7834	5201	6525	4877	3178	5431
COAD	6981	6418	9998			
ESCA	7992	5646	6959	8233	5207	6212
HNSC	9090	7496	8976	9065	7006	8013
KICH	8998	7044	8252	9594	7125	7902
KIRC	8838	7351	9056	9575	7543	8681
KIRP	9169	7388	8629	9311	7025	8267
LIHC	7466	6848	9560	8896	6851	7720
LUAD	8255	7354	9898	8150	6843	8848
LUSC	8535	7265	9447	8844	6710	8166
PAAD	3759	616	1169	1529	442	1748
PCPG	6303	4033	5295	4102	3110	5652
PRAD	7422	5932	8039	4997	1600	2374
READ	6938	6225	9672			
SKCM	5983	631	857			
STAD	8921	6761	8041	8947	6731	7855
THCA	8763	7244	8702	9064	7065	8056
UCEC	9182	7171	8436	9338	7069	8201

Tabla 3.3: Número de genes afectados por miARNs sobre e infrarregulados. La columna *Común* muestra el número de genes que son dianas de ambos grupos de miARNs: los sobre e infrarregulados. El número total de genes que son dianas al menos para un miARN es **12084**.

La Tabla 3.4 muestra el número de términos GO asociados a genes que fueron regulados por miARNs en ambos sentidos. Como se puede observar, para la mayoría de los tipos de cáncer, todos los términos GO incluidos en el estudio fueron representados por estos genes. Obviamente, en este escenario,

las metodologías de los AS no tienen sentido en la interpretación funcional de los resultados. Esta situación se trata generalmente por métodos "ad hoc", tales como el incremento del punto de corte del valor p, para que un menor número de miARNs sean clasificados como diferencialmente expresados y por consiguiente, grupos más pequeños de genes necesiten ser interpretados.

En los estudios genómicos de cáncer se espera que un gran número de miARNs estén diferencialmente expresados, sin embargo, también son frecuentes los experimentos donde se detectan muy pocos o incluso ningún miARN diferencialmente expresado, debido por ejemplo a las restriciones del tamaño muestral. En tales casos, las metodologías de AS no son aplicables, pero el método de análisis de grupos de genes, como el que aquí se presenta, podría permitir a los investigadores extraer conclusiones de interés sobre los datos.

3.4.2. Nivel de gen

Tras el análisis de la expresión diferencial de miARNs, la ecuación 1 fue utilizada para resumir los valores de p y el signo del estadístico de contraste en una única ordenación estadística. Entonces, la ecuación fue aplicada para traducir esta evidencia de expresión diferencial del miARN a una escala de inhibición diferencial de genes. Para cada gen, este índice transferido condensa la información de los miARNs de los cuales son diana, conservando dos características adecuadas para la interpretación funcional del experimento: considera el efecto múltiple de cancelación de los miARNs e incorpora el efecto aditivo de pequeños eventos de inhibición.

	No pareado			Pareado		
ID	Infra	Común	Sobre	Infra	Común	Sobre
BLCA	5169	5169	5169	5169	5168	5168
BRCA	5169	5169	5169	5169	5169	5169
CESC	5169	5168	5168	5144	5138	5160
COAD	5168	5168	5169			
ESCA	5169	5168	5168	5169	5167	5167
HNSC	5169	5169	5169	5169	5169	5169
KICH	5169	5169	5169	5169	5169	5169
KIRC	5169	5169	5169	5169	5169	5169
KIRP	5169	5169	5169	5169	5169	5169
LIHC	5169	5169	5169	5169	5169	5169
LUAD	5169	5169	5169	5169	5169	5169
LUSC	5169	5169	5169	5169	5169	5169
PAAD	5129	4578	4590	4870	4681	4915
PCPG	5166	5161	5164	5150	5146	5165
PRAD	5169	5169	5169	5159	4981	4990
READ	5168	5168	5169			
SKCM	5169	4385	4385			
STAD	5169	5169	5169	5169	5169	5169
THCA	5169	5169	5169	5169	5169	5169
UCEC	5169	5169	5169	5169	5169	5169

Tabla 3.4: Número de términos GO asociados a genes diana de miARNs sobre e infrarregulados. La mayoría de los términos GO están descritos en los grupos casos y control al mismo tiempo, tal como se comprueba en la columna *Común*. El número total de términos GO anotados por los genes diana es **5169**.

Por ejemplo, el gen *GPR162* es diana de dos miARNs: *hsa-miR-22-3p* y *hsa-miR-214-3p*. En el análisis pareado de cáncer renal (KIRCH), se detectó sobreexpresión de *hsa-miR-22-3p* en muestras tumorales (con un valor de p de 5.6 x 10-30), mientras que hsa-miR-214-3p estaba infraexpresado (con un valor de p de 1.8 x 10-29). Los índices de sobreexpresión derivados de la aplicación de la ecuación 1 fueron 67.34 para *hsa-miR-22-3p* y -66.61 para *hsa-miR-214-3p*, lo que indica que hay evidencia para diferencias similares de expresión en estos dos miARNs, pero en "direcciones" opuestas. Luego, el gen *GPR162* debe inhibirse en los casos por el *miARN hsa-miR-22-3p* con la misma fuerza que se inhibe en los controles por el miARN *hsa-miR-214-3p*. Por lo tanto, nuestra interpretación es que ambos efectos de inhibición

se anulan entre sí y así, el gen *GPR162* es considerado irrelevante en el proceso del cáncer en términos de la acción del miARN. Esta cancelación se refleja en el índice transferido al gen obtenido con la ecuación 3.2 y que produce un valor de inhibición diferencial próximo a 0 para este gen: 0.73. Por otra parte, cuando se utiliza el modelo de regresión logística indicado en la ecuación 3.3 para realizar un análisis de grupo de genes del índice transferido a gen, GPR162 no apoyará el enriquecimiento de cualquiera de las funciones en las que está involucrado.

El efecto acumulativo de varios eventos débiles de expresión diferencial en miARN también se puede apreciar, por ejemplo en los resultados producidos por *GREB1*, gen regulador del crecimiento del cáncer. Este gen es diana de 16 miARNs, ninguno de los cuales tiene un valor absoluto de inhibición diferencial superior a 10 en el análisis de los datos del carcinoma esofágico (ESCA). Sin embargo, la suma de los 16 valores, supone un valor de inhibición diferencial de -53.65 para el gen, lo que indica una fuerte inhibición en muestras normales en comparación con los tumores. Llegamos a la conclusión de que *GREB1* generalmente se regula en los tejidos normales por la acción combinada de muchos miARNs y que esta regulación se pierde en tumores ESCA, y por lo tanto puede afectar el crecimiento del cáncer. Respecto el análisis de grupos de genes, *GREB1* estará asociado a los términos GO, a los cuales pertenece por estar inhibido por la acción del miARN en los controles o, equivalentemente, como desregulado en los casos.

3.4.3. Nivel de grupo de genes

Una vez que la evidencia de la expresión diferencial del miARN es transferida a los genes, el índice de ordenación de la inhibición diferencial puede ser fácilmente analizado en términos de grupos de genes utilizando un enfoque basado en la regresión logística (Montaner et al. 2009; Sartor & others 2009; Montaner & Dopazo 2010).

La Tabla 3.5 muestra el número de términos GO enriquecidos en valores positivos y negativos del índice de transferencia. En nuestro análisis, los valores positivos del índice de transferencia pertenecen a genes que son dianas de miARNs y están sobreexpresados en el cáncer. Estos genes están generalmente más inhibidos en muestras de tumores debido al efecto de los miARNs. Por lo tanto, los términos GO enriquecidos en los valores positivos del índice de transferencia a gen, representan funciones biológicas que están globalmente más inhibidas o interceptadas por el efecto del miARN en los casos que en los controles.

Del mismo modo, términos GO enriquecidos en los valores negativos del índice de transferencia a gen, representan los que tienen mayores tasas de interceptación en las muestras control que en las muestras tumorales. La interpretación biológica de este segundo grupo de funciones es que normalmente están controlados por la acción del miARN en el tejido normal y que esta coordinación se pierde en el tejido afectado, causando la desregulación de la función en estado de cáncer. Por lo tanto, en este trabajo nos referimos a los términos GO enriquecidos en los valores positivos del índice de transferencia como inhibidos o interceptados en las células cancerosas y los enriquecidos en los valores negativos, desregulados en los estados de cán-

cer. La Figura 3.2 describe y resume los principales parámetros y pasos en nuestra metodología.

En general, los patrones de inhibición o desregulación de los términos GO encontrados en los análisis pareados y no pareados están fuertemente correlacionados positivamente (ver resultados suplementarios en Apéndice 1), lo que refleja la consistencia de nuestro enfoque. A pesar de ello, el número de términos GO enriquecidos en los análisis pareados y no pareados son diferentes, lo que puede reflejar la variabilidad interindividual del papel que juegan los miARNs en el cáncer. No se ha encontrado ningún patrón de asociación entre el tamaño del término GO (número de genes en el bloque) y los niveles de significación (ver Apéndice 1), lo que indica la falta de sesgo en este aspecto del método.

No son muchos los términos GO enriquecidos que son compartidos entre los tipos de cáncer (ver resultados suplementarios en Apéndice 1). Esto era esperable debido a la gran cantidad de diferencias en los tejidos normales y tumorales, recogidos en los diferentes experimentos del proyecto The Cancer Genome Atlas. Pero también puede reflejar el papel específico de los miARNs que juegan en el desarrollo del cáncer. La mayoría de los términos enriquecidos compartidos a través de diferentes tipos de cáncer están relacionados con el desarrollo celular, ampliamente conocidos por estar relacionados con la evolución del cáncer. Por otro lado, la mayoría de los términos de GO que se enriquecen individualmente en los diferentes tipos de cáncer específicos están relacionados con el desarrollo celular, la adhesión, la señalización y la proliferación. Todos ellos claros procesos asociados al cáncer.

Por ejemplo, en nuestro análisis pareado, el componente celular *endoplas-*

mic reticulum lumen (GO: 0005788) está desregulado en los tumores BLCA, CESC y UCEC, todos ellos estrechamente relacionados con carcinomas urogenitales. El perfil completo de los grupos de genes en estudios pareados y no pareados, para los 5169 términos GO está disponible en nuestros materiales complementarios (ver resultados en Apéndice 1), donde también se incluyen comparaciones entre subgrupos pareados y no pareados.

	No pareado			Pareado		
ID	Desr.	noDif	Inh.	Desr.	noDif	Inh.
BLCA	2	5167	0	2	5167	0
BRCA	3	5166	0	0	5167	2
CESC	0	5169	0	1	5167	1
COAD	18	4930	221			
ESCA	2	5167	0	1	5168	0
HNSC	53	5116	0	0	5169	0
KICH	1	5167	1	30	5138	1
KIRC	0	5159	10	5	5163	1
KIRP	4	5165	0	13	5155	1
LIHC	7	5080	82	0	5169	0
LUAD	0	5169	0	0	5169	0
LUSC	0	5169	0	0	5169	0
PAAD	3	5165	1	0	5169	0
PCPG	0	5169	0	0	5166	3
PRAD	0	5168	1	1	5168	0
READ	0	5157	12			
SKCM	121	5043	5			
STAD	5	5164	0	0	5169	0
THCA	2	5167	0	2	5167	0
UCEC	89	5080	0	9	5160	0

Tabla 3.5: Número de términos GO significativos en el análisis de enriquecimiento funcional en estudios pareados y no pareados. Las columnas **Inh.** indican el número de términos con un coeficiente α **positivo** en el análisis de regresión logística. Esos son los términos inhibidos o interceptados en los casos. Las columnas **Desr.** indican el número de términos con un coeficiente α **negativo** en el análisis de regresión logística. Esos son los términos inhibidos en controles o *desregulados* en los casos. Las columnas **noDif** indícan el número de términos GO con un coeficiente no significativo de la pendiente del modelo de regresión.

Para evaluar el método descrito y los resultados obtenidos, se realizaron los siguientes análisis complementarios:

- Estimación del error de tipo I para la metodología propuesta.
- Perfil funcional en genes expresados.
- Comparación de resultados funcionales entre diversas propuestas metodológicas.
- Similitud funcional en grupos de tumores.

3.4.3.1. Estimación del error de tipo I

La estimación de la tasa del error de tipo I para la metodología propuesta, se determinó utilizando una estrategia de permutación de los datos. Se generó una tabla con todos los pares miARN-genes diana y a continuación, aleatoriamente se permutaron las columnas que representaban a los genes. El procedimiento se repitió 100 veces. Medianas y percentiles fueron utilizados para describir el porcentaje de resultados significativos.

La cantidad de *falsos positivos* (FP) encontrados en el análisis de permutación está por debajo del umbral esperado de acuerdo con el punto de corte seleccionado para su valor de p (menor del 5 %). (Tablas 3.6 y 3.7).

Cáncer	Mediana	Percentil 5	Percentil 95
BLCA	0	0	0.037
BRCA	0	0	0.055
CESC	0.037	0	0.405
ESCA	0.018	0	0.018

Cáncer	Mediana	Percentil 5	Percentil 95
HNSC	0.11	0.037	0.883
KICH	0.258	0.055	0.662
KIRC	0	0	0.055
KIRP	1.031	0.368	2.024
LIHC	0	0	0.147
LUAD	1.104	0.276	1.987
LUSC	1.288	0.57	2.319
PAAD	0	0	0.018
PCPG	0	0	0.055
PRAD	0.147	0	0.865
STAD	0.092	0	0.736
THCA	0	0	0
UCEC	0.129	0	0.645

Tabla 3.6: Estimación del error de tipo I en estudios pareados.

Cáncer	Mediana	Percentil 5	Percentil 95
BLCA	0	0	0.037
BRCA	0	0	0
CESC	0.202	0.018	0.386
COAD	4.803	3.607	5.332
ESCA	0.755	0.35	1.104
HNSC	0.681	0	2.3
KICH	0	0	0.037

Cáncer	Mediana	Percentil 5	Percentil 95
KIRC	0.202	0.037	1.822
KIRP	0.883	0.294	1.307
LIHC	4.141	3.257	6.294
LUAD	0	0	0
LUSC	0.883	0.074	2.079
PAAD	0.018	0	0.129
PCPG	0.018	0	0.055
PRAD	0	0	0.037
READ	0.294	0.018	0.626
SKCM	1.619	0.773	2.687
STAD	0.212	0.018	0.883
THCA	0	0	0.055
UCEC	2.843	1.012	4.734

Tabla 3.7: Estimación del error de tipo I en estudios no pareados.

3.4.3.2. Perfil funcional en genes expresados

Los métodos y programas presentados anteriormente se han orientado a la interpretación de los datos de expresión de miARNs. Sin embargo, en ocasiones también están disponibles los datos de expresión génica para el mismo grupo de sujetos que forman la muestra. En esta situación, los investigadores podrían considerar de interés, la restricción de esta interpretación funcional sólo para aquellos grupos de genes regulados por los miARNs que efectivamente están expresados.

La aplicación de este abordaje es trivial a partir de la modificación de algunas de las funciones generadas para esta metodología y que están disponibles en el paquete *mdgsa* (Montaner & Dopazo 2010) de *Bioconductor*.

Métodos

Si disponemos de una lista de genes expresados, la ecuación (3.2) puede ser modificada para incluir esta información:

$$t_i = \begin{cases} \sum_{j \in G_i} r_j & \text{si el gen } i \text{ está expresado} \\ 0 & \text{si el gen } i \text{ no está expresado} \end{cases}$$

Esta variación no cambia el método descrito, pero sí considera la no expresión de los grupos de genes regulados por miARNs.

Resultados

Como ejemplo de aplicación del enfoque descrito anteriormente, se utilizaron los datos de expresión génica y miARN de KICH (cáncer renal cromófobo) del *TCGA*.

Para ambos grupos de datos, los niveles de expresión normalizados fueron descargados desde este repositorio. Para determinar si un gen estaba expresado, se estableció como criterio que presentara un número de conteos normalizados superior a 1 en todas las muestras del estudio. Con ello, 8821 (73 %) genes de los 12084 que constituían las dianas de los miARNs, fueron descritos como expresados. La interpretación funcional se realizó siguiendo los procedimientos indicados en la metodología.

En la evaluación global de resultados, se determinó la correlación entre los resultados de la interpretación funcional con y sin la selección de los genes expresados, obteniendo coeficientes de correlación de 0.5 en los estudios pareados y 0.41 en los estudios no pareados. Ambos coeficientes fueron significativos (nivel de confianza del 95 %).

Pocos términos GO, como *histone modification* (GO:0016570) y *homophilic cell adhesion via plasma membrane adhesion molecules* (GO:0007156) resultaron significativos en ambos análisis, mientras que muchos otros fueron significativos sólo en uno de los dos escenarios evaluados (todos los genes o sólo los genes expresados). Esto puede ser un buen indicador de las diferencias fundamentales entre los dos enfoques aplicados.

Todos los programas y resultados detallados están disponibles en el Apéndice 1.

3.4.3.3. Comparación de resultados funcionales

Existen diversos procedimientos para la interpretación funcional de estudios de miARNs, aunque algunos de ellos no son directamente comparables con la metodología que hemos propuesto, bien porque se siguió un método de dos pasos (AS) o bien porque la propuesta metodológica incorpora variaciones en su estructura.

Una de las estrategias presentada por Godard & Eyll (2015) anota a los miARNs, la funcionalidad de sus genes-diana y a continuación aplica una prueba hipergeométrica estándar para los miARNs expresados diferencialmente. Para comparar nuestra metodología con esta propuesta, se utilizaron

modelos de regresión logística con el paquete *mdgsa* (Montaner & Dopazo 2010). Esta estrategia retiene las características de interés de la metodología propuesta por Godard, al tiempo que incorpora los beneficios del enfoque GSA sobre los AS.

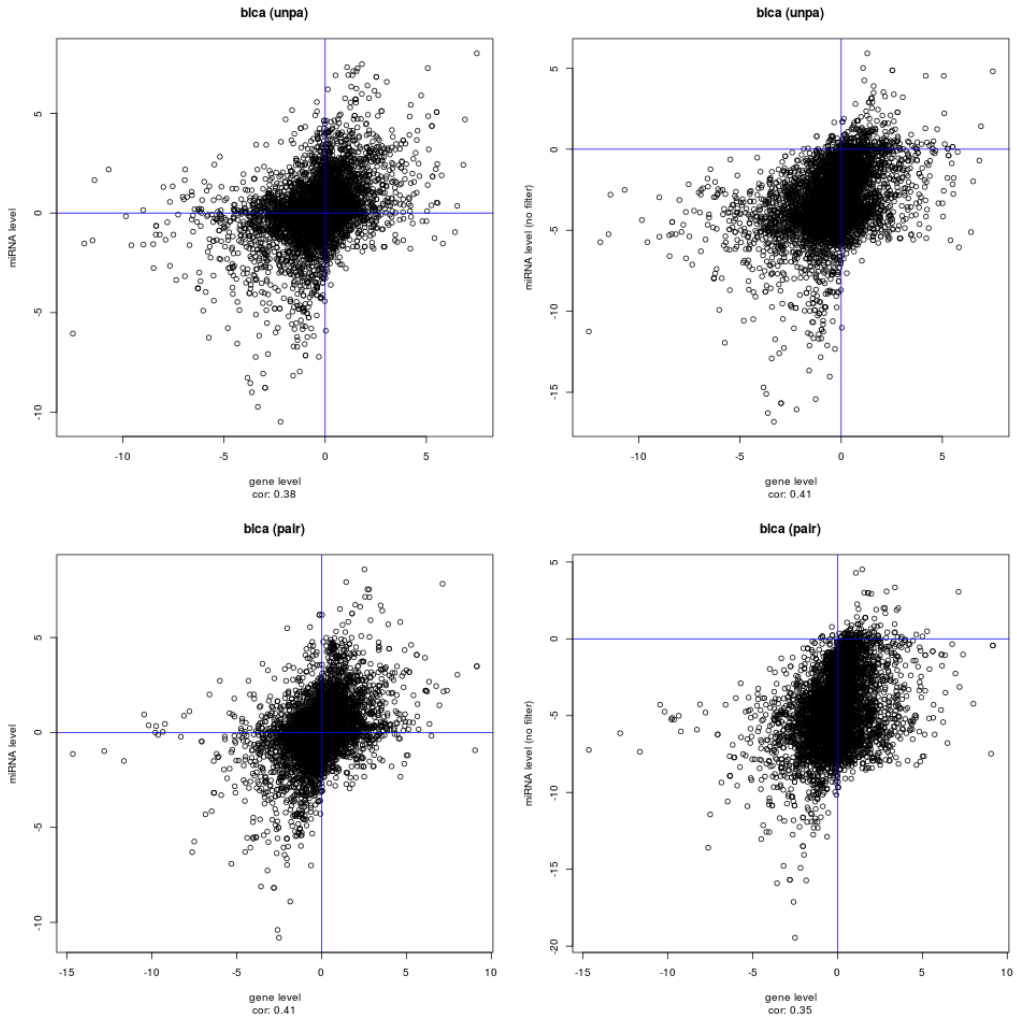

Figura 3.4: Comparación de resultados funcionales para BLCA

Los resultados funcionales obtenidos en este análisis a nivel de miARN (generalización de Godard) y los que están en el nivel de genes (después de la transferencia propuesta originalmente) presentan una correlación significativamente positiva. Esto indica que, en general ambas metodologías deben

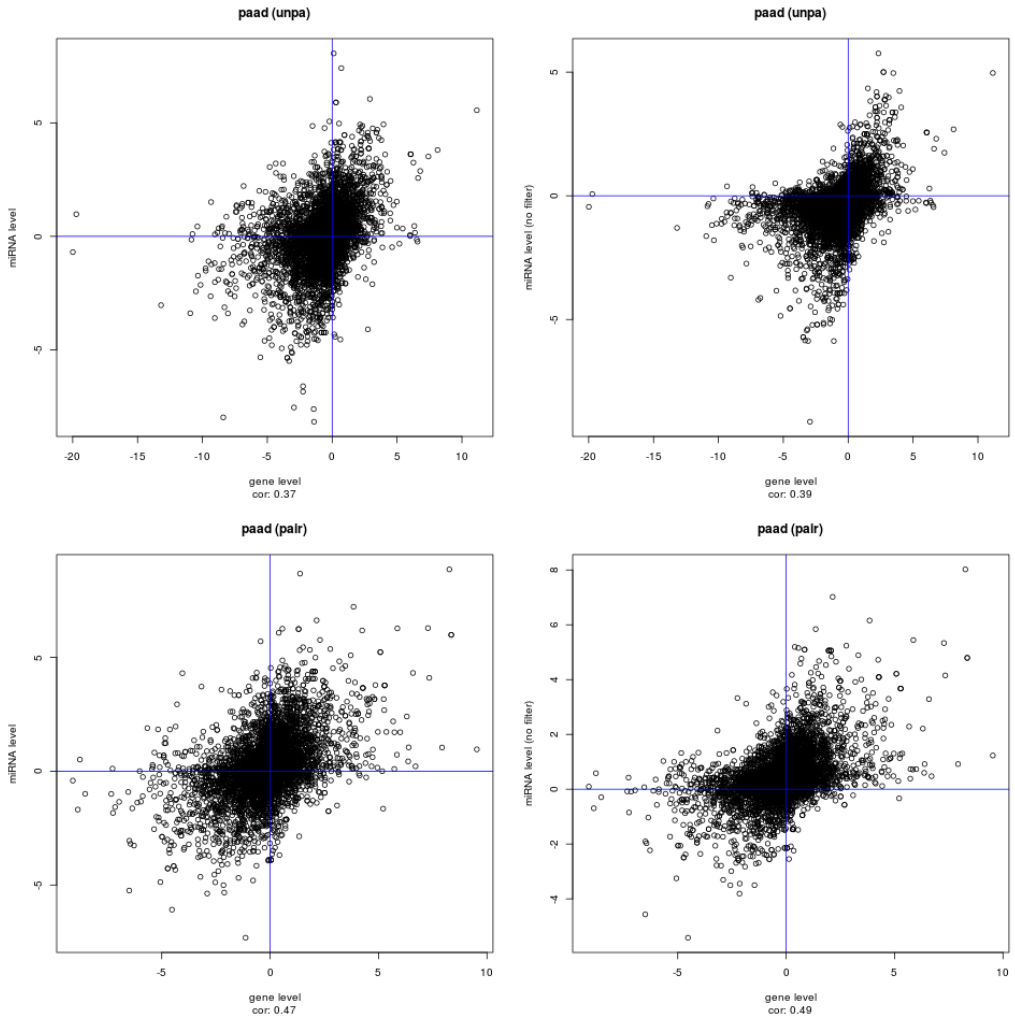

Figura 3.5: Comparación de resultados funcionales para PAAD

proporcionar resultados similares, aún obteniendo magnitudes de correlación no muy fuertes, debido fundamentalmente a que estos procedimientos son explícitamente diferentes.

Para cada tipo de cáncer, los resultados se han representado mediante gráficos de dispersión donde cada función está representada por un punto, mostrando la correlación entre los resultados funcionales a nivel de gen después de la *transferencia* y los resultados del análisis GSA llevado a cabo a nivel de miARN (paradigma de Godard), respectivamente indicados en los ejes X e Y de cada gráfico.

En las Figuras 3.4 y 3.5 se presentan los resultados de la evaluación de esta comparación de métodos para los tumores BLCA y PAAD. Para el resto de tumores se ha obtenido un patrón similar al descrito anteriormente (ver Apéndice 1).

3.4.3.4. Similitud funcional en grupos de tumores

Para conocer si existen similitudes funcionales entre los tumores estudiados, se realizó un análisis de conglomerados a partir de los resultados de los respectivos análisis de miARNs.

Se siguió el siguiente procedimiento:

1. Se generó un indicador para cada término funcional según la expresión 3.4, utilizando el paquete *mdgsa* (Montaner & Dopazo 2010):

$$indicador = signo\ (log(OR)) \cdot\ -1 \cdot\ log(valor\ p) \qquad (3.4)$$

2. Inicialmente se exploraron las relaciones entre grupos de tumores mediante análisis de componentes principales.
3. Para detectar similitudes funcionales se aplicó un procedimiento de clustering jerárquico, método completo y la distancia para valorar la proximidad de los grupos fue el coeficiente de correlación de Pearson. A continuación se determinaron los valores de probabilidad (valores de p) para cada agrupación, utilizando técnicas bootstrap de remuestreo que permitieron comprobar la robustez de los grupos detectados. Se obtuvieron dos tipos de indicadores: AU (*Approximately Unbiased*) y BPR (*Bootstrap Probability*). El primero de ellos fue calculado por remuestreo *bootstrap multiscale* y el segundo se determinó mediante remuestreo *bootstrap normal*. Estos métodos están implementados en el paquete *pvclust* (Suzuki & Shimodaira 2006) de *R*.

Los análisis de componentes principales se realizaron de forma separada para los estudios con muestras pareadas y no pareadas. Además en cada uno de estos dos grupos se evaluaron las tres ontologías de la Gene Ontology referidas en el análisis de enriquecimiento funcional: procesos biológicos (PB), funciones moleculares (FM) y componentes celulares (CC). Los resultados fueron representados gráficamente para sus tres primeras componentes.

Las Figuras 3.6, 3.7 y 3.8 corresponden a los estudios pareados y las Figuras 3.9, 3.10 y 3.11 a los estudios no pareados. Su exploración permitió detectar grupos de tumores que mantienen un comportamiento funcional similar como los tumores KIRP, KICH y KIRC (variantes de cáncer de riñón). También quedan agrupados los tumores de pulmón: LUSC y LUAD así como los tumores KICH (riñón) y THCA (tiroides), habiendo asociaciones descritas

entre ambos tipos de cáncer (Oh et al. 2015). No se detectó ningún estudio que presentara un diferente comportamiento del resto de tumores en las distintas ontologías y escenarios con muestras pareadas y no pareadas. Con el análisis de clustering se confirman estas relaciones y además se determinaron indicadores de robustez de las agrupaciones detectadas mediante técnicas de remuestreo.

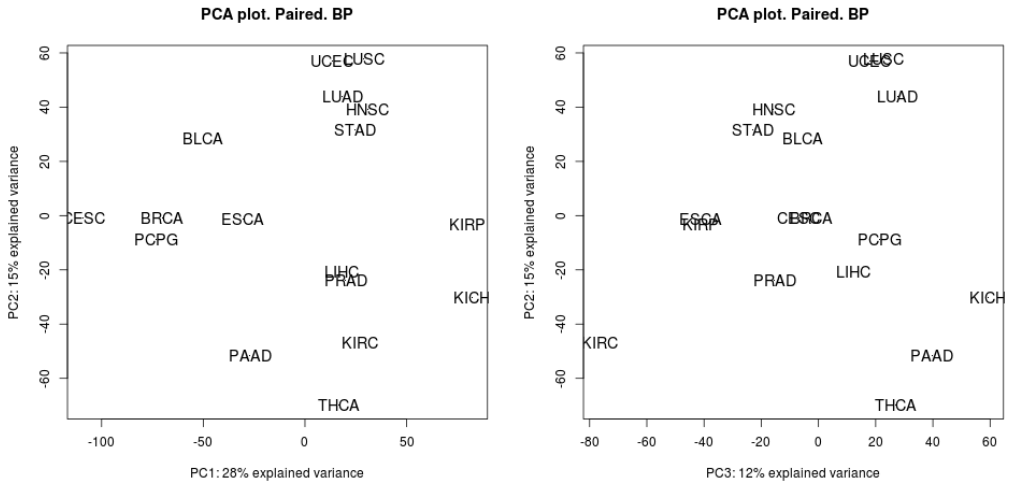

Figura 3.6: Análisis de componentes principales de los resultados del enriquecimiento funcional (procesos biológicos). Estudios pareados.

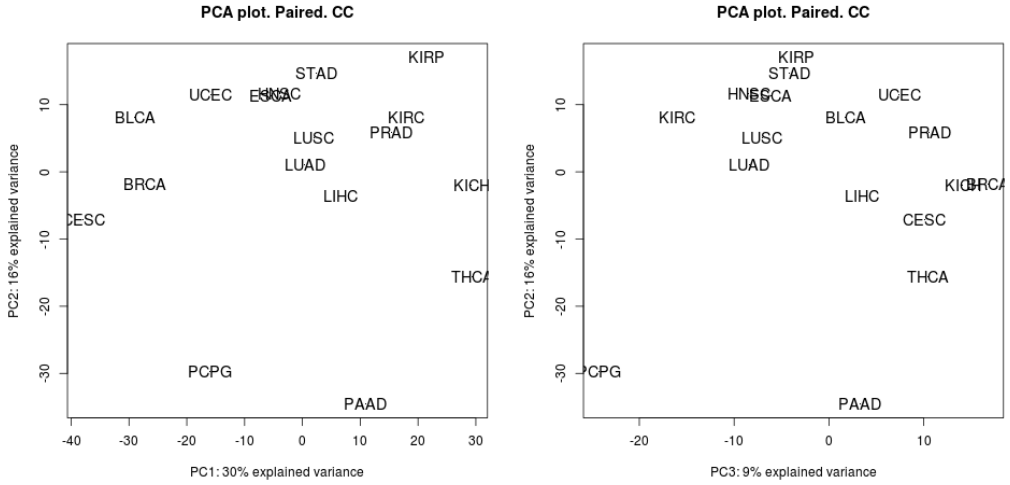

Figura 3.7: Análisis de componentes principales de los resultados del enriquecimiento funcional (componentes celulares). Estudios pareados.

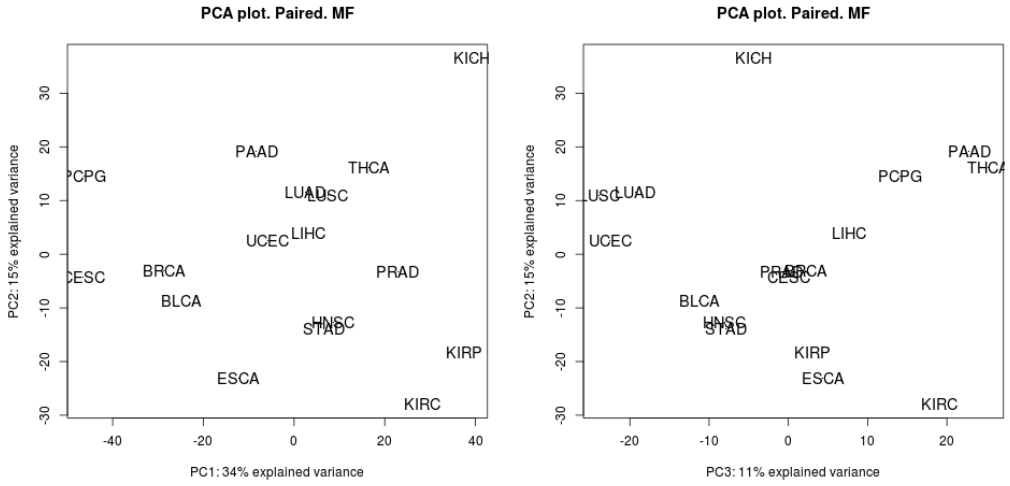

Figura 3.8: Análisis de componentes principales de los resultados del enriquecimiento funcional (funciones moleculares). Estudios pareados.

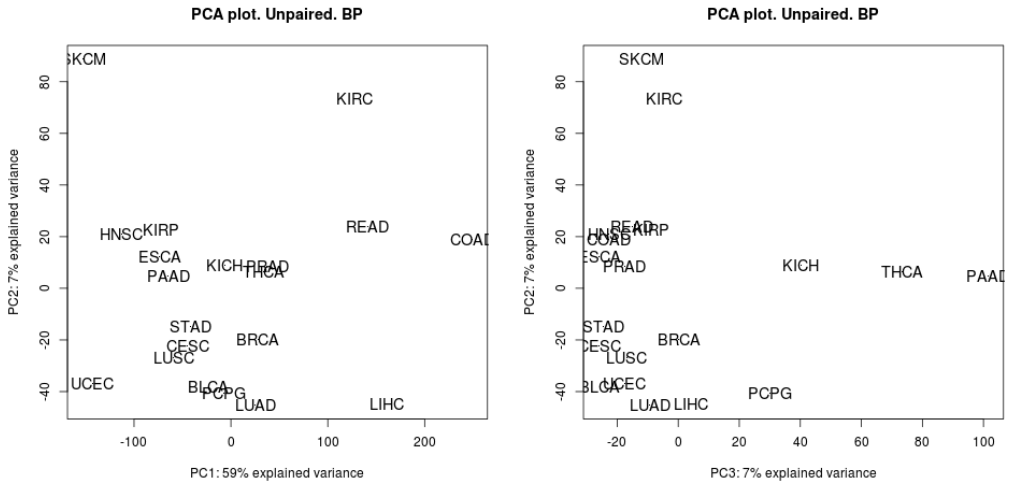

Figura 3.9: Análisis de componentes principales de los resultados del enriquecimiento funcional (procesos biológicos). Estudios no pareados.

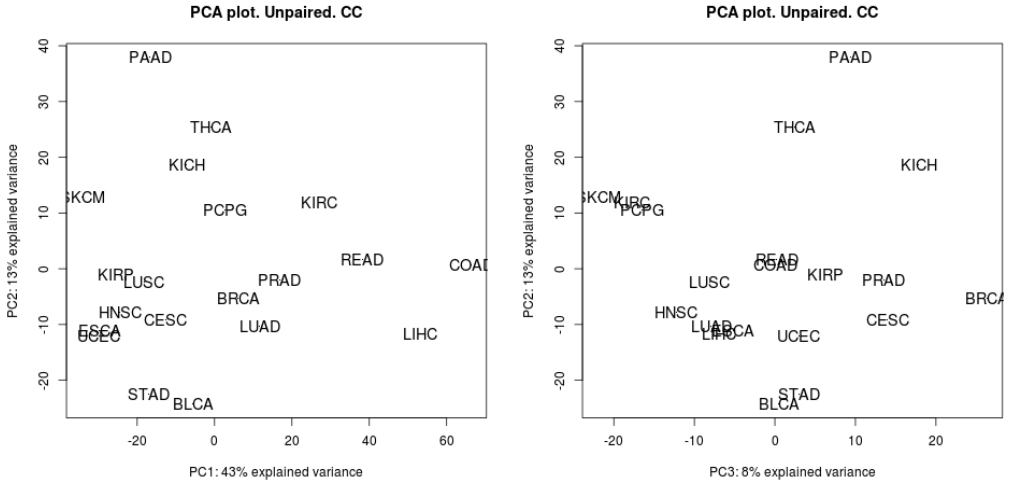

Figura 3.10: Análisis de componentes principales de los resultados del enriquecimiento funcional (componentes celulares). Estudios no pareados.

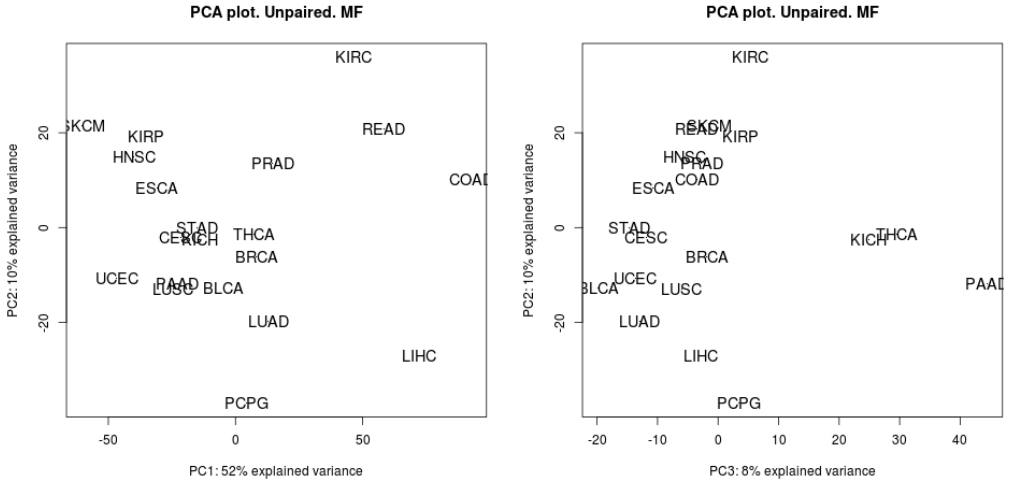

Figura 3.11: Análisis de componentes principales de los resultados del enriquecimiento funcional (funciones moleculares). Estudios no pareados.

Los análisis de clustering también se realizaron independientemente para estudios con muestras pareadas y no pareadas. Además en cada uno de estos dos grupos se evaluaron las tres ontologías de la Gene Ontology utilizadas en el análisis de enriquecimiento funcional. Las Figuras 3.12 a 3.17 muestran las agrupaciones detectadas. Cada una de ellas incluye los indicadores obtenidos por técnicas de remuestreo para comprobar el nivel de robustez del clúster: valores en color rojo corresponden a AU y en verde a los obtenidos por BPR.

Se detectó similitud funcional en los siguientes grupos:

1. Agrupación de tumores por órganos.

    - Para los *estudios con muestras pareadas* los resultados de la agrupación se muestran en las Figuras 3.12, 3.13 y 3.14. En cualquiera de las ontologías, los tumores que afectan a un mismo órgano quedan agrupados. De este modo, los cánceres KICH, KIRC, KIRP que son de riñón con diferentes variantes, quedan próximos en los árboles de clustering. Lo mismo ocurre con LUAD y LUSC, que son tumores de pulmón. Según la ontología, el nivel de significación de estas agrupaciones varía, así por ejemplo los perfiles funcionales de los tumores KIRC y KIRP son claramente significativos y por lo tanto similares funcionalmente, cuando la evaluación se realiza en las funciones moleculares y los procesos biológicos, aunque no ocurre los mismo con los componentes celulares. Por otra parte, los tumores de pulmón LUAD y LUSC son funcionalmente similares en cualquiera de las agrupaciones realizadas por ontología.
    - Para los *estudios con muestras no pareadas*, las Figuras 3.15, 3.16

y 3.17 muestran los resultados de las agrupaciones. No se mantiene el patrón de agrupación por órgano descrito anteriormente en los estudios donde el diseño sí incluyó muestras pareadas.

2. Agrupación de tumores por funcionalidad del órgano afectado. Se comprobó si los tumores de órganos que participan en funciones comunes, estaban agrupados a partir de los resultados del enriquecimiento. Es decir, si los tumores de órganos del aparato digestivo como son COAD (colon), ESCA (esófago), STAD (estómago), READ (recto), PAAD (páncreas), LIHC (hígado) se diferenciaban del resto. Análogamente para tumores asociados al aparato urogenital: KICH, KIRC, KIRP (riñón), PRAD (próstata), UCEC (endometrio), CESC (cuello uterino) o bien otros tumores que apuntaran a órganos comunes en algunos aparatos "funcionales". En general, estos tumores no se presentan en bloque de una forma completamente diferenciada del resto, aunque sí hay subgrupos con una cooperativa funcionalidad de los órganos afectados:

- En los estudios *pareados*, se aprecia que los CC y las MF, agrupan bien los tumores urogenitales de riñón (KIRC, KIRP) y próstata (PRAD) frente el resto. Los tumores KICH (riñón) y THCA (tiroides) también se agrupan en cada una de las ontologías, habiendo asociaciones descritas entre ambos tipos de cáncer (Oh et al. 2015). En todas las ontologías, se detectan similitudes funcionales entre los tumores CESC (cuello uterino) y BRCA (mama).
- Para los estudios *no pareados*, los tumores COAD (colon) y READ (recto) forman un conglomerado "digestivo" en cualquiera de las agru-

paciones de las diferentes ontologías. Se repite de nuevo el patrón de similitud entre los tumores KICH (riñón) y THCA (tiroides) en todas las ontologías.

En definitiva, el estudio de la similitud de los patrones funcionales confirma relaciones descritas entre tumores y revela otras relaciones cuyo estudio puede ser de interés para una mejor comprensión del funcionamiento de los diferentes tipos de cáncer.

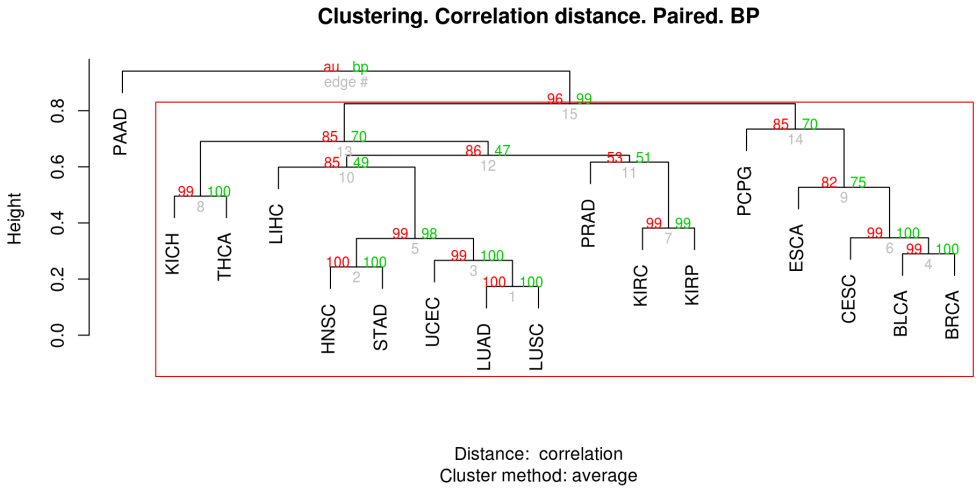

Figura 3.12: Análisis de clustering de resultados del enriquecimiento funcional (procesos biológicos). Estudios pareados.

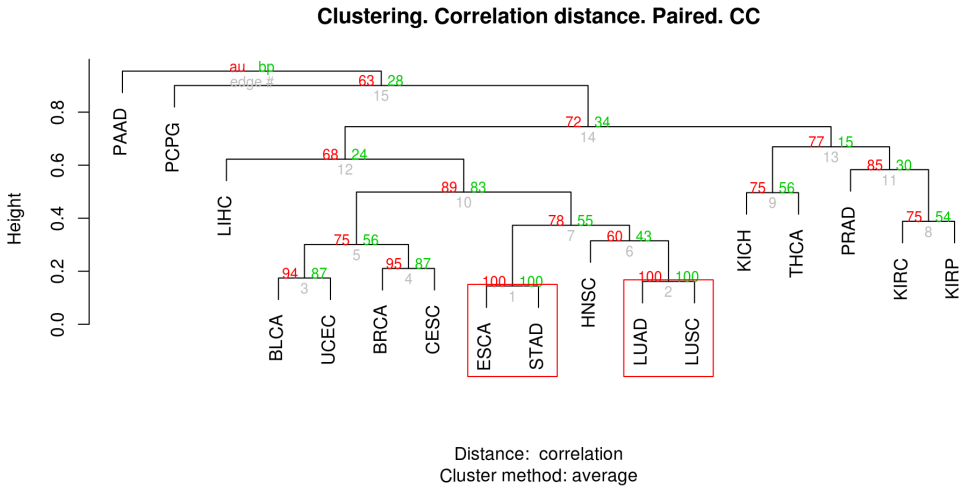

Figura 3.13: Análisis de clustering de resultados del enriquecimiento funcional (componentes celulares). Estudios pareados.

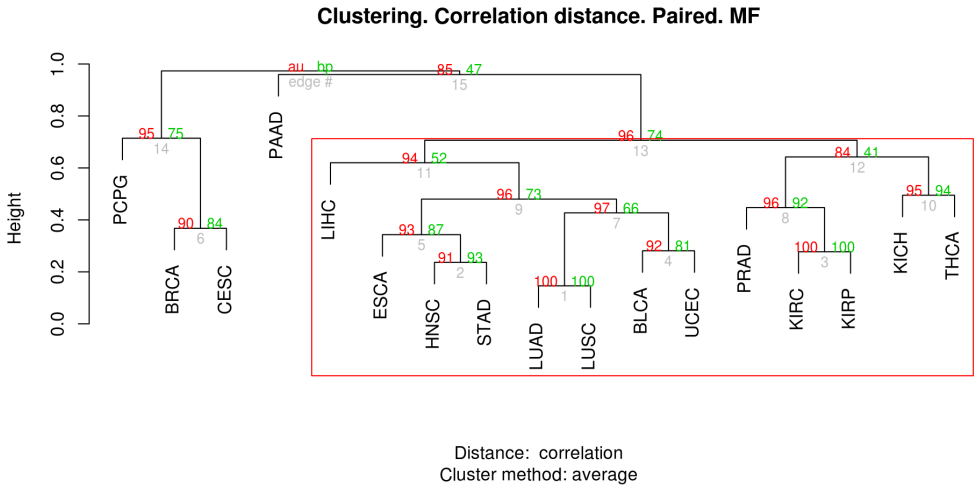

Figura 3.14: Análisis de clustering de resultados del enriquecimiento funcional (funciones moleculares). Estudios pareados.

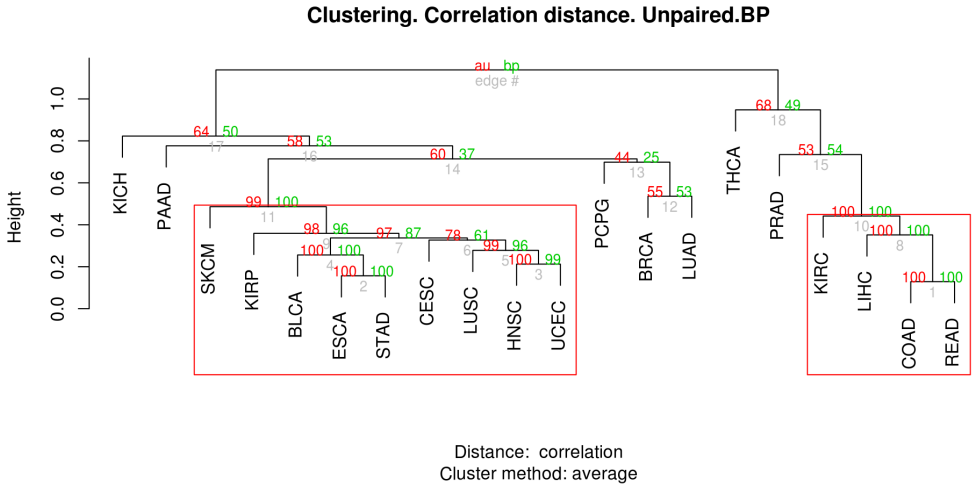

Figura 3.15: Análisis de clustering de resultados del enriquecimiento funcional (procesos biológicos). Estudios no pareados.

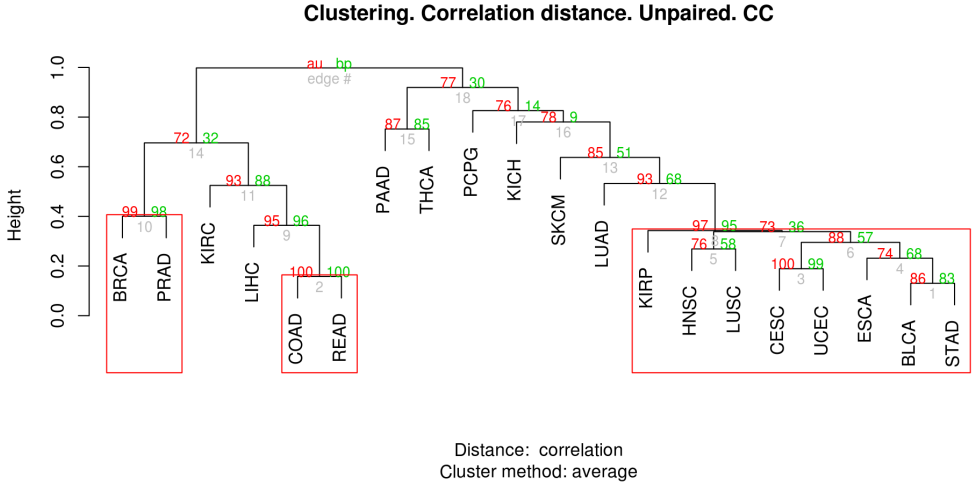

Figura 3.16: Análisis de clustering de resultados del enriquecimiento funcional (componentes celulares). Estudios no pareados.

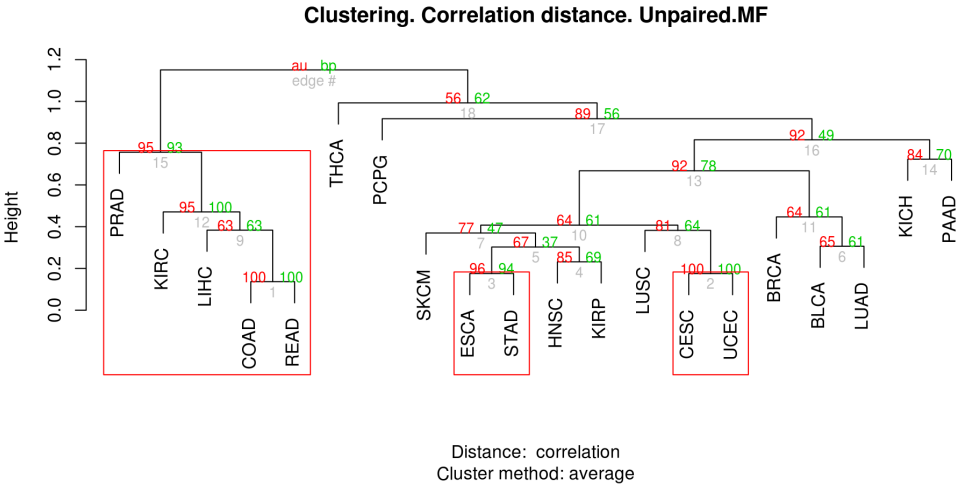

Figura 3.17: Análisis de clustering de resultados del enriquecimiento funcional (funciones moleculares). Estudios no pareados.

3.5. Discusión

Hemos introducido un nuevo enfoque para la interpretación funcional de los estudios de miARN que está diseñado principalmente para conocer los efectos de la expresión diferencial de miARN en grupos de genes.

Nuestra propuesta se basa en el paradigma del análisis de grupos de genes, el cual extiende las metologías de los análisis de sobrerrepresentación. Constituye un marco general aplicable a la mayoría de los escenarios genómicos, incluso cuando no hay (o hay demasiados) miARNs que se expresan diferencialmente, por lo tanto, este algoritmo elimina la arbitrariedad de los procedimientos *ad hoc* actuales. Pero lo más importante, es que nuestro algoritmo puede abarcar acontecimientos biológicamente relevantes que son abandonados por los demás, lo que representa un paso adelante en la modelización de la regulación miARN-gen.

En primer lugar, nuestro enfoque considera los efectos de cancelación que

surgen cuando un gen es interceptado por diferentes grupos de miARNs dentro de cada condición biológica. En segundo lugar, es capaz de incorporar el efecto aditivo causado, cuando varios inhibidores de miARNs débiles ejercen su influencia en el mismo gen.

Estas importantes ventajas son posibles gracias a una innovadora idea introducida en este trabajo: que la expresión diferencial de los miARNs puede ser transferida a nivel del gen como un indicador de inhibición diferencial.

Si la transferencia de miARN a gen incorpora la cancelación y la suma de efectos, la metodología de grupos de genes realiza el mismo papel en el nivel funcional. Un término GO no es considerado enriquecido o cancelado, si la mitad de sus genes son inhibidos en los casos y la otra mitad en los controles. Pero también la consideración del efecto aditivo reaparece a nivel de función: muchos genes desregulados o inhibidos débilmente, los cuales serían relevantes cuando están aislados, adquieren importancia si se anotan sistemáticamente bajo la misma función biológica.

Además del análisis aquí presentado, la metodología de regresión logística desarrollada en trabajos anteriores (Montaner & Dopazo 2010) permite que el algoritmo se pueda extender de diversas formas con gran interés. Por ejemplo, la importancia relativa de miARNs, genes, o la relación miARN-gen puede ser fácilmente ponderada durante la etapa de transferencia o en el ajuste del modelo logístico. Por lo tanto, la fiabilidad de las dianas de los miARN, los niveles absolutos de expresión génica o incluso la natural pérdida funcional de los miARNs (Carbonell & others 2012), pueden ser directamente valoradas por nuestro modelo.

Por otra parte, la información genómica adicional se puede incorporar en

el uso de nuestro marco multidimensional. Por ejemplo, la integración de análisis de GSA de regulación de miARN y expresión génica es sencilla una vez que el problema de la transferencia se resuelve utilizando la metodología explicada anteriormente. También la flexibilidad de nuestro enfoque y su implementación en un software de fácil utilización, hace su uso independiente del algoritmo de diferencial expresión utilizado a nivel de miARN. Diferentes procedimientos estadísticos (*NOISeq*, Tarazona et al. 2012; *DESeq*, Anders & Huber 2012) o incluso indicadores como el *fold change* pueden sustituir el método propuesto en *edgeR* y que se ha utilizado en el análisis. Asímismo se pueden utilizar diferentes bases de datos que incluyan la información de las dianas de los miARNs.

Hemos ilustrado nuestra metodología novedosa (García-García et al. 2016) utilizando una extensa serie de estudios con datos de cáncer, pero aquí simplemente se presentan algunos genes o funciones desreguladas como una prueba de concepto. Los resultados completos están disponibles en los datos suplementarios (ver Apéndice 1). Las ideas introducidas pueden ser fácilmente extrapolables a otros procesos de regulación de genes como los relacionados con los factores de transcripción.

Capítulo 4

Metaanálisis funcional en estudios genómicos

4.1. Introducción

4.1.1. Introducción al metaanálisis

Las *revisiones sistemáticas* y los *metaanálisis* se han consolidado como herramientas metodológicas que ofrecen información con elevado nivel de calidad y rigor científico (González et al. 2011; Catalá-López & Tobías 2013). La revisión sistemática integra toda la información empírica disponible sobre un tema específico de interés o una pregunta de investigación. El metaanálisis es la metodología estadística que permite combinar resultados sobre los efectos procedentes de diversos estudios individuales que previamente han sido identificados y valorados críticamente desde la revisión sistemática.

Los procedimientos de metaanálisis fueron inicialmente aplicados en las ciencias sociales y en psicología (Glass 1976). A partir de la década de los 80, se comenzó a aplicar de forma creciente en medicina y a partir de los 90 son frecuentes los artículos que describen resultados de metaanálisis en publicaciones médicas (Normand 1999). Posteriormente han sido utilizados en diferentes áreas científicas, incluyendo la Genómica.

El metaanálisis permite obtener una medida combinada del efecto de interés con una mayor precisión que la ofrecida por los estudios individuales que han sido detectados en una revisión sistemática previa, y por lo tanto ofrece una mayor potencia estadística, cuantificando la variabilidad de los estudios individuales. Sin embargo, una inadecuada selección de los métodos estadísticos e interpretación de resultados, puede suponer una limitación en su uso e interpretación (Catalá-López & Tobías 2014).

4.1.2. Metaanálisis de datos genómicos

La aparición de las tecnologías de alto rendimiento (microarrays y secuenciación masiva) ha motivado la generación de una gran cantidad de datos genómicos. La comunidad científica ha demandado la creación de diversos repositorios que permitan la organización sistemática y el acceso libre de estos datos procedentes de diferentes estudios. Entre ellos: *Gene Expression Omnibus* (*GEO*, http:// www. ncbi. nlm. nih. gov/ geo) (Barrett et al. 2007), *ArrayExpress* (http:// www. ebi. ac. uk/ microarray-as/ ae) (Parkinson et al. 2005) y *Sequence Read Archive* (*SRA*, http://www.ncbi.nlm.nih.gov/sra) (Kodama et al. 2012). También algunas instituciones ofrecen sus datos a los investigadores: *The Cancer Genome Atlas* (McCain 2006) y *European Genome-phenome Archive* (*EGA*, https://www.ebi.ac.uk/ega/home) (Lappalainen et al. 2015).

El uso de estos conjuntos de datos proporciona una importante fuente de información en investigación genómica, ofreciendo datos adicionales para los análisis y evaluación de metodologías, verificación y reproducibilidad de resultados. Aunque el coste de estas tecnologías ha ido decreciendo, todavía sigue siendo una limitación su uso generalizado. Por ello, la mayoría de los experimentos realizados, incluyen un reducido número de muestras biológicas y por lo tanto los análisis tienden a carecer de poder de detección. La aplicación de técnicas de metaanálisis constituye un abordaje global e integrador en Genómica.

Las primeras aplicaciones de metaanálisis de datos genómicos procedentes de microarrays se produjeron en estudios de expresión génica (Rhodes et al. 2002; Grützmann et al. 2005) y en estudios de asociación entre polimorfismos

de un solo nucleótido (*SNPs*) y enfermedad (Zeggini et al. 2008; Evangelou & Ioannidis 2013). Existen diversas revisiones sobre las ventajas y limitaciones de los diferentes métodos (Choi et al. 2003; Hong & Breitling 2008; Ramasamy et al. 2008; Campain & Yang 2010) que con la aparición de la secuenciación masiva han sido revisados y reajustados (Rau et al. 2014).

La mayoría de los métodos de metaánalisis de estudios genómicos, se centran en un nivel de gen o variante. Sin embargo, el uso de procedimientos de metaanálisis en un nivel de función, amplía el número de escenarios donde interpretar los resultados de estudios genómicos procedentes de diferentes tecnologías, proporcionando una mejor comprensión biológica y clínica.

4.1.3. Metaanálisis funcional de datos genómicos

En numerosos experimentos que utilizan tecnologías de alto rendimiento, se han realizado un gran esfuerzo para detectar grupos de genes con funciones comunes. La identificación de estos grupos suele ser inconsistente entre estudios independientes, debido a varios factores como el ruido de la tecnología utilizada, las características del experimento y el reducido tamaño muestral. Por lo tanto, la combinación de información de múltiples estudios puede mejorar la detección de verdaderas clases de genes enriquecidas.

Existen pocos métodos orientados al metaanálisis de estudios genómicos a nivel de función. Shen & Tseng (2010) desarrollaron y evaluaron un metaanálisis de enriquecimiento de rutas de señalización con datos de microarrays. Propusieron tres procedimientos basados en el popular método *GSEA* (Subramanian et al. 2005) para análisis de enriquecimiento y que denominaron:

MAPE G (metaánalisis de enriquecimiento de rutas a nivel de gen), *MAPE P* (metaánalisis de enriquecimiento de rutas a nivel de ruta) y *MAPE I* (metaánalisis de enriquecimiento de rutas integrando niveles gen y ruta). En todos ellos se distinguen dos fases: en la primera se realiza el análisis de expresión diferencial, obteniéndose estadísticos a nivel de gen para cada estudio individual. En la segunda etapa se aborda el análisis de enriquecimiento y el metaanálisis en diferentes órdenes: con *MAPE G* se realiza primero el metaanálisis de los resultados de la expresión diferencial a nivel de gen y a continuación se determina el análisis de enriquecimiento; con *MAPE P* se obtiene un análisis de enriquecimiento para cada estudio individual y entonces combina los resultados a nivel de grupo de genes. Con el tercer método, *MAPE I*, se combinan los resultados finales de los dos procedimientos descritos para mejorar la potencia del abordaje metodológico.

Posteriormente, Chen et al. (2013) presentaron una metodología de metaanálisis con un enfoque bayesiano que permitía de forma flexible, la inclusión de diversas fuentes de variabilidad procedentes de los estudios individuales. Esta propuesta proporciona una modelización conjunta de los datos de expresión de los distintos estudios y la información de los grupos de genes.

En el primero de estos procedimientos hay una clara orientación al uso de rutas de señalización en datos de microarrays. En el segundo enfoque no se muestra esta limitación en la anotación funcional, aunque se mantiene el marco de aplicación a los datos de expresión génica. La metodología que presentamos sobre metaanálisis funcional, permite el uso de cualquier información funcional procedente de una base de datos conocida o una anotación propia elaborada por los investigadores. Por otra parte, es aplicable tanto a

datos de expresión génica de microarrays como secuenciación masiva e incluso a otro tipo de unidad como miARNs. También es posible aplicar nuestro metaanálisis funcional en otros escenarios ómicos, en los que disponemos estudios metabolómicos o proteómicos donde estamos interesados en conocer funcionalidades comunes de interés. En definitiva, este nuevo enfoque posibilita un flexible abordaje funcional que proporciona a los investigadores una mejor interpretación de los resultados de estudios genómicos.

4.2. Datos

Para la evaluación del método que proponemos, se utilizaron datos procedentes de dos grupos de estudios diferentes:

4.2.1. Estudios de expresión génica en enfermedades dermatológicas: psoriasis y dermatitis

La psoriasis y la dermatitis son dos de las enfermedades dermatológicas más comunes. Presentan un impacto considerable en la calidad de la vida de las personas afectadas (Rapp et al. 1999) y su alta prevalencia supone un importante problema de Salud Pública, con su consecuente coste económico en el sistema sanitario (Javitz et al. 2002).

La detección de patrones comunes en ambas enfermedades, como es la interacción entre la base génica, la base inmunológica y el entorno para el desarrollo de las mismas, aportaría información relevante para diseñar abordajes clínicos más precisos. Sin embargo, debido a que presentan una base

inmunológica heterogénea, (Bieber 2008; DaVeiga 2012) el conocimiento de estas enfermedades es todavía incompleto y por ello actualmente no existe un tratamiento definitivo (Mu et al. 2014).

Un mejor conocimiento de la base genética de ambas enfermedades proporcionará una mayor comprensión de las mismas y el descubrimiento de los mecanismos biológicos implicados en ellas. Esta información se podría aplicar en una asignación más específica de los diagnósticos y en el desarrollo de herramientas de pronóstico, así como en el diseño de tratamientos individualizados más efectivos y con menor impacto para el paciente, dentro del marco de la *Medicina Personalizada*.

Desde el repositorio *GEO* se seleccionaron estudios de microarrays referentes a psoriasis y dermatitis (series). Los criterios de inclusión inicial fueron los siguientes:

- Estudios realizados en humanos.
- Diseño experimental caso-control, donde los casos presentan piel lesionada y los controles refieren piel sin lesiones.

Se obtuvieron un total de 65 series para psoriasis y 36 para dermatitis. A partir de esta información inicial se refinó la búsqueda, incluyendo únicamente aquellos estudios que contuvieran muestras sin tratamiento médico o alérgeno que pudiera alterar la expresión génica, y que además presentaran un formato de identificador de sonda con una equivalencia para los identificadores de *Ensembl* (Flicek et al. 2014). *Ensembl* es un proyecto de bases de datos de genomas y entre sus funcionalidades se encuentra la posibilidad de descargar tablas de correspondencias de identificadores. De este modo, es

posible seleccionar un organismo y características determinadas, y *Ensembl* proporciona relaciones entre términos GO, identificadores de dominios de proteínas, identificadores de sondas de multitud de microarrays, etc.

La importancia de una revisión sistemática con definidos criterios de interés es fundamental en este tipo de abordajes. El conjunto final de estudios seleccionados incluyó 26 series: 17 correspondendientes a psoriasis, 6 a dermatitis atópica, 1 a dermatitis de contacto y 2 incluyen casos de psoriasis y dermatitis atópica. Las características de cada uno de los estudios pueden observarse en la Tabla 4.1.

En cada uno de estos estudios se incluyeron comparaciones entre diferentes grupos experimentales (no siempre disponemos de una única comparación entre enfermedad frente control, sino que pueden presentarse varios grupos de enfermedad). Por ello, a partir de la selección de las 26 series descritas anteriormente, se evaluaron 41 comparaciones finales entre grupos enfermo y control. De ellas, 18 correspondían a dermatitis atópica y 23 a psoriasis.

4.2.2. Estudios de miARNs en tumores

En este segundo grupo de estudios, se ha trabajado con datos procedentes del proyecto *The Cancer Genome Atlas*, que ya se utilizaron en el capítulo anterior en los estudios de enriquecimiento de miARNs. De modo que obtendremos diferentes interpretaciones de los resultados obtenidos de los mismos datos, empleando distintos abordajes funcionales: en el método descrito en el capítulo anterior, la interpretación funcional se realiza sobre cada uno de los estudios de miARNs y con esta segunda propuesta metodológica, la

Estudio	Enfermedad	Plataforma y tipo de *microarray*
GSE2737	PS	*Affymetrix Human Genome U95A/U95 Version 2*
GSE6710	PS	*Affymetrix Human Genome U133A*
GSE52471	PS	*Affymetrix Human Genome U133A 2.0*
GSE26866	PS	*Affymetrix Human Genome U133A 2.0*
GSE11903	PS	*Affymetrix Human Genome U133A 2.0*
GSE30768	PS	*Affymetrix Human Genome U133A 2.0*
GSE32407	PS	*Affymetrix Human Genome U133A 2.0*
GSE14905	PS	*Affymetrix Human Genome U133 Plus*
GSE41662	PS	*Affymetrix Human Genome U133 Plus 2.0*
GSE41663	PS	*Affymetrix Human Genome U133 Plus 2.0*
GSE34248	PS	*Affymetrix Human Genome U133 Plus 2.0*
GSE13355	PS	*Affymetrix Human Genome U133 Plus 2.0*
GSE30999	PS	*Affymetrix Human Genome U133 Plus 2.0*
GSE40263	PS	*Affymetrix Human Gene 1.0 ST*
GSE31835	PS	*Illumina HumanMethylation27 BeadChip*
GSE18686	PS	*Illumina HumanHT-12 V3.0 expression beadchip*
GSE53431	PS	*Illumina HumanHT-12 V4.0 expression beadchip*
GSE53431	PS	*Illumina HumanHT-12 V4.0 expression beadchip*
GSE16161	PS/DA	*Affymetrix Human Genome U133/U133A Plus 2.0*
GSE26952	PS/DA	*Sentrix HumanRef-8 Expression BeadChip*
GSE5667	DA	*Affymetrix Human Genome U133A/B*
GSE6012	DA	*Affymetrix Human Genome U133A*
GSE27887	DA	*Affymetrix Human Genome U133 Plus 2.0*
GSE32924	DA	*Affymetrix Human Genome U133 Plus 2.0*
GSE36842	DA	*Affymetrix Human Genome U133 Plus 2.0*
GSE12511	DA	*Print_730*
GSE6281	DC	*Affymetrix Human Genome U133 Plus 2.0*

Tabla 4.1: Conjunto de estudios de piel seleccionados de Gene Expression Omnibus. Las enfermedades corresponden a psoriasis (PS), dermatitis atópica (DA) y dermatitis de contacto (DC).

evaluación funcional se aplica sobre el conjunto de todos los estudios.

Se descargaron y analizaron 20 de estos estudios, habiendo seleccionado aquellos que incluían información de expresión de miARN medida con tecnología *Illumina HiSeq* (Bentley & others 2008), y que contienen tanto muestras tumorales como sanas. La Tabla 4.2 muestra los datos descargados de los diferentes estudios y el número de muestras incluidas en cada uno de ellos.

ID	total	casos	controles	casos-controles pareados	descripción
BLCA	271	252	19	19	Bladder Urothelial Carcinoma
BRCA	807	720	87	86	Breast invasive carcinoma
CESC	218	215	3	3	Cervical squamous cell carcinoma
COAD	243	235	8	0	Colon adenocarcinoma
ESCA	113	102	11	11	Esophageal carcinoma
HNSC	519	475	44	43	Head and Neck squamous cell carcinoma
KICH	91	66	25	25	Kidney Chromophobe
KIRC	311	240	71	68	Kidney renal clear cell carcinoma
KIRP	245	211	34	34	Kidney renal papillary cell carcinoma
LIHC	283	233	50	49	Liver hepatocellular carcinoma
LUAD	474	428	46	39	Lung adenocarcinoma
LUSC	376	331	45	45	Lung squamous cell carcinoma
PAAD	100	96	4	4	Pancreatic adenocarcinoma
PCPG	182	179	3	3	Pheochromocytoma and Paraganglioma
PRAD	117	100	17	17	Prostate adenocarcinoma
READ	93	90	3	0	Rectum adenocarcinoma
SKCM	75	74	1	0	Skin Cutaneous Melanoma
STAD	345	306	39	39	Stomach adenocarcinoma
THCA	558	499	59	59	Thyroid carcinoma
UCEC	418	386	32	19	Uterine Corpus Endometrial Carcinoma

Tabla 4.2: Conjunto de estudios de tumores seleccionados de TCGA (The Cancer Genome Atlas). Las columnas indican: identificador de enfermedad en TCGA, número total de muestras en el análisis, número de muestras tumorales, número de muestras control (tejido normal sólido), número de muestras pareadas disponibles en el el grupo de datos y tipo de cáncer.

4.3. Métodos

Los métodos presentados en metaanálisis funcional están enmarcados en la secuencia de pasos de la siguiente estrategia de análisis:

1. Revisión sistemática y selección de estudios.

2. Análisis primario:

- Procesamiento de los datos.
- Análisis de expresión diferencial.
- Análisis de enriquecimiento de grupos de genes.

3. Metaanálisis a nivel de función:

    - Configuración y exploración de matrices de entrada.
    - Análisis de heterogeneidad y determinación de la medida combinada del efecto.
    - Análisis de sensibilidad y evaluación de sesgos.
    - Representación e interpretación de resultados.

4.3.1. Revisión sistemática y selección de estudios

La validez de un metaanálisis depende en buena medida, de la identificación y selección de los estudios originales. La definición de criterios de inclusión y exclusión de los estudios, un diseño muestral adecuado y la valoración de la calidad de los estudios son elementos necesarios para la obtención de robustos e interpretables indicadores del efecto de interés. La forma de evaluar cada uno de estos aspectos se ha basado en las indicaciones consensuadas en *PRISMA* (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, http://www.prisma-statement.org).

Los datos de estos estudios se obtienen de diversas fuentes, siendo los repositorios de datos de acceso público como *GEO* o proyectos como *The Cancer Genome Atlas* (http://cancergenome.nih.gov) dos recursos de interés donde buscar estudios primarios que incluyan datos ómicos.

4.3.2. Análisis primario

4.3.2.1. Procesamiento de los datos

- En los estudios de las enfermedades de la piel, los datos normalizados de expresión proceden del repositorio *GEO*. El procesamiento incluyó la exploración de estas matrices de niveles de expresión, traducción de identificadores a una misma referencia (Ensembl Gene ID) y el promedio de valores en sondas repetidas, con el objeto de estandarizar y facilitar los análisis posteriores sobre el conjunto de estos estudios.

- Para los estudios de tumores, las matrices con los conteos normalizados de la expresión de miARN fueron descargadas del portal de datos *The Cancer Genome Atlas*. Tras su exploración, se completó el procesamiento de los datos para preparar los elementos de entrada de la siguiente fase del análisis.

4.3.2.2. Análisis de expresión diferencial

En ambos grupos de estudios se compararon la condición enfermedad frente el control:

- En las enfermedades dermatológicas, el nivel de expresión entre los grupos se evaluó con el uso de modelos lineales implementados en el paquete *limma* (Smyth 2005) de *Bioconductor*. Los contrastes enfermo-control se realizaron considerando los diferentes tipos de muestras de los estudios:

- *Lesionado* (piel lesionada en individuos enfermos) frente *control* (piel de individuos sanos).
- *No lesionado* (piel no lesionada en individuos enfermos) frente *control*.
- *Lesionado* frente *no lesionado*.

De modo que para un mismo estudio, es posible disponer de varios contrastes si están presentes los tres tipos descritos de muestras: *lesionado*, *no lesionado* y *control*.

Para cada comparación se obtuvieron los contrastes estadísticos y los valores de p, para la detección de la expresión diferencial a nivel de gen.

- Para los tumores, con el análisis de expresión diferencial se compararon las muestras de tumores primarios respecto el tejido normal sólido, utilizando un enfoque no pareado para los 20 conjuntos de datos. Además, también se realizó un análisis pareado para 17 de ellos: los estudios que contenían muestras tumorales y normales de la misma persona. Estos análisis a nivel de miARN fueron realizados utilizando el paquete *edgeR* (Robinson & others 2010) de *Bioconductor*. Para cada comparación se determinaron los correspondientes contrastes estadísticos y valores de p, a nivel de miARN.

4.3.2.3. Análisis de enriquecimiento de grupos de genes

A partir de los resultados del análisis anterior, disponemos de una lista de todos los genes ordenados por su patrón de diferencia de expresión. En el caso de los estudios de las enfermedades de la piel, esta lista de genes se ordena según el nivel de expresión diferencial entre la condición enfermedad y control. Para los estudios de tumores, la relación de genes está ordenada por el nivel de inhibición de los miARNs que regulan a estos genes (previamente obtuvimos la diferencia de expresión de los miARNs que regulaban a estos genes y transferimos esta información a nivel gen, tal como se describe en el método presentado en el capítulo anterior).

El segundo input que necesitamos para realizar un análisis de enriquecimiento de grupo de genes es la información de la base de datos sobre la que queremos interrogar a nuestros genes. La flexibilidad del procedimiento permite el uso de cualquier tipo de anotación funcional. En la evaluación de la metodología, descargamos esta información de las bases de datos Ensembl, Reactome y KEGG para el procesamiento (exploración, propagación y filtrado de términos funcionales) y uso de las anotaciones descritas.

Con los genes ordenados por un criterio de interés y la anotación funcional de estos genes en distintas bases de datos, se aplicó un método de enriquecimiento funcional de grupos de genes basado en modelos de regresión logística (Montaner et al. 2009; Sartor & others 2009; Montaner & Dopazo 2010).

4.3.3. Metaanálisis funcional

La configuración y exploración de matrices de la medida del efecto y su varianza, es el primer paso de esta propuesta metodológica. Tras disponer de los datos de partida, la evaluación de la variabilidad entre los estudios posibilita la selección del procedimiento de combinación de los efectos. A continuación, el análisis de sensibilidad y la evaluación de sesgos de las medidas combinadas garantizan la robustez de los resultados obtenidos. Por último, la representación de los resultados del metaánalisis funcional y su interpretación dan respuesta a la evaluación funcional de los estudios analizados.

4.3.3.1. Configuración y exploración de matrices de entrada

A partir de los resultados del enriquecimiento funcional con los modelos logísticos, obtenemos una medida del efecto entre enfermos y controles que se representa mediante logaritmos de los *odds ratios* (razón de ventajas) entre ambas condiciones. Para cada una de las funciones evaluadas en la base de datos de interés, disponemos de esta medida de efecto y su varianza. Ambos elementos constituyen los datos de entrada necesarios para el metaanálisis:

1. Matriz de logaritmos de los *odds ratios*.
2. Matriz de varianzas de los logaritmos de los *odds ratios*.

La exploración de estas matrices permitirá un mejor conocimiento de los datos, incluyendo la descripción de la magnitud del efecto y su variabilidad, la detección de valores atípicos y una cuantificación de los valores perdidos

procedentes de la combinación de los distintos estudios (Figura 4.1 y Figura 4.2). Esta valoración orientará sobre la necesidad de aplicar técnicas de imputación de valores perdidos o bien la eliminación de aquellas funciones donde no esté disponible la medida del efecto en un número determinado de estudios.

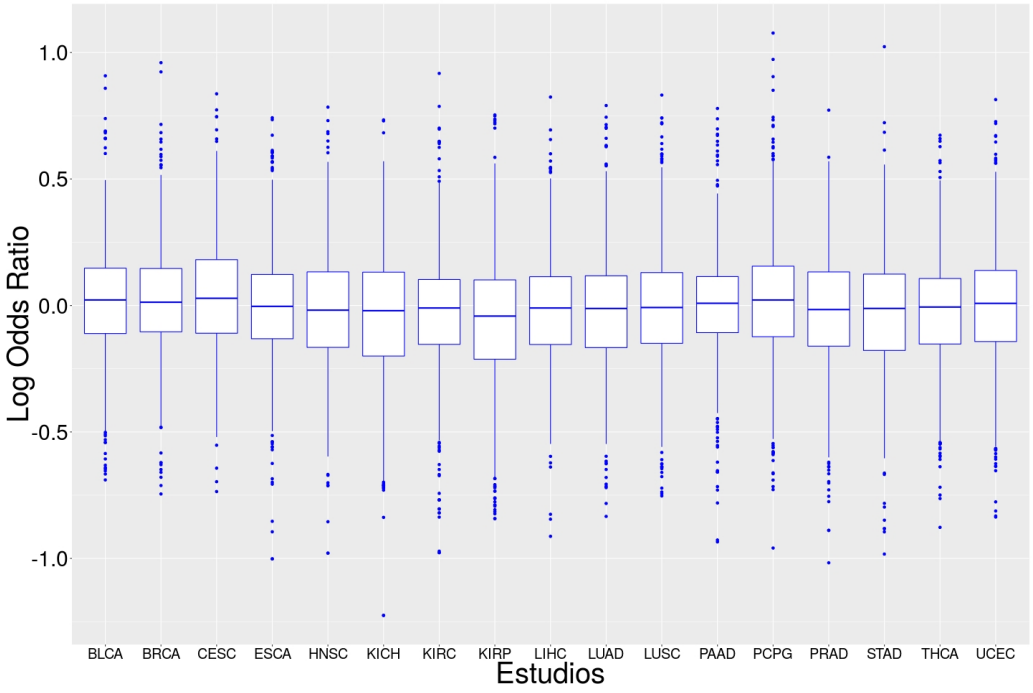

Figura 4.1: Distribución de la medida del efecto por estudio

4.3.3.2. Análisis de heterogeneidad y determinación de la medida combinada del efecto

Para cada una de las funciones de la base de datos seleccionada, se realiza un metaanálisis que combina el efecto medido de todos los estudios analizados. La combinación y presentación de resultados se lleva a cabo utilizando diversas técnicas estadísticas, cuya elección depende fundamentalmente del tipo de medida de resultado/efecto que se ha utilizado y de la valoración del

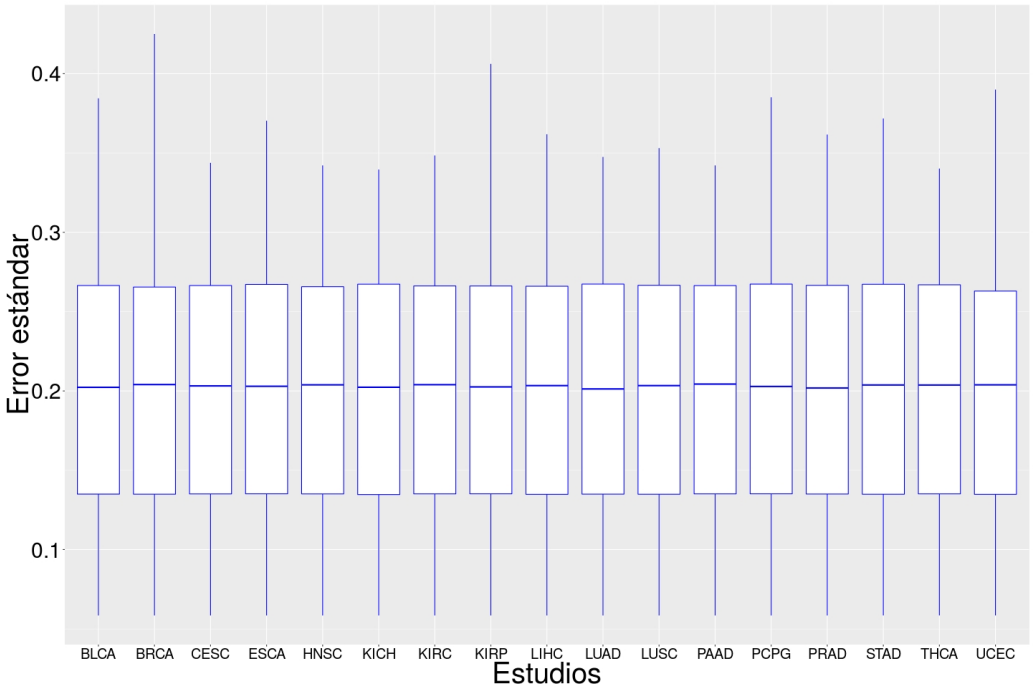

Figura 4.2: Distribución de la varianza de la medida del efecto por estudio

grado de heterogeneidad de los resultados de los estudios.

En un metaanálisis, primero se evalúa la heterogeneidad de los estudios y en función de esta información se decide el modelo de estimación de la variabilidad del efecto. Sin embargo, al trabajar con miles de metaanálisis simultáneamente (tantos como funciones), consideramos la información a priori que disponemos de los estudios para la selección de la técnica de combinación del efecto y tras su ejecución computacional, valoramos los resultados del estudio de heterogeneidad para confirmar la idoneidad del procedimiento seleccionado.

En la obtención de la medida resumen, se ponderan los resultados de los estudios individuales por la inversa de su varianza, de modo que aquellos estudios con mayor variabialidad tendrán un aporte menor sobre el efecto combinado. Si hemos detectado heterogeneidad entre los estudios, ésta puede

ser considerada utilizando un *modelo de efectos aleatorios* o bien no ser incluida, si se utiliza un *modelo de efectos fijos*, donde todos los estudios estiman el mismo efecto y las diferencias observadas se deben únicamente al azar.

De modo que el modelo de efectos fijos asume la existencia de un único efecto en la población y no considera la variabilidad de los resultados entre los estudios primarios, mientras que el modelo de efectos aleatorios incorpora la posible heterogeneidad de los efectos entre los distintos estudios. En este último caso, la ponderación en la determinación de un efecto combinado incluye tanto la variabilidad entre-estudios como la variabilidad intra-estudio. En el contexto genómico, la heterogeneidad entre estudios suele ser habitual por el empleo de distintas plataformas de tecnologías de alto rendimiento, tamaños muestrales y tipos de contrastes entre grupos experimentales. Por ello un modelo de efectos aleatorios se ajustaría mejor a las características de estos estudios.

En el metaanálisis de cada término funcional se utilizaron las funciones incluidas en el paquete *metafor* (Viechtbauer 2010) de R, que incorpora la implementación del modelo de efectos fijos (*FE*) y diferentes modelos de efectos aleatorios:

- *DL*, DerSimonian & Laird (1986).
- *HE*, Hedges et al. (2008).
- *HS*, Schmidt & Hunter (2014).
- *SJ*, Sidik & Jonkman (2005), Sidik & Jonkman (2007).
- *PM*, Paule & Mandel (1982).

Tras aplicar esta metodología a un grupo de estudios, considerando una determinada base de datos biológicos (KEGG, Reactome, Gene Ontology…), obtenemos los estimadores de la medida del efecto en cada una de las funciones estudiadas y una serie de indicadores que valoran la heterogeneidad del metaanálisis realizado. En la Tabla 4.3, se detalla esta información.

ID	*LI 95 %*	*LOR*	*LS 95 %*	*P*	*FDR*	*SE*
GO:0005112	-0.394	-0.238	-0.082	0.003	0.011	0.079
GO:0097472	-0.341	-0.211	-0.081	0.002	0.007	0.066
GO:0005227	0.068	0.221	0.375	0.005	0.017	0.078

ID	*QE*	*QEp*	τ^2	I^2	H^2
GO:0005112	22.25	0.135	0.03	28.089	1.391
GO:0097472	28.916	0.025	0.034	44.667	1.807
GO:0005227	20.262	0.209	0.022	21.034	1.266

Tabla 4.3: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis.

Para una mejor comprensión de los procedimientos utilizados, revisamos cada uno de los indicadores obtenidos en el metaanálisis correspondiente a la función de la Gene Ontology *GO:0005112* (*Notch binding*: función molecular que describe la interacción con la proteína Notch) y que se incluyen en la primera línea de la Tabla 4.3:

- *QE* y *QEp* representan el estadístico de contraste y el valor *p* respectivamente de la prueba de Der Simonian y Laird (DerSimonian & Laird 1986), utilizada para detectar la presencia de heterogeneidad entre los estudios. La hipotésis nula apunta a la presencia de homogeneidad entre los estudios. Con un nivel de confianza del 95 %, en las funciones de la Tabla 4.3 se detecta presencia de heterogeneidad y se confirma la adecuación de un modelo de efectos aleatorios que incorpore esta variabilidad. Además completamos este análisis de la heterogeneidad con varias pruebas que se describen posteriormente.

- *LOR* es la estimación del efecto combinado de todos los estudios. El *logaritmo del odds ratio* para el término GO:0005112 es -0.238. El signo negativo indica mayor presencia de genes con nivel alto de expresión en la segunda clase experimental respecto la primera clase, en la comparación valorada (*enfermo* frente a *control*). La magnitud de este indicador cuantifica la sobrerrepresentación de la función en un grupo frente el otro. Esta función tiene un *LOR* -0.238 y por lo tanto la sobrerrepresentación de la función es mayor en el grupo control que en el grupo enfermo.

- La estimación del efecto estudiado se acompaña de su *intervalo de confianza al 95 %* construido con la variabilidad estimada en el modelo seleccionado (*SE*: 0.079). [-0.394, -0.082] es el *intervalo de confianza* descrito. La no inclusión del valor 0 en el intervalo, confirmaría la significatividad del *LOR*.

- El valor de p (P) informa del nivel de significación de un efecto combinado nulo. Para el término *GO:0005112*, en media, el nivel de sobrerrepresentación es mayor en el grupo control que en el grupo enfermo, siendo significativo con un valor de p de 0.003. Sin embargo, este valor de p no contempla el escenario de multiplicidad que disponemos. Por ello, es necesario ajustar este indicador con algún procedimiento de corrección. El *FDR* 0.011 es el valor corregido de p por el método de la *tasa de falsos descubrimientos* (*FDR*, Benjamini & Hochberg 1995).

- τ^2 es la estimación de la heterogeneidad entre los estudios. Será 0 cuando utilicemos un modelo de efectos fijos. Para la función *GO:0005112* habíamos detectado la presencia de heterogeneidad entre los estudios primarios y su estimación es 0.03.

- Hay otras medidas que facilitan la interpretación de la estimación de la heterogeneidad (Higgins & Thompson 2002): el estadístico I^2 estima (en porcentaje) la relación entre la variabilidad entre estudios y el total de la variabilidad, siendo del 28.089 % para el término *GO:0005112*. Por otra parte, el estadístico H^2 es el cociente entre la variabilidad total y la variabilidad en el muestreo, de modo que cuando la estimación de τ^2 sea 0, entonces H^2 será 1. Para la función *GO:0005112* este indicador muestra un valor de 1.391.

El test Q propuesto por Der Simonian y Lard (DerSimonian & Laird 1986) presenta una baja potencia cuando el número de estudios primarios es pequeño. Por ello, el análisis de la heterogeneidad entre estudios se

acompaña de varios métodos complementarios que se resumen gráficamente en las siguientes representaciones:

Gráficos de embudo

Mediante los *gráficos de embudo* se evalúa la variabilidad de los distintos estudios, así como la presencia de sesgos. En este tipo de representaciones se muestra la magnitud del efecto medido (eje X) frente a una medida de precisión (eje Y), como la desviación estándar o el inverso de la varianza. Cada punto representa un estudio primario y el diagnóstico del gráfico se realiza tras la valoración de la nube de puntos (Sterne & Egger 2001).

En ausencia de sesgos y heterogeneidad, se espera que los puntos se distribuyan en forma de embudo, donde la mayoría de los puntos caerían dentro de la región de confianza de la estimación del efecto. Con otras medidas de precisión para el eje Y, la forma esperada del embudo puede ser diferente.

La Figura 4.3 presenta la relación entre el efecto estudiado y distintos indicadores de la variabilidad para la función GO:0005227. En el eje X se muestran los valores del logaritmo del *odds ratio* y en el eje Y: el error estándar, la varianza en el muestreo y sus respectivos valores inversos (medidas de precisión). En todos los gráficos se repite el mismo patrón:

- Hay dos estudios que se separan de la región de confianza.
- En los dos primeros gráficos, no se presentan grandes cambios en la distribución de la variabilidad (error estándar o varianza) en función del tamaño del efecto. Los estudios con LOR positivos, muestran un ligero incremento de su error estándar y cuando son evaluados mediante

los gráficos que incorporan los valores inversos del error estándar o varianza, se aprecia con mayor claridad, el despunte de dos estudios.

- El número de estudios con efecto positivo y negativo es similar.

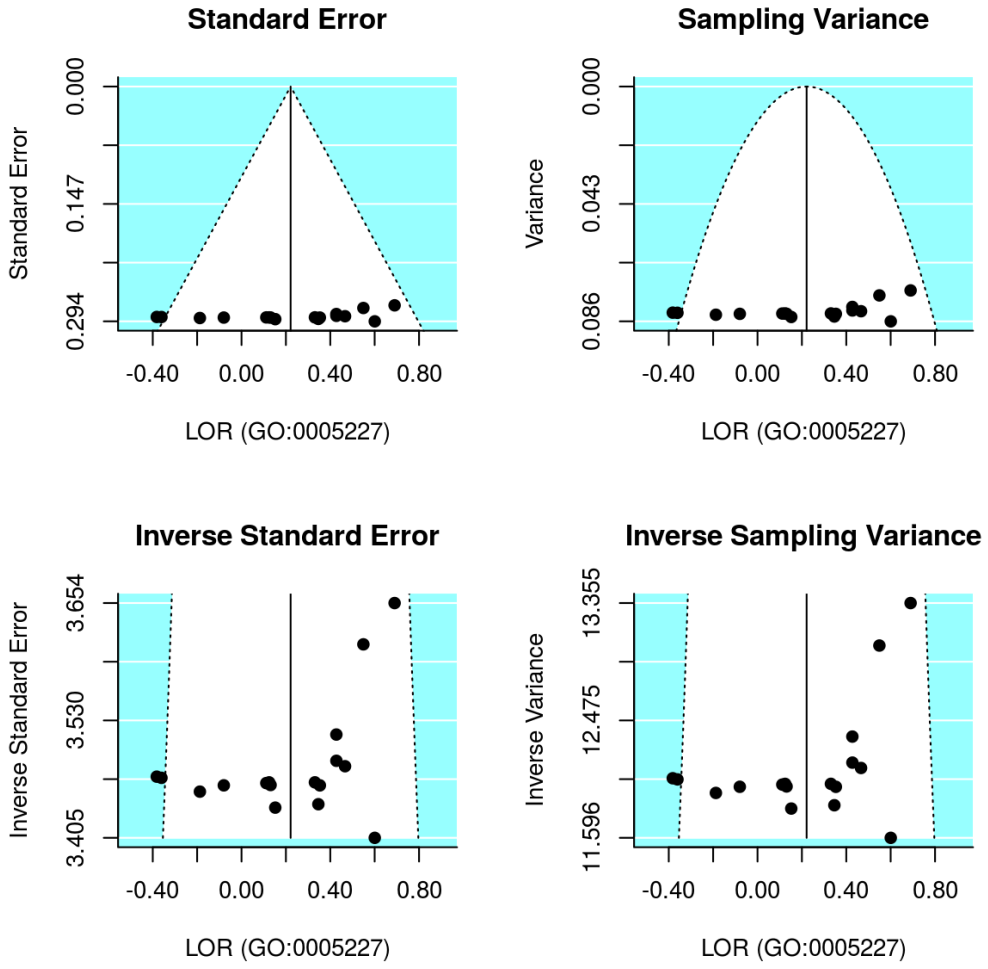

Figura 4.3: Variabilidad del efecto estudiado en la función GO:0005227

Gráficos radial

Los *gráficos radial* (R. Galbraith 1988b; R. Galbraith 1988a; Galbraith 1994) se utilizan para valorar la consistencia de los efectos según su nivel de precisión.

La Figura 4.4 muestra una consistencia de la precisión para los distintos efectos medidos en los estudios. En el eje X se muestra la inversa de los errores estándares (precisión) y en el eje Y el tamaño de los efectos observados estandarizados por su correspondiente error estándar. En la derecha del gráfico, se incluye un arco donde se representan los efectos observados sin estandarizar, de modo que se determina la magnitud del efecto observado para un estudio específico, siguiendo la proyección de la línea con origen en (0,0), pasando por el estudio de interés y llegando a la proyección en el arco.

Por otra parte, en la Figura 4.5 se distinguen dos patrones de variabilidad distintos para la ruta de señalización hsa05146 (*amoebiasis*): hay un primer grupo de estudios con una menor precisión, mientras que en un segundo grupo la precisión es mayor, aunque algunos de los estudios quedan fuera del intervalo de confianza descrito. La presencia de inconsistencia en el nivel de precisión debe ser estudiada para determinar el posible origen de diferencias entre grupos.

La Figura 4.6 incluye dos gráficos radial para una misma función (GO:0001105), utilizando dos métodos diferentes de estimación de la variabilidad del efecto combinado, uno con efectos fijos y otro aleatorios. Se observan diferencias de patrones y magnitud de la variabilidad según el modelo utilizado.

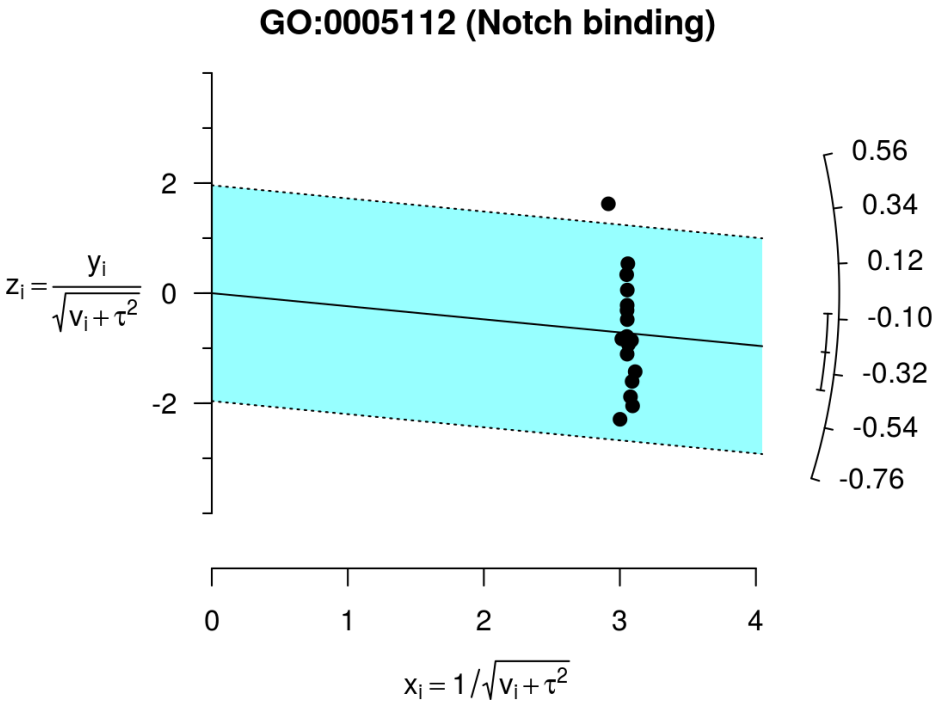

Figura 4.4: Variabilidad del efecto estudiado en la función GO:0005112

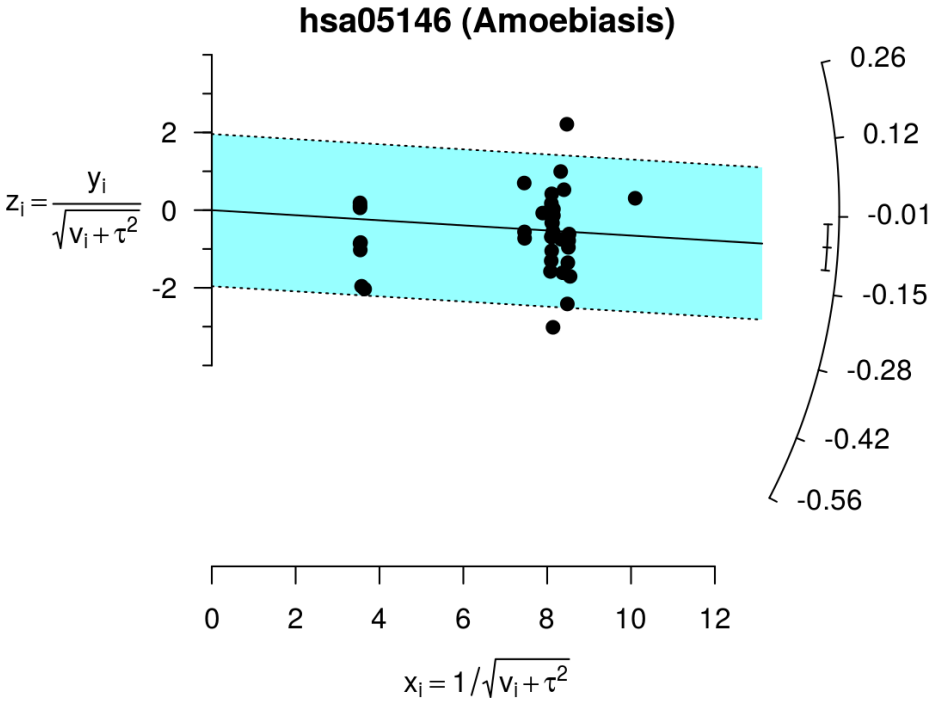

Figura 4.5: Variabilidad del efecto estudiado en la ruta hsa05146

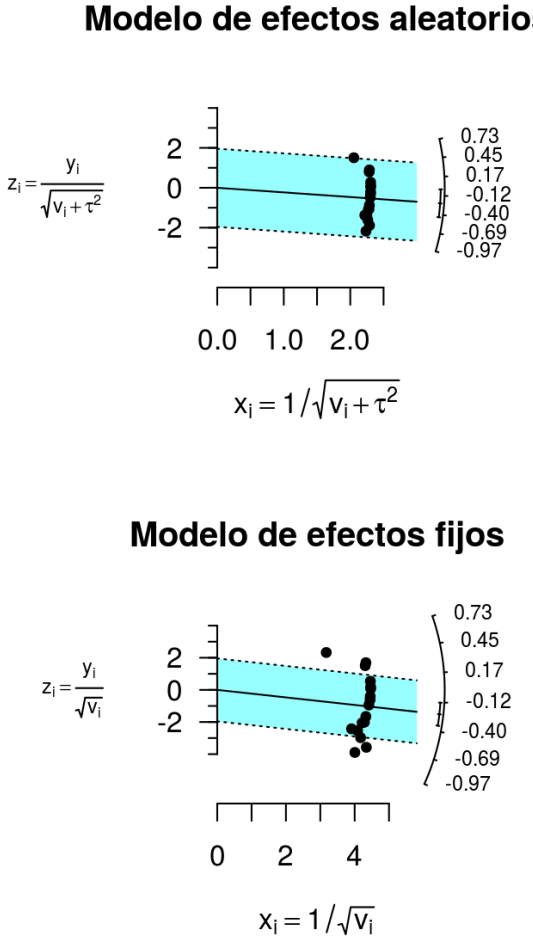

Figura 4.6: Gráficos radial de la función GO:0005227 por métodos de estimación del efecto

Los procedimientos anteriores permiten el análisis de la heterogeneidad para cada una de las funciones evaluadas. También presenta un gran interés, el conocimiento de la distribución de los indicadores descritos en el conjunto de funciones de la base de datos para cada uno de los métodos de estimación del efecto combinado y para ello se utilizaron los diagramas de cajas que se muestran en las Figuras 7 a 11.

Estos resultados permiten comparar los distintos estimadores y elegir el método que mejor se ajuste a los datos de los estudios tratados.

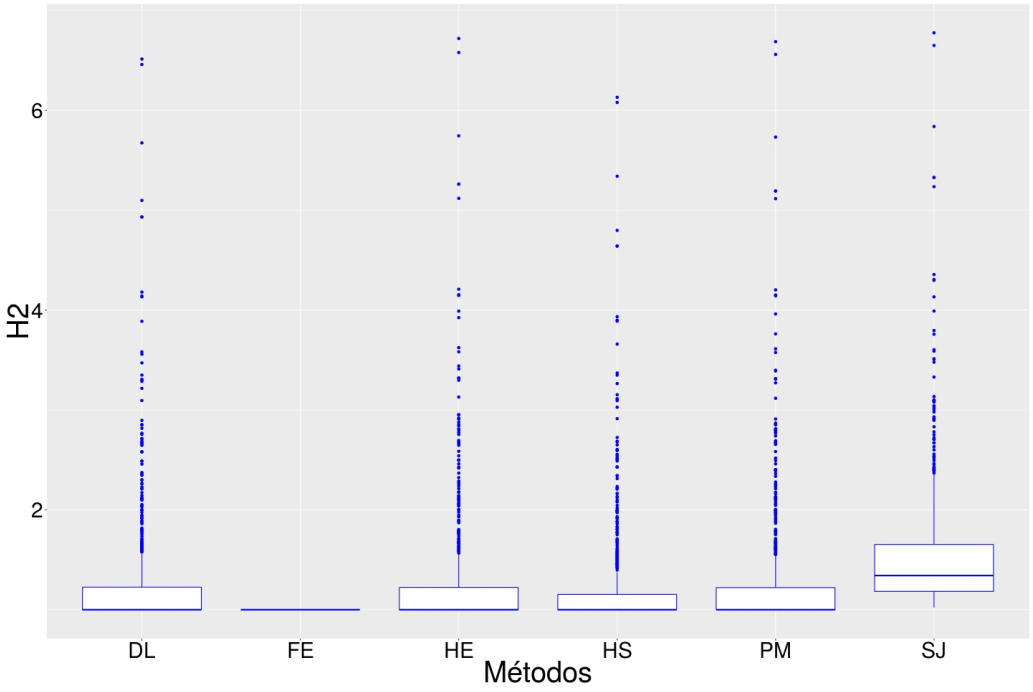

Figura 4.7: Análisis global de la heterogeneidad. Distribución de H^2 por métodos de estimación del efecto

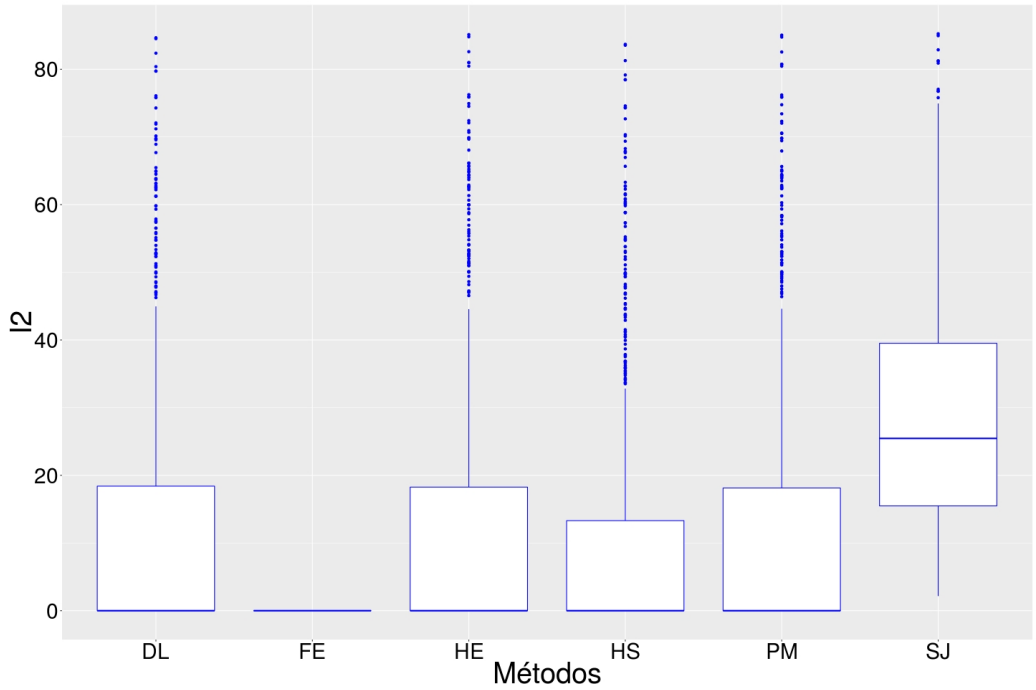

Figura 4.8: Análisis global de la heterogeneidad. Distribución de I^2 por métodos de estimación del efecto

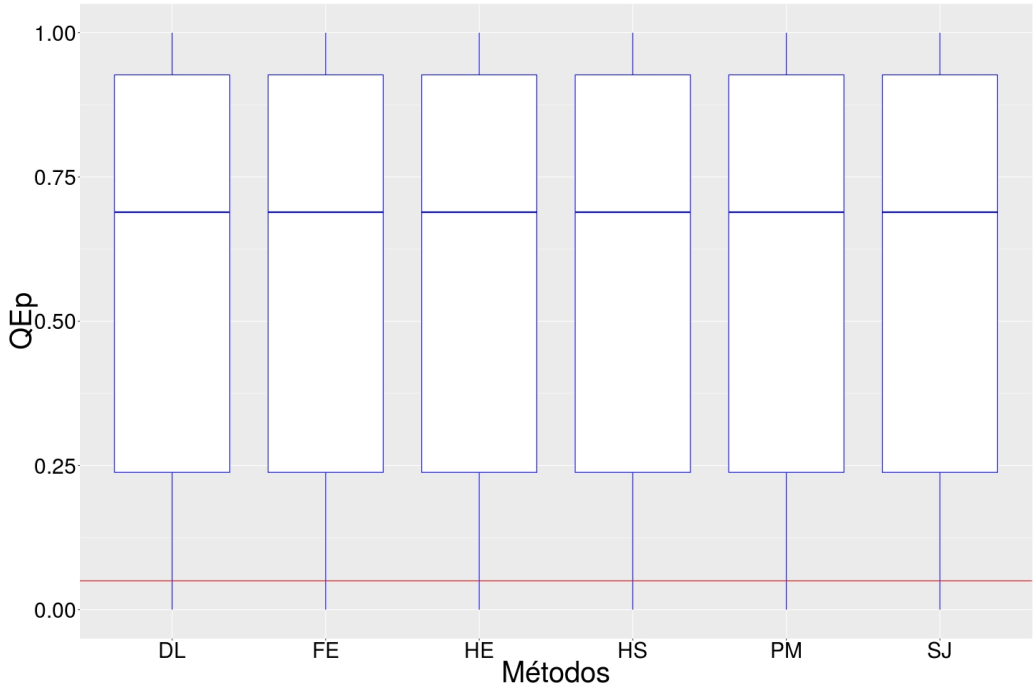

Figura 4.9: Análisis global de la heterogeneidad. Distribución de *QEp* por métodos de estimación del efecto

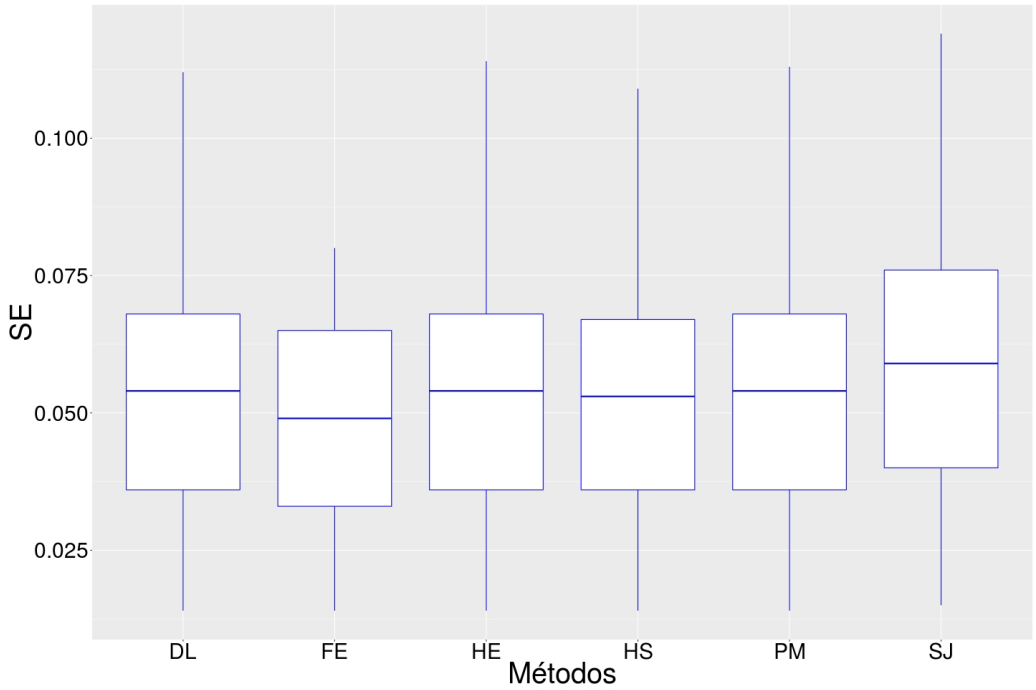

Figura 4.10: Análisis global de la heterogeneidad. Distribución de *SE* por métodos de estimación del efecto

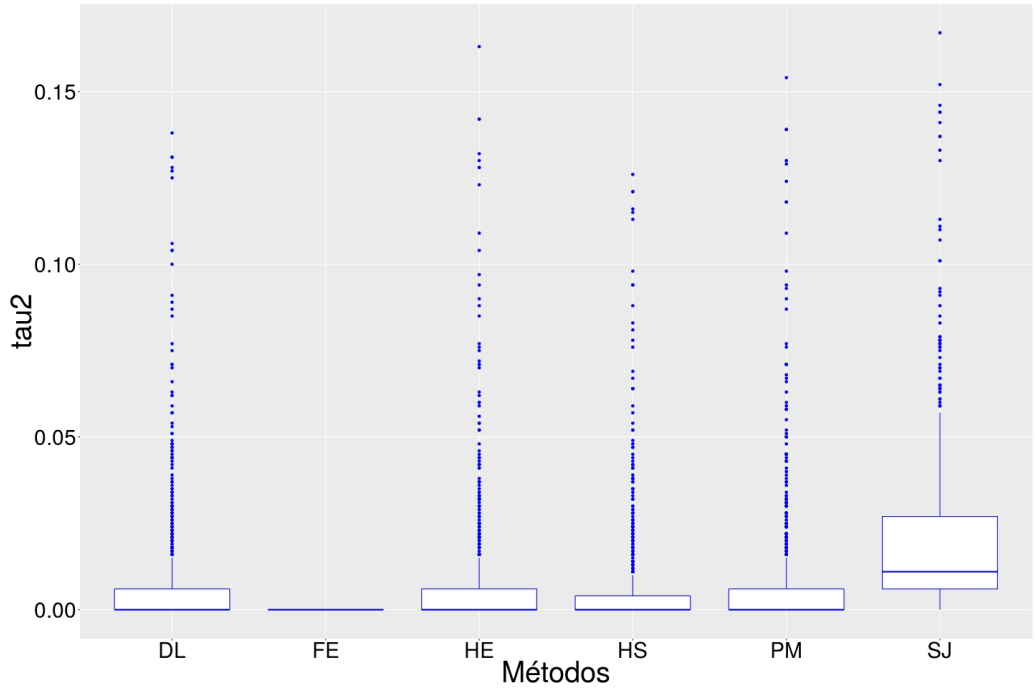

Figura 4.11: Análisis global de la heterogeneidad. Distribución de τ^2 por métodos de estimación del efecto

4.3.3.3. Análisis de sensibilidad y evaluación de sesgos

Análisis de sensibilidad

Tras la realización del metaanálisis, la influencia de cada uno de los estudios iniciales es evaluada mediante un *análisis de sensibilidad*. Para ello, reproducimos el metaanálisis excluyendo el estudio que queremos revisar. Si los resultados son similares a los obtenidos con el total de los estudios, se estará garantizando la robustez del metaanálisis.

El método proporciona una doble información sobre el nivel de sensibilidad de los estudios:

- Por una parte se detalla el impacto de cada uno de los estudios en el metaanálisis de cada función.
- También se ofrece una evaluación global del análisis de sensibilidad por métodos, considerando todas las funciones de la base de datos seleccionada.

En el primer caso, este proceso se desarrolla para cada una de las funciones evaluadas de la base de datos seleccionada. Así por ejemplo, en la Tabla 4.4 se muestran los resultados del metaanálisis con el método DL para un término GO específico. Tras la replicación del análisis una vez que hemos extraído cada uno de los estudios primarios, se determinan los indicadores que informan sobre la estimación del efecto y su variabilidad:

Estudios	*LOR*	*SE*	*ZVAL*	*P*	*LI 95 %*	*LS 95 %*
BLCA	-0.09	0.033	-2.702	0.007	-0.156	-0.025

Estudios	*LOR*	*SE*	*ZVAL*	*P*	*LI 95 %*	*LS 95 %*
BRCA	-0.099	0.034	-2.939	0.003	-0.165	-0.033
CESC	-0.103	0.033	-3.156	0.002	-0.167	-0.039
ESCA	-0.085	0.032	-2.654	0.008	-0.147	-0.022

Estudios	*Q*	*Qp*	τ^2	I^2	H^2
BLCA	17.595	0.285	0.003	14.751	1.173
BRCA	17.803	0.273	0.003	15.745	1.187
CESC	16.874	0.326	0.002	11.106	1.125
ESCA	16.098	0.376	0.001	6.819	1.073

Tabla 4.4: Indicadores del análisis de sensibilidad de los estudios en cada función.

A partir del análisis de sensibilidad específico de cada función, se ha determinado el error estándar de las estimaciones de todas las réplicas del análisis donde se evaluaba el impacto de cada estudio. A continuación se repitió este proceso para todas las funciones de la base de datos evaluada. De modo que disponemos de una distribución de errores estándares para cada uno de los métodos de estimación del efecto. Esta aproximación permite comparar el nivel global de sensibilidad que ofrecen los diversos métodos utilizados. La representación gráfica de este análisis global se muestra en la Figura 4.12. En este caso todos los métodos de estimación del efecto combinado presentan una variabilidad común y magnitudes similares exceptuando el procedimiento *SJ* que presenta una distribución de los errores estándares con valores de

mayor magnitud y por lo tanto, sería el menos robusto.

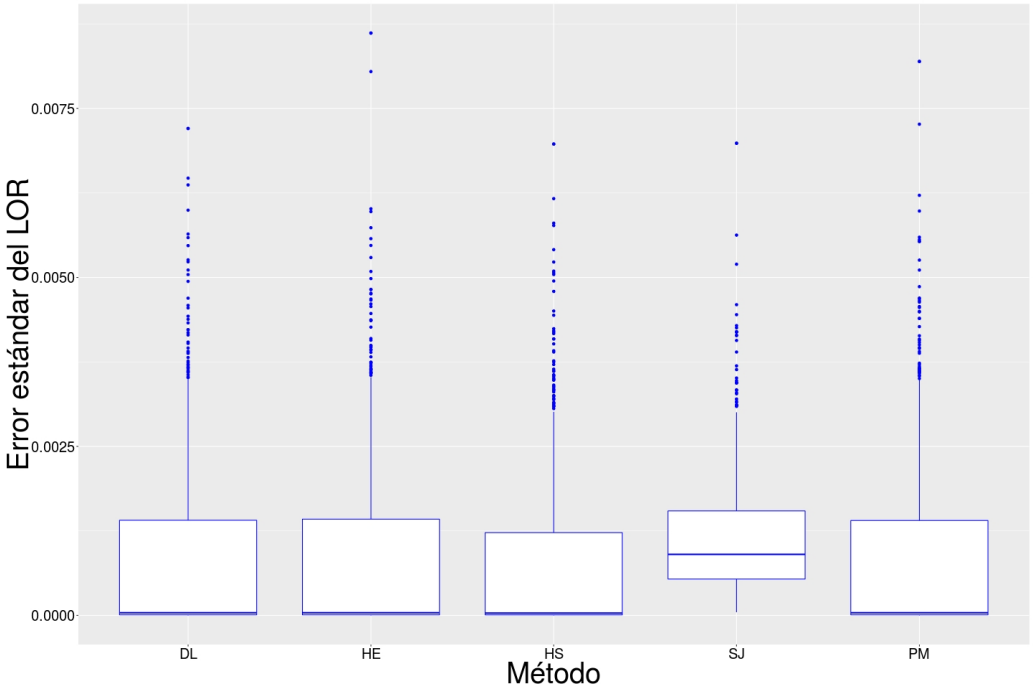

Figura 4.12: Análisis de sensibilidad por métodos de estimación del efecto

Evaluación de sesgos

La revisión de posibles sesgos es importante para garantizar una adecuada interpretación de los resultados. Los gráficos de embudo, presentados en el apartado anterior, además de informarnos de la presencia de heterogeneidad entre los estudios, nos advierten de un posible sesgo de publicación. Se espera la obtención de un efecto de magnitud similar para todos los estudios, alrededor de la recta horizontal y con mayor variabilidad si el tamaño muestral es pequeño, de modo que los puntos se distribuirían en forma de embudo invertido. Si existe un sesgo de publicación, entonces la nube de puntos aparecería deformada en alguno de sus extremos.

Esta detección de asimetrías en los patrones de variabilidad no siempre es sencilla, especialmente si el número de estudios es pequeño. En ocasiones

esta asimetría puede ser debida a la calidad de los estudios. Hay algunos procedimientos que estiman el número de estudios que faltan de un metaanálisis para configurar una distribución simétrica, como el método del "ajuste y relleno". Es una técnica no paramétrica que aumenta los datos observados, examinando la sensibilidad de los resultados (Duval & Tweedie 2000a; Duval & Tweedie 2000b; Rothstein et al. 2006).

Los resultados de la aplicación de este procedimiento se presentan en la Figura 4.13. Los puntos en color negro son los datos observados en los estudios y los puntos en color blanco son los "aumentados". En el gráfico se observa un patrón asimétrico en el nivel de precisión según la medida del efecto.

Análisis de estudios influyentes

La presencia de algunos estudios cuya magnitud y variabilidad son muy diferentes del resto pueden producir una fuerte influencia en el metaanálisis. La exclusión de este estudio en el análisis permite considerar los cambios en el modelo ajustado y valorar su influencia. El diagnóstico de datos influyentes está incorporado en los modelos de regresión (Belsey et al. 1980; Cook & Weisberg 1982) y ha sido adaptado al contexto del metaanálisis para identificar estudios con clara influencia en el modelo.

En la Figura 4.14 se presentan varias medidas de diagnóstico de datos influyentes (residuos estandarizados, las distancias de Cook, covarianza de ratios, estimaciones de τ^2,...) para la función *GO:0005112*. En el grupo de 17 estudios integrados en el análisis, todos ellos presentan un peso similar. Tras la revisión de los gráficos se detecta una influencia mayor en el estudio 12, tal como apunta el descenso de τ^2 o *QE*, coincidente con un incremento de los residuos estandarizados o la distancia de Cook.

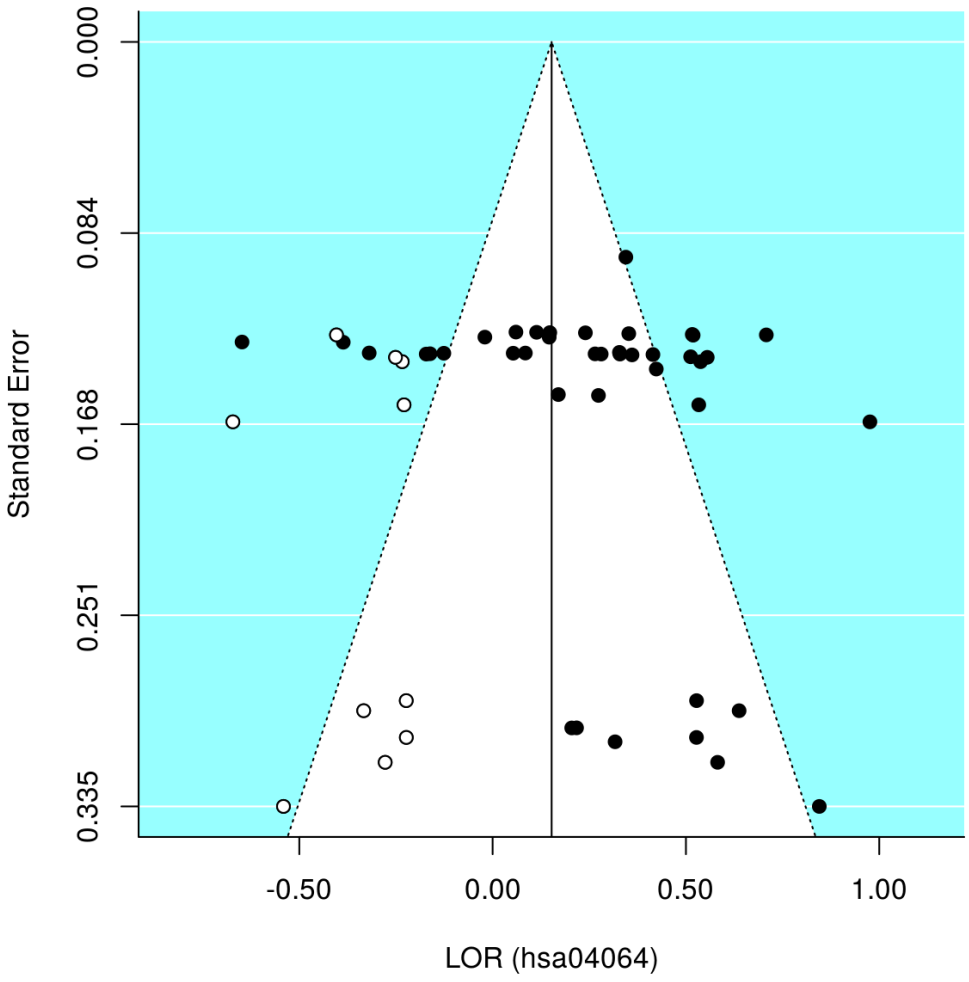

Figura 4.13: Evaluación de sesgos en la ruta hsa04064

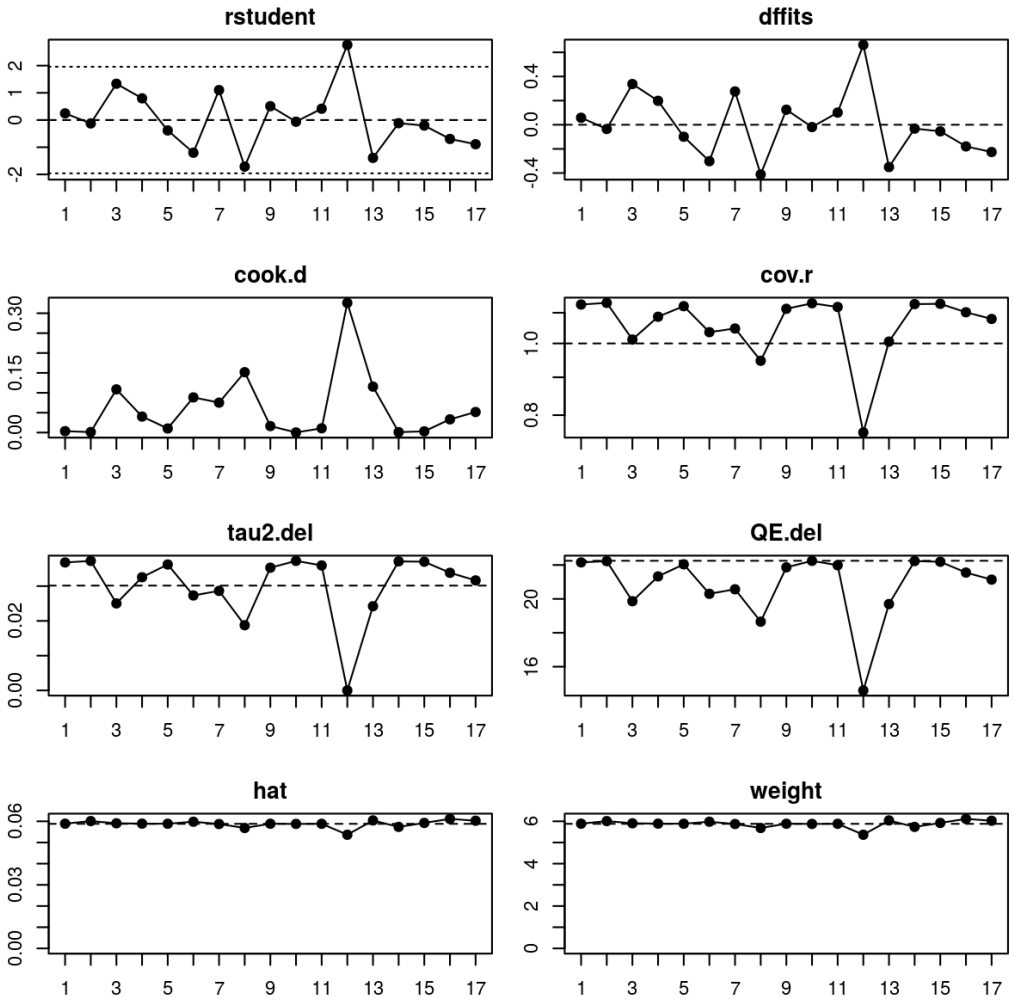

Figura 4.14: Análisis de estudios influyentes para la función GO:0005112

4.3.3.4. Representación e interpretación de resultados

La metodología propuesta para el metaanálisis funcional proporciona resultados globales sobre el conjunto de las anotaciones y resultados específicos a nivel de función, ofreciendo una evaluación de la integración de los estudios en el modelo. Su representación gráfica favorece la rápida interpretación de los resultados del metaanálisis.

Resultados globales del metaanálisis

El primer resultado global de nuestra metodología es una tabla resumen que informa sobre el desarrollo y obtención de resultados significativos tras la combinación de la información de los estudios iniciales.

En la Tabla 4.5, cada una de las filas muestra los resultados de un método de estimación de la variabilidad del efecto medido. Por defecto, generamos un metaanálisis por cada método y función. Ello permite la comparabilidad y la confirmación de la selección adecuada del método.

Así por ejemplo, en la primera fila de esta tabla, aparecen los resultados referentes a la aplicación del método de *DerSimonian-Laird* (DL):

- Las columnas 2 y 3 muestran el número de rutas KEGG sobrerrepresentadas en cada clase experimental: 79 rutas sobrerrepresentadas en el grupo de *enfermos* (términos KEGG sobrerrepresentados en grupos de genes que tienen un nivel de expresión alto en el grupo de enfermos) y 223 rutas sobrerrepresentadas en el grupo de *controles* (términos KEGG sobrerrepresentados en grupos de genes que tienen un nivel de expresión alto en la clase control).

- Las columnas 4 y 5 representan el número de rutas KEGG significativas y sobrerrepresentadas en cada grupo experimental: 27 rutas sobrerrepresentadas en el grupo de *enfermos* y 163 rutas sobrerrepresentadas en el grupo de *controles*. Los valores de p que indican la significación de cada función en el metaanálisis, han sido corregidos por el método de la *tasa de falsos descubrimientos* (*FDR*, Benjamini & Hochberg (1995)).

- Las columnas 6 y 7 combinan dos criterios de selección de términos funcionales: aquellas funciones que son significativas y además tienen efecto de sobrerrepresentación entre grupos con mayor magnitud. En este caso se describen las rutas de señalización significativas y con un valor absoluto del logaritmo del *odds ratio* mayor de 0.5.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	79	223	27	163	0	3
HE	78	223	27	165	0	3
HS	79	223	27	164	0	3
SJ	77	224	21	158	0	3
PM	79	223	25	163	0	3
FE	84	216	45	173	0	3

Tabla 4.5: Resultados globales del metaanálisis funcional con rutas de señalización KEGG.

La interpretación de los resultados es análoga para el resto de estimadores.

La inspección de la Tabla 4.5 de los resultados globales del metaanálisis conduce a la selección de resultados de un modelo determinado de metaanálisis. Siguiendo con el método elegido anteriormente (*DL*), se han obtenido resultados que informan sobre todas las funciones valoradas en el procedimiento. En la Tabla 4.6 se muestra la estructura de resultados por función.

ID	*LI 95 %*	*LOR*	*LS 95 %*	*P*	*FDR*
hsa00010	-0.213	-0.093	0.027	0.129	0.229
hsa00020	-0.270	-0.133	0.005	0.058	0.124
hsa00040	-0.438	-0.275	-0.112	0.001	0.006
hsa04064	0.125	0.244	0.363	0	0.001

Tabla 4.6: Resultados específicos del metaanálisis funcional con rutas de señalización KEGG. Estimadores de la medida del efecto e indicadores de heterogeneidad.

En la representación gráfica de todos los resultados del metaanálisis funcional, para un método específico de estimación de la variabilidad, empleamos los *gráficos volcán*. Son diagramas de puntos utilizados para describir resultados de análisis con datos numerosos como los ómicos y detectar cambios de interés. Muestran simultáneamente una magnitud de cambio frente a una medida de significación estadística. En el eje X se representan los logaritmos de los *odds ratios* y en el eje Y el nivel de significación (valor negativo del logaritmo en base 10 del valor de p).

En el contexto del metaanálisis, cada punto representa una de las funciones evaluadas en la base de datos seleccionada. Así por ejemplo en la Figura 4.15, las funciones que superan el valor 1.30103 en el eje Y son resultados significativos en el metaanálisis (valor ajustado de $p < 0.05$). Las funciones están coloreadas en rojo cuando el efecto combinado es negativo y verde cuando la medida es positiva. En azul se representan las funciones no significativas en el metaanálisis. La representación confirma una presencia balanceada de funciones con efecto negativo (sobrerrepresentadas en el grupo de controles) y positivo (sobrerrepresentadas en el grupo de enfermos). Las funciones significativas presentan generalmente una magnitud mayor del efecto combinado.

Específicamente para las funciones de la Gene Ontology, se utilizan grafos dirigidos que representan los términos significativos, ofreciendo además una detallada información de las relaciones entre las funciones. En la Figura 4.16, los nodos coloreados corresponden a las funciones moleculares sobrerrepresentadas en el grupo de enfermos. El resto de nodos sin color, se incluyen en el grafo para una mejor comprensión de las relaciones entre las funciones.

Resultados específicos del metaanálisis a nivel de función

Para visualizar el efecto aportado de cada estudio en la estimación del efecto global en una función determinada, utilizamos los *gráficos de bosque*. Este gráfico representa un bosque donde los árboles serían los estudios primarios del metaanálisis y donde se resumen todos los resultados relevantes de la síntesis cuantitativa.

En la Figura 4.17 se muestran los resultados para el término *GO:0005112* (*Notch binding*) referido anteriormente. Elementos de interés en la representación gráfica:

Figura 4.15: Resultados del metaanálisis de funciones moleculares en estudios pareados de tumores

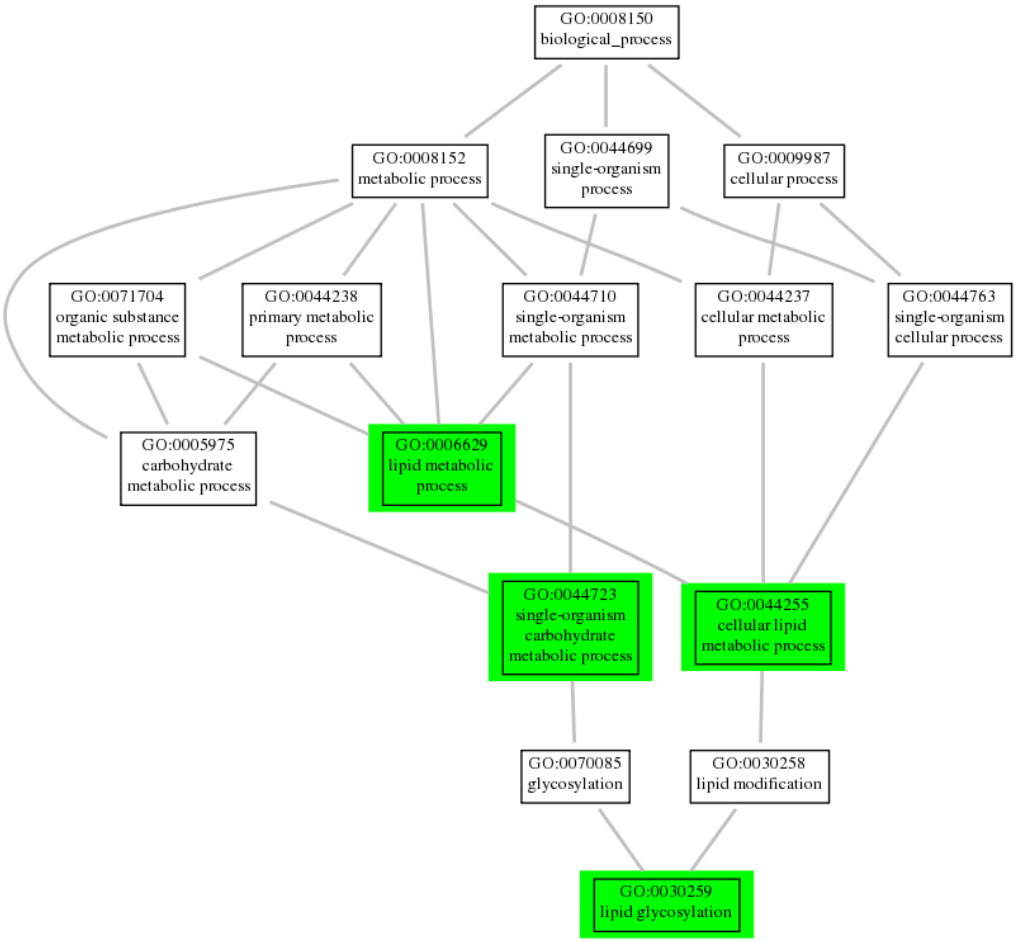

Figura 4.16: Funciones moleculares significativas con sobrerrepresentación en el grupo de enfermos

- A la izquierda del gráfico se enumeran los estudios incluidos en el metaanálisis.
- En la parte derecha se incluye la estimación de la medida resumen individual de cada estudio y su intervalo de confianza (95 %).
- En el centro del gráfico se visualiza la medida del efecto (cuadrado en color negro) cuyo tamaño es proporcional a la precisión de las estimaciones, de modo que una mayor variabilidad se visualizaría con una figura de menor tamaño. El cuadrado está ubicado dentro de un segmento que representa los extremos de su intervalo de confianza.
- En la parte inferior, el resultado global del metaanálisis se representa con un rombo en color verde. Su posición respecto a la línea de efecto nulo nos informa sobre la significación estadística del resultado global, mientras que su anchura nos proporciona una idea de su precisión (su intervalo de confianza). Para esta función, observamos que la estimación del efecto global es negativo (-0.24).
- También en la parte inferior derecha de esta figura se indica el valor de significación de los intervalos de confianza (habitualmente 95 %) y a la izquierda el modelo de análisis de datos que se ha utilizado. En este caso se indica que el modelo incluyó el estimador de DerSimonian-Laird (modelo de efectos aleatorios).

En la Figura 4.18 se muestran los resultados del metaanálisis para la función *GO:0005227* (*calcium activated cation channel activity*) donde la medida resumen del efecto es positiva.

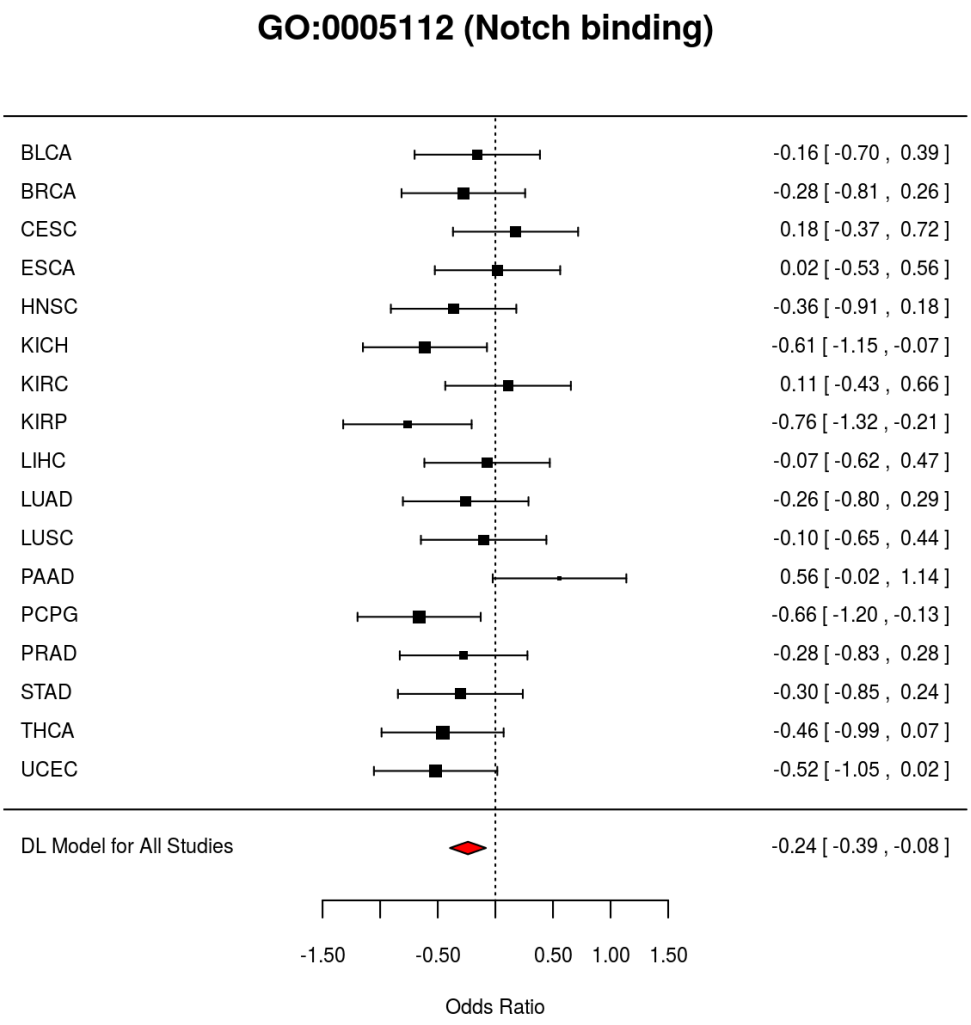

Figura 4.17: Distribución del efecto para la función GO:0005112

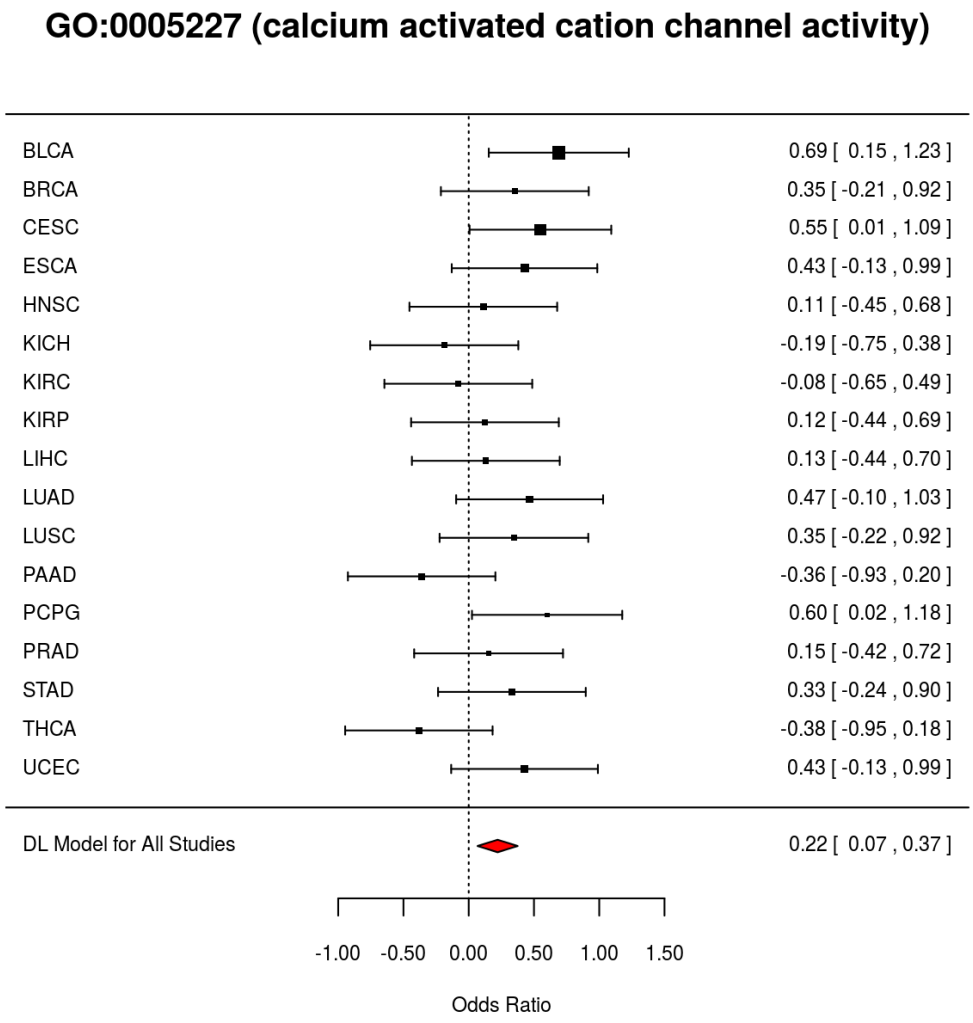

Figura 4.18: Distribución del efecto para la función GO:0005227

4.4. Resultados

Para cada uno de los dos grupos de estudios descritos (enfermedades de la piel y tumores) se siguieron los pasos descritos en la Figura 4.19.

Figura 4.19: Estrategia de análisis en estudios de metaanálisis

Los resultados del metaanálisis de las funciones incluidas para cada base de datos seleccionada (KEGG, Reactome y Gene Ontology), se han organizado en dos niveles :

- Resultados globales del metaanálisis para el conjunto de funciones incluidas en la base de datos seleccionada.
- Resultados específicos del metaanálisis de cada una de las funciones evaluadas.

Los resultados globales del metaanálisis nos proporcionan indicadores y representaciones gráficas para conocer cómo es la medida combinada del efecto en el conjunto de las funciones de la bases de datos de interés. También se incluyen los resultados globales por método de estimación de la variabilidad del efecto entre los estudios, que permiten confirmar la elección del procedimiento más adecuado.

Por otra parte, tras conocer un escenario global de resultados, la metodología propuesta proporciona un grupo de resultados específicos de cada función que posibilita la revisión exhaustiva del peso de cada estudio, la presencia de variabilidad y sesgos, así como la interpretación de la información obtenida.

4.4.1. Enfermedades dermatológicas

A continuación se muestran los resultados del metaanálisis utilizando dos bases de datos de rutas: KEGG y Reactome.

4.4.1.1. KEGG

En la Tabla 4.7 se describen los resultados globales obtenidos para los diferentes métodos de metaanálisis evaluados. Todos los modelos presentados son de efectos aleatorios (*DL, HE, HS, SJ, PM*) excepto *FE* que corresponde a un modelo de efectos fijos.

Para cada método de estimación de la heterogeneidad de los datos, se presenta el número de rutas que están sobrerrepresentadas en los grupos de los enfermos y los controles respectivamente. Por ejemplo, utilizando el estima-

dor *DL* (DerSimonian-Laird), obtenemos 79 términos KEGG sobrerrepresentados en grupos de genes que tienen un nivel de expresión alto en el grupo de enfermos de psoriasis o dermatitis. Por otra parte, se detectan 223 rutas de señalización en las que están participando genes con un nivel alto de expresión en el grupo de controles (o bien genes con un nivel bajo de expresión en el grupo enfermo). Esta estimación incorpora la información de los diferentes estudios incluidos en el análisis, de ellos 27 y 163 son respectivamente significativos en el grupo enfermo y control (columnas *Sig.E* y *Sig.C*). Las columnas 6 y 7 muestran el número de rutas KEGG significativas y que además presentan una magnitud del $LOR > 0.05$ para cada grupo experimental (*Sig.LOR.E* y *Sig.LOR.C*). La interpretación de los resultados es análoga para el resto de estimadores.

Cuando indicamos que hay sobrerrepresentación de estas 3 funciones en los controles, queremos decir que detectamos un conjunto de genes que participan en esas rutas de señalización y que además tienen un patrón de expresión alto en el grupo control, o lo que sería equivalente, un conjunto de genes con un patrón de expresión bajo en el grupo enfermo. Esta información sobre la alteración de una vía asociada a un grupo experimental de interés puede ser relevante en diferentes situaciones como la planificación de posteriores estudios, la detección de dianas terapéuticas o la confirmación de hipótesis iniciales.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	79	223	27	163	0	3
HE	78	224	27	165	0	3

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
HS	79	223	27	164	0	3
SJ	77	225	21	158	0	3
PM	79	223	25	163	0	3
FE	85	217	45	173	0	3

Tabla 4.7: Resultados globales del metaanálisis funcional con rutas de señalización KEGG utilizando diferentes métodos de estimación de la variabilidad del efecto medido.

En la Tabla 4.8 se detalla el nivel de información de los resultados. Se presentan los estimadores de la medida del efecto y los indicadores de heterogeneidad en el metaanálisis con el estimador *DL* (DerSimonian-Laird), para las tres funciones con mayor significación y magnitud del efecto.

La disposición de diferentes niveles de información en los resultados del metaanálisis permite la revisión desde lo más general a lo más específico, facilitando la interpretación de los resultados.

ID	*Nombre*	*LI 95 %*	*LOR*	*LS 95 %*
hsa00730	Thiamine metabolism	-0.943	-0.761	-0.579
hsa00785	Lipoic acid metabolism	-0.754	-0.576	-0.398
hsa01220	Degradation of arom. compounds	-0.889	-0.534	-0.179

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
hsa00730	0	0	51.360	0.016	0.093	0.105	37.695	1.605

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
hsa00785	0	0	57.583	0.035	0.091	0.099	30.535	1.440
hsa01220	0.003	0.006	58.562	0.001	0.181	0.458	50.480	2.019

Tabla 4.8: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con rutas de señalización KEGG.

4.4.1.2. Reactome

En la Tabla 4.9 se describen los resultados globales obtenidos para los diferentes métodos de metaanálisis evaluados.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	963	824	425	379	103	83
HE	966	821	511	425	136	104
HS	964	823	443	383	111	87
SJ	956	831	409	358	115	90
FE	995	792	679	508	189	142

Tabla 4.9: Resultados globales del metaanálisis funcional con rutas de Reactome utilizando diferentes métodos de estimación de la variabilidad del efecto medido.

Cuando nos centramos en uno de los métodos de estimación del efecto (DerSimonian-Laird), observamos con detalle las rutas con una significativa

sobrerrepresentación. En la Tabla 4.10 se muestran 3 de las 186 rutas significativas de Reactome con mayor magnitud de sobrerrepresentación en los grupos experimentales evaluados. La exploración del conjunto de las rutas significativas proporciona una interpretación de los resultados de los estudios transcriptómicos considerados.

ID	*Nombre*
R-HSA-111367	SLBP independent Processing of Histone Pre-mRNAs
R-HSA-111446	Activation of BIM and translocation to mitochondria
R-HSA-111448	Activation of NOXA and translocation to mitochondria

ID	*LI 95 %*	*LOR*	*LS 95 %*
R-HSA-111367	0.218	0.527	0.836
R-HSA-111446	0.098	0.522	0.946
R-HSA-111448	0.259	0.787	1.315

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
R-HSA-111367	0.001	0.003	119.513	0	0.158	0.582	73.225	3.735
R-HSA-111446	0.016	0.038	170.452	0	0.216	1.216	81.226	5.327
R-HSA-111448	0.003	0.01	616.201	0	0.269	2.642	93.509	15.405

Tabla 4.10: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con rutas de Reactome.

La revisión pormenorizada de esta información permite la detección de un posible sesgo en la selección de estudios o la presencia de heterogeneidad por alguna causa específica.

Los resultados detallados para los diferentes estimadores del efecto combinado, en cada una de las funciones evaluadas de la base de datos KEGG y Reactome, están disponibles en los enlaces descritos del Apéndice 2.

4.4.2. Tumores

Por las características del diseño utilizado se realizó un metaanálisis de los estudios pareados y otro para los estudios no pareados, abordando en cada escenario las siguientes bases de datos:

4.4.2.1. Gene Ontology. Componentes celulares

Las Tablas 4.11 y 4.12 muestran los resultados globales obtenidos para los diferentes métodos de metaanálisis evaluados, respectivamente en estudios de tumores con muestras pareadas y no pareadas. Ambas tablas presentan diferentes niveles de información: desde el nivel más general donde se indican los componentes celulares con mayor sobrerrepresentación en cada grupo experimental (columnas 2 y 3), a un nivel más preciso donde se describen los términos significativos (columnas 4 y 5) y un nivel más fino donde se combinan diferentes criterios de detección: significación < 0.05 y magnitud del efecto > 0.5 (columnas 6 y 7).

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	225	246	83	128	0	3
HE	225	246	83	128	0	3
HS	225	246	85	130	0	3
SJ	226	245	71	116	0	3
PM	225	246	83	128	0	3
FE	225	246	98	137	0	4

Tabla 4.11: Resultados del metaanálisis funcional en estudios *pareados* de tumores con componentes celulares. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	222	249	65	85	0	1
HE	222	249	65	86	0	1
HS	222	249	66	87	0	1
SJ	222	249	54	74	0	1
PM	222	249	65	85	0	1
FE	222	249	89	118	0	1

Tabla 4.12: Resultados del metaanálisis funcional en estudios *no pareados* de tumores con componentes celulares. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

Tras la selección de un método de estimación del efecto, por ejemplo el estimador *DL* (DerSimonian-Laird) en un modelo de efecto aleatorios, detectamos 3 términos funcionales con una mayor sobrerrepresentación en el grupo control respecto los enfermos en estudios pareados (*Podosome*, *Beta-catenin destruction complex* y *Myofilament*) y una función con mayor sobrerrepresentación también en el grupo control respecto los enfermos en estudios no pareados, siendo coincidentes estos términos funcionales con los obtenidos en los estudios pareados. Esta información se describe en las Tablas 4.13 y 4.14.

ID	*Nombre*	*LI 95 %*	*LOR*	*LS 95 %*
GO:0002102	podosome	-0.709	-0.566	-0.423
GO:0030877	beta-catenin destr. complex	-0.782	-0.591	-0.4
GO:0036379	myofilament	-0.694	-0.54	-0.387

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
GO:0002102	0	0	21.987	0.144	0.073	0.025	27.228	1.374
GO:0030877	0	0	27.298	0.038	0.098	0.067	41.388	1.706
GO:0036379	0	0	13.896	0.606	0.078	0	0	1

Tabla 4.13: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con componentes celulares. Estudios *pareados*.

ID	*Nombre*	*LI 95 %*	*LOR*	*LS 95 %*
GO:0030877	beta-catenin destr. complex	-0.668	-0.524	-0.381

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
GO:0030877	0	0	20.792	0.348	0.073	0.009	8.621	1.094

Tabla 4.14: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con componentes celulares. Estudios *no pareados*.

Los componentes celulares significativos y con mayor magnitud del efecto en cada uno de los dos grupos de estudios, se representan en las Figuras 4.20 y 4.21 utilizando grafos dirigidos que muestran la relaciones entre los diferentes términos GO.

4.4.2.2. Gene Ontology. Funciones moleculares

Las Tablas 4.15 y 4.16 muestran las funciones moleculares con una mayor sobrerrepresentación en enfermos y controles. En los estudios de tumores pareados se detectaron 2 funciones significativas con sobrerrepresentación en controles y con una magnitud del efecto > 0.05 (Figura 4.22). En los estudios no pareados de tumores, no se detectaron resultados significativos con una magnitud del efecto descrito, aunque sí encontramos funciones significativas con una magnitud menor.

Cuando indicamos que hay sobrerrepresentación de estas 2 funciones en los

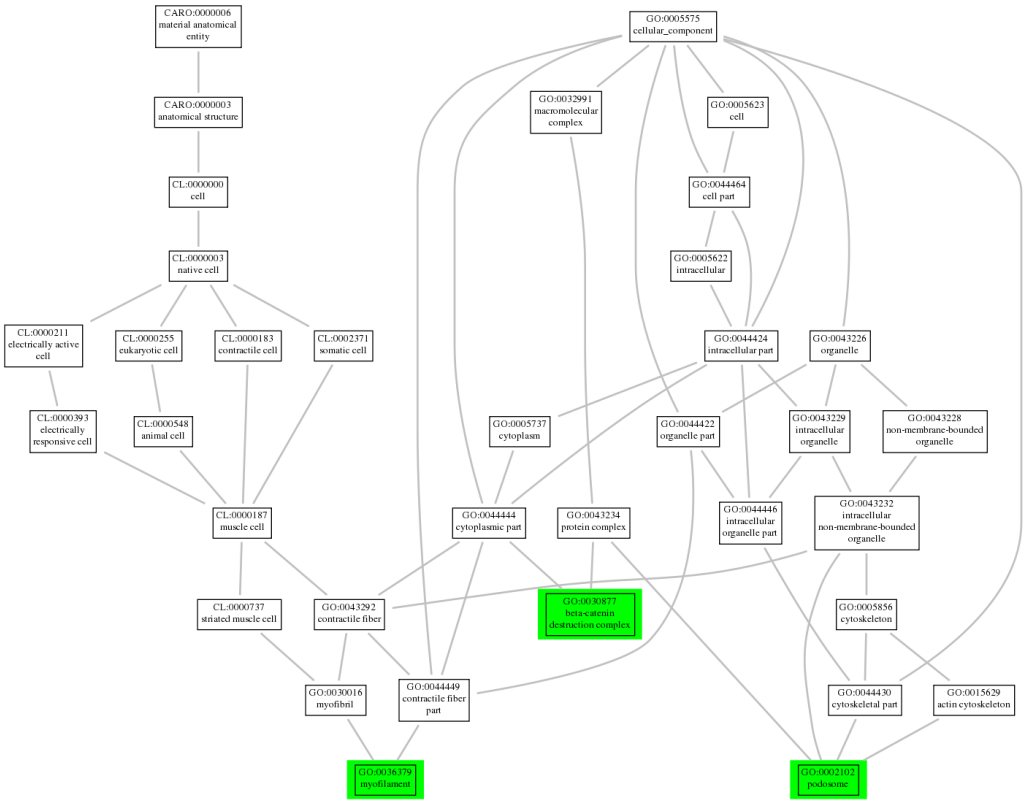

Figura 4.20: Componentes celulares significativos y con mayor magnitud del efecto en estudios pareados

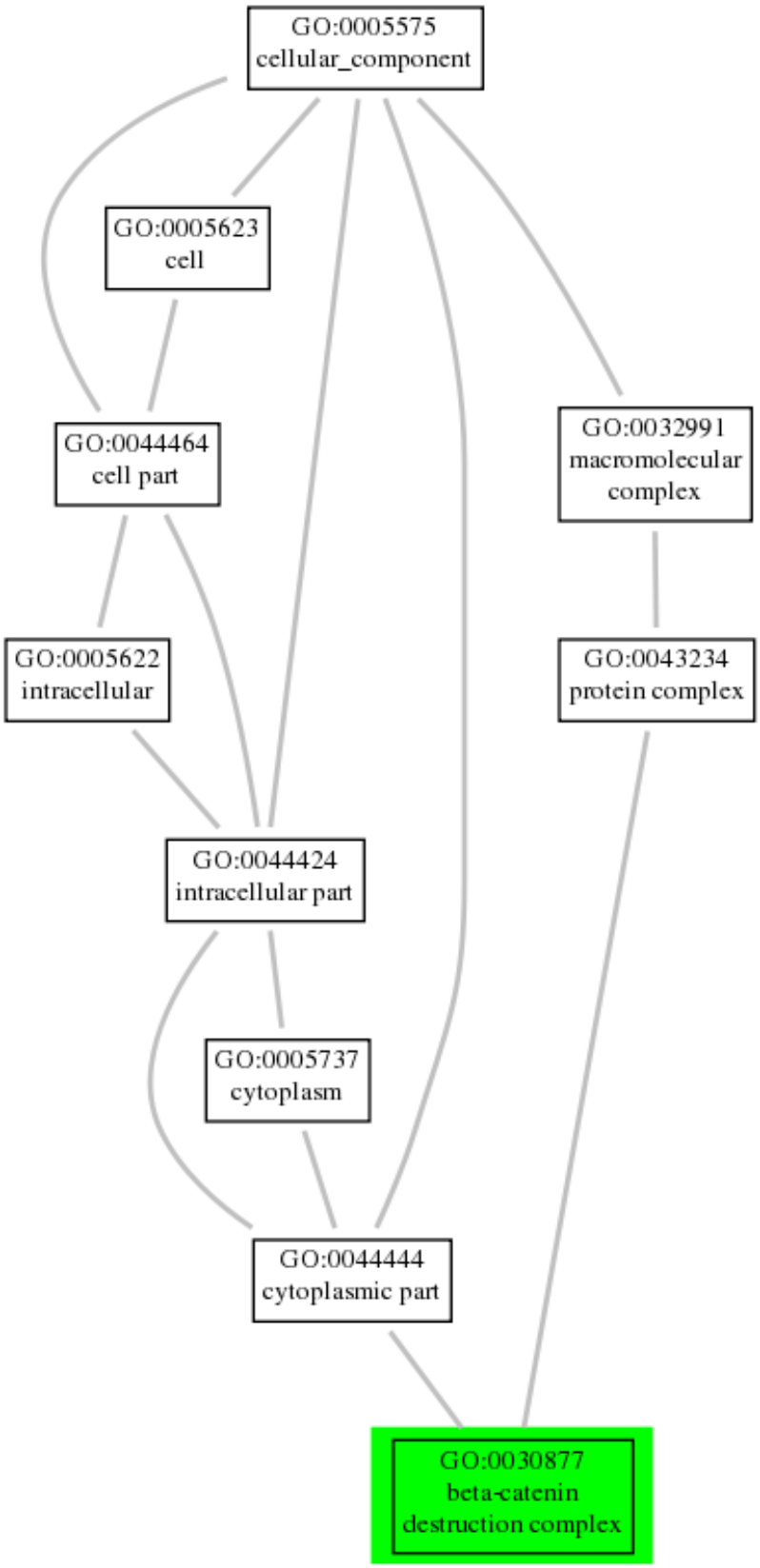

Figura 4.21: Componentes celulares significativos y con mayor magnitud del efecto en estudios no pareados

controles, queremos decir que detectamos un conjunto de genes que realizan esta función y que además tienen un patrón de expresión alto en el grupo control, o lo que sería equivalente, un conjunto de genes con un patrón de expresión bajo en el grupo enfermo.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	351	398	120	136	0	2
HE	351	398	120	136	0	2
HS	351	398	123	139	0	2
SJ	352	397	105	117	0	2
PM	351	398	120	136	0	2
FE	351	398	133	169	0	2

Tabla 4.15: Resultados del metaanálisis funcional en estudios *pareados* de tumores con funciones moleculares. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	362	387	68	87	0	0
HE	362	387	70	91	0	0
HS	362	387	72	95	0	0
SJ	362	387	55	61	0	0
PM	362	387	69	91	0	0
FE	362	387	98	136	0	0

Tabla 4.16: Resultados del metaanálisis funcional en estudios *no pareados* de tumores con funciones moleculares. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

ID	*Nombre*	*LI 95 %*	*LOR*	*LS 95 %*
GO:0004559	alpha-mannosidase	-0.685	-0.512	-0.34
GO:0015923	mannosidase	-0.685	-0.512	-0.34

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
GO:0004559	0	0	24.53	0.079	0.088	0.046	34.773	1.533
GO:0015923	0	0	24.53	0.079	0.088	0.046	34.773	1.533

Tabla 4.17: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con funciones moleculares. Estudios *pareados*.

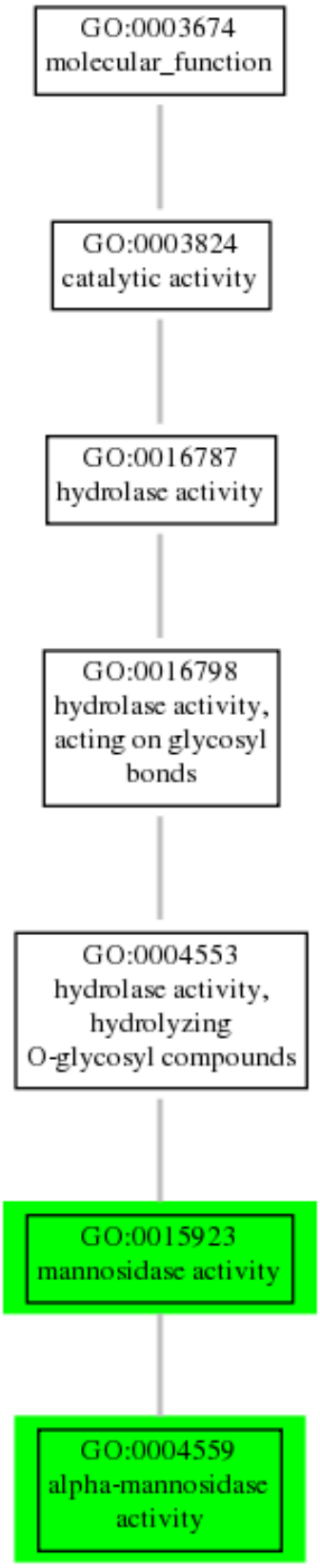

Figura 4.22: Funciones moleculares significativas y con mayor magnitud del efecto en estudios pareados

4.4.2.3. Gene Ontology. Procesos biológicos

Las Tabla 4.18 y Tabla 4.19 muestran los resultados del metaanálisis con diferentes métodos de estimación de la variabilidad del efecto. En general el número de procesos biológicos sobrerrepresentados en los grupos de enfermos y controles es similar a lo largo de los resultados de los diferentes métodos utilizados. Las diferencias encontradas entre ellos se deben a la específica forma de estimar la variabilidad y determinar el efecto combinado por cada método.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	1652	2561	460	1121	0	7
HE	1653	2561	462	1121	0	7
HS	1651	2563	473	1136	0	7
SJ	1653	2561	355	944	0	6
PM	1652	2561	461	1121	0	7
FE	1649	2565	532	1245	0	7

Tabla 4.18: Resultados del metaanálisis funcional en estudios *pareados* de tumores con procesos biológicos. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
DL	1693	2519	334	626	0	1
HE	1693	2519	333	620	0	1

Métodos	*E*	*C*	*Sig.E*	*Sig.C*	*Sig.LOR.E*	*Sig.LOR.C*
HS	1693	2519	339	646	0	1
SJ	1693	2521	216	420	0	1
PM	1693	2519	334	627	0	1
FE	1691	2521	463	1020	0	1

Tabla 4.19: Resultados del metaanálisis funcional en estudios *no pareados* de tumores con procesos biológicos. Modelos de efectos aleatorios y fijos para la estimación de la variabilidad del efecto medido.

Las Tablas 4.20 y 4.21 detallan los procesos biológicos significativos de mayor magnitud en el conjunto de los resultados obtenidos en el metaanálisis con el estimador *DerSimonian-Laird*.

ID	*Nombre*
GO:0006684	sphingomyelin metabolic process
GO:0035020	regulation of Rac protein signal transduction
GO:0043923	positive regulation by host of viral transcription
GO:0048340	paraxial mesoderm morphogenesis
GO:0071364	cellular response to epidermal growth factor stimulus
GO:0071599	otic vesicle development
GO:0090190	positive regulation of branching involved in ureteric bud morph.

ID	*LI 95 %*	*LOR*	*LS 95 %*
GO:0006684	-0.774	-0.624	-0.474
GO:0035020	-0.678	-0.517	-0.357
GO:0043923	-0.715	-0.56	-0.406
GO:0048340	-0.655	-0.5	-0.345
GO:0071364	-0.681	-0.511	-0.34
GO:0071599	-0.658	-0.516	-0.374
GO:0090190	-0.622	-0.504	-0.387

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
GO:0006684	0	0	11.58	0.772	0.077	0	0	1
GO:0035020	0	0	21.133	0.173	0.082	0.028	24.288	1.321
GO:0043923	0	0	16.739	0.403	0.079	0.005	4.416	1.046
GO:0048340	0	0	17.974	0.325	0.079	0.012	10.984	1.123
GO:0071364	0	0	26.551	0.047	0.087	0.051	39.738	1.659
GO:0071599	0	0	12.585	0.703	0.072	0	0	1
GO:0090190	0	0	6.129	0.987	0.06	0	0	1

Tabla 4.20: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con procesos biológicos. Estudios *pareados*.

En la Figura 4.23 se representa la función *sphingomyelin metabolic process* con una sobrerrepresentación significativa y de mayor magnitud en el grupo de controles respecto el grupo de enfermos, en estudios no pareados.

ID	*Nombre*	*LI 95 %*	*LOR*	*LS 95 %*
GO:0006684	sphingomyelin metabolic process	-0.706	-0.504	-0.302

ID	*P*	*FDR*	*QE*	*QEp*	*SE*	τ^2	I^2	H^2
GO:0006684	0	0	40.246	0.003	0.103	0.111	52.79	2.118

Tabla 4.21: Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con procesos biológicos. Estudios *no pareados*.

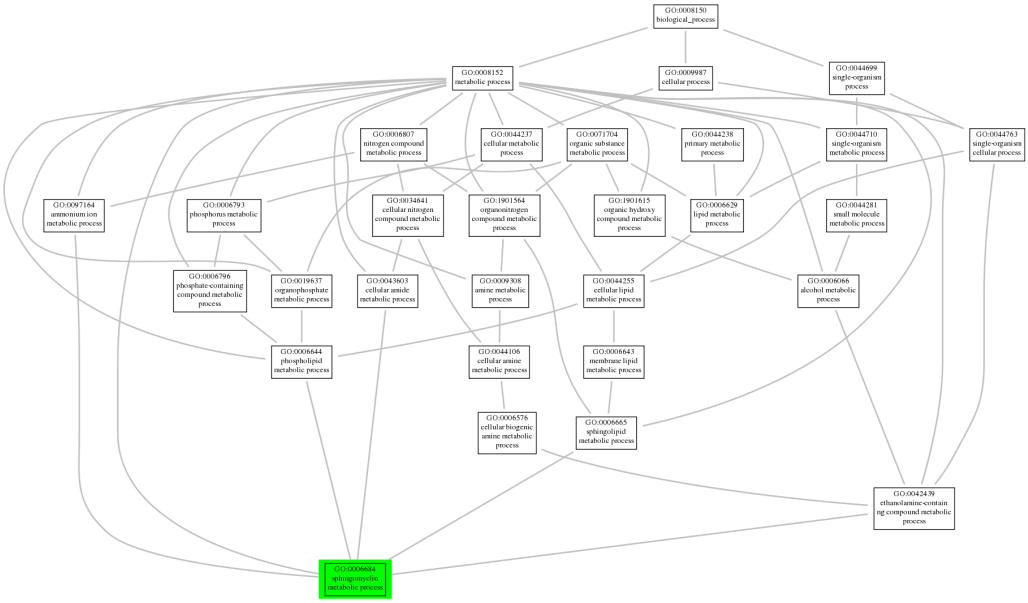

Figura 4.23: Procesos biológicos significativos y con mayor magnitud del efecto en estudios no pareados

Los resultados detallados para los diferentes estimadores del efecto combinado, en cada una de las funciones evaluadas de las bases de datos Gene

Ontology y también rutas KEGG y Reactome, están disponibles en los enlaces descritos del Aṕendice 2.

4.5. Discusión

La metodología descrita está orientada a un análisis de enriquecimiento de genes y metaanálisis donde la respuesta es *binaria* (sobrerrepresentación de una función en enfermos o sanos) y por ello la medida de efecto fue la razón de ventajas. Aunque este escenario es habitual en Genómica, la propuesta metodológica funcionaría también para respuestas continuas en otros contextos diferentes, ajustando algunas de las fases del método propuesto.

En los dos grupos de estudios se han utilizado diferentes bases de datos conocidas, sin embargo la metodología es aplicable a cualquier tipo de anotación funcional donde se describa la relación entre la unidad biológica de estudio y las funciones de interés. Esta flexibilidad posibilita la combinación de la información funcional no sólo en estudios transcriptómicos, también en otros tipos de estudios ómicos como metabolómicos, proteómicos o cualquier otro donde dispongamos de una anotación funcional.

La profusión de resultados funcionales suele ser un problema importante en la interpretación de los estudios genómicos. Por ello, la organización y presentación de resultados es un elemento crucial para que una metodología sea empleada y su uso sea optimizado. Nuestra propuesta integra la presencia de diferentes niveles de resultados, permitiendo revisar los aspectos funcionales comunes en varios estudios de interés, desde lo más general a lo más específico. De modo que con una primera y rápida inspección, es posible

determinar si hay efectos funcionales comunes en los estudios primarios. A continuación, podemos conocer cuáles son las funciones con mayor robustez en su significación y magnitud. Y por último revisar con detalle cada uno de los procesos que intervienen y evalúan el metaanálisis realizado para cada término funcional.

A lo largo de este capítulo, hemos presentado una metodología que integra información genómica, produciendo resultados que mejoran el conocimiento funcional, siendo de gran utilidad a la comunidad científica tanto en la confirmación de funcionalidades comunes ya conocidas, como en el descubrimiento de nuevas relaciones funcionales en un conjunto de estudios genómicos.

Capítulo 5

Discusión general y conclusiones

5.1. Discusión general

En Genómica y Biología Molecular, el creciente volumen de datos procedentes de tecnologías de alto rendimiento y la ingente cantidad de información accesible en bases de datos biológicos, han producido un escenario de datos masivos de gran interés para la comunidad científica, que precisa de metodologías que traten conjuntamente esta información.

Esta tesis se centra en el desarrollo de nuevos métodos de interpretación funcional de estudios genómicos, capaces de abordar estos problemas y que permitirán a los investigadores una mejor comprensión de los resultados de estos estudios en el contexto funcional.

Concretamente en el Capítulo 3 se describe una nueva propuesta metodológica para caracterizar funcionalmente los resultados en estudios de miARNs. La inexistencia de bases de datos que de forma sistematizada relacionen funciones con miARNs, supone un problema para la aplicación directa de los métodos habituales de enriquecimiento funcional. Para ofrecer una interpretación funcional en este tipo de estudios, transferimos la información de la expresión diferencial de los miARNs a genes, obteniendo una nueva ordenación en términos de inhibición y sobre esta ordenación aplicamos métodos de enriquecimiento de grupos de genes que nos proporcionan una interesante información sobre qué conjunto de genes comparten funciones y presentan un común patrón de inhibición de miARNs.

En diversos ámbitos científicos, se necesita conocer aspectos funcionales comunes en una serie de estudios genómicos. Para ofrecer soluciones en estos escenarios, en el Capítulo 4 presentamos un método de metaanálisis funcio-

nal capaz de responder a este tipo de preguntas: ¿qué procesos biológicos son comunes en diversos estudios transcriptómicos sobre resistencia a fármacos?, ¿qué vías de señalización aparecen sobrerrepresentadas en una enfermedad a lo largo de diferentes estudios metabolómicos?, ¿hay funciones compartidas en varios grupos de enfermedades que nos proporcionen una caracterización funcional global?, ¿conocemos una sobrerrepresentación común de funciones en un grupo de estudios de una enfermedad que nos oriente sobre una posible diana terapéutica?, ¿es posible detectar anómalos comportamientos de variabilidad en varios estudios genómicos con un mismo objetivo y diseño experimental?.

Este abordaje metodológico proporciona la combinación de la información de una serie de experimentos, posibilitando un marco adecuado para la evaluación de nuestro estudio de interés en relación con otros estudios con un diseño y características comunes. Asímismo también constituye una herramienta de valoración crítica de un conjunto de estudios genómicos.

Esperamos que nuestras propuestas metodológicas sean de utilidad a la comunidad científica y sirvan para aproximar de forma directa los resultados de estudios genómicos a la práctica investigadora y clínica, potenciando el valor traslacional de estos estudios.

5.2. Conclusiones

Las conclusiones de esta tesis se han resumido y organizado de acuerdo con los objetivos definidos en el Capítulo 2.

Los métodos desarrollados y presentados en esta tesis permiten combinar la información disponible en bases de datos biológicos y clínicos con la información experimental de estudios genómicos, facilitando a los investigadores la interpretación de los resultados obtenidos.

5.2.1. Generación de métodos de interpretación funcional para los resultados de estudios de miARNs

1. Hemos introducido un nuevo enfoque para la interpretación funcional de los estudios de miARN que está diseñado principalmente para conocer los efectos de la expresión diferencial de miARN en grupos de funciones o vías.
2. Nuestra propuesta se basa en el paradigma del análisis de grupos de genes, el cual extiende las metologías de análisis de sobrerrepresentación. Constituye un marco general aplicable a la mayoría de los escenarios genómicos, incluso cuando no hay (o hay demasiados) miARNs que se expresan diferencialmente, por lo tanto, este algoritmo elimina la arbitrariedad de los procedimientos *ad hoc* actuales. Pero lo más importante, es que nuestro algoritmo puede abarcar acontecimientos biológicamente relevantes que son abandonados por los demás, lo que representa

un paso adelante en la modelización de la regulación miARN-gen.

3. Nuestro enfoque considera los efectos de cancelación que surgen cuando un gen es interceptado por diferentes grupos de miARNs dentro de cada condición biológica y además es capaz de incorporar el efecto aditivo causado cuando varios miARNs débiles ejercen su influencia inhibidora en el mismo gen.
4. El algoritmo de análisis permite la inclusión de información relativa a los miARNs, genes y la relación entre ambos, de modo que fácilmente se puede ponderar en la etapa de transferencia o en el ajuste del modelo logístico.
5. Asímismo, la información genómica adicional se puede incorporar en el uso de nuestro marco multidimensional: por ejemplo, la integración de análisis de GSA de regulación de miARN y expresión génica es sencilla una vez que el problema de la transferencia se resuelve utilizando la metodología explicada anteriormente.
6. También la flexibilidad de nuestro enfoque y su implementación en un software de fácil utilización, hace su uso independiente del algoritmo de la expresión diferencial utilizado a nivel de miARN. Diferentes pruebas estadísticas o incluso fold-changes pueden ser empleados. Asímismo se pueden utilizar diversas bases de datos que incluyan la información de las dianas de los miARNs.

5.2.2. Desarrollo de métodos de metaanálisis funcional para estudios genómicos

1. Presentamos una metodología eficaz en la detección de alteraciones génicas relevantes en diversos escenarios, con capacidad de confirmar funcionalidades ya descritas y mostrar nuevas relaciones funcionales que pueden ser de interés para el inicio de nuevos estudios.
2. El método es flexible en el uso de información funcional: términos GOs, rutas de señalización o cualquier otra función que incorporemos en el el análisis de grupos de genes de cada estudio.
3. La metodología propuesta también es flexible en la selección de distintos modelos que estimen adecuadamente la variabilidad entre los estudios (diferentes modelos de efectos fijos y aleatorios).
4. Los estudios sobre los que hemos aplicado esta metodología se han centrado en la Transcriptómica de genes y miARNs pero los procedimientos son directamente aplicables a estudios de Metabolómica, Proteómica o cualquier otra área donde dispongamos de una anotación funcional vinculada al elemento biológico del estudio.
5. La combinación de los modelos logísticos multivariantes como método de enriquecimiento funcional y el posterior metaanálisis de los resultados obtenidos con estos modelos, permite la integración de información de las diversas variables explicativas utilizadas en los modelos logísticos también a nivel de metaanálisis.
6. El uso del metaanálisis funcional en estudios genómicos constituye una herramienta eficaz en la evaluación de diferentes estudios, constituyendo un recurso crítico de interés.

7. La perspectiva global que ofrece el metaanálisis funcional en la evaluación conjunta de patologías, proporciona una información funcional relevante en el diseño de nuevos abordajes clínicos y en la optimización del uso de recursos en el sistema sanitario.

5.3. Continuación del trabajo

Los métodos presentados en esta tesis han generado diversas ideas de continuación del trabajo, incluyendo los siguientes puntos de interés:

- Implementación de los métodos desarrollados en herramientas web y en el paquete *mdgsa* de Bioconductor con el objetivo de difundir y aproximar su uso a todos los perfiles de usuarios. Se incorporarán nuevos ejemplos de aplicaciones en diferentes áreas.

- Extensión de la metodología de enriquecimiento funcional en estudios de miARNs a los factores de transcripción, en el marco de la regulación génica.

- Desarrollo de métodos multivariantes en metaanálisis funcional que posibiliten la inclusión de información de otras variables de interés en el conjunto de los estudios seleccionados.

- Estrategias de integración que permitan la modelización de datos de distintas tecnologías para un mismo estudio y en un conjunto de estudios.

Apéndice 1. Análisis de enriquecimiento funcional en estudios de miARNs

Metodología de análisis de enriquecimiento funcional en estudios de miARNs: resultados detallados, programas y material suplementario.

- La implementación de esta propuesta metodológica se elaboró mediante la librería *mdgsa* http:// bioconductor.org/packages/mdgsa de *Bioconductor* (Gentleman & others 2004). El software utilizado fue *R* (R Core Team 2016). Todos los programas están disponibles en https://github.com/dmontaner-papers/gsa4mirna para la reproducibilidad de la investigación realizada.

- El material suplementario generado en el desarrollo metodológico está accesible en http://dmontaner-papers.github.io/gsa4mirna

Apéndice 2. Metaanálisis funcional en estudios genómicos

Metododología de metaanálisis funcional en estudios genómicos: resultados detallados, programas y material suplementario.

- La implementación de esta propuesta metodológica se elaboró mediante la librería *mdgsa* http:// bioconductor.org/packages/mdgsa de *Bioconductor* (Gentleman & others 2004). El software utilizado fue *R* (R Core Team 2016). Todos los programas están disponibles en https://github.com/fgardos-papers/funcmeta para la reproducibilidad de la investigación realizada.

- El material suplementario generado en el desarrollo metodológico está accesible en http://fgardos-papers.github.io/funcmeta

Referencias

Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J., 2004. FatiGO: A web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4), pp.578–580.

Al-Shahrour, F. et al., 2007. FatiGO+: A functional profiling tool for genomic data. integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic acids research*, 35(suppl 2), pp.W91–W96.

Alemán, A. et al., 2014. A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. *Nucleic acids research*, 42(W1), pp.W83–W87.

Alemán, A. et al., 2014. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic acids research*, 42(W1), pp.W88–W93.

Alonso, R. et al., 2015. Babelomics 5.0: Functional interpretation for new generations of genomic data. *Nucleic acids research*, p.gkv384.

Anders, S. & Huber, W., 2012. Differential expression of rNA-seq data at the gene level–the dESeq package. *Heidelberg, Germany: European Molecular*

Biology Laboratory (EMBL).

Ashburner, M. & others, 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1), pp.25–29.

Barrett, T. et al., 2007. NCBI gEO: Mining tens of millions of expression profiles - database and tools update. *Nucleic acids research*, 35(suppl 1), pp.D760–D765.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2), pp.281–297.

Belsey, D.A., Kuh, E. & Welsch, R.E., 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*, John Wiley.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, 57(1), pp.289–300.

Benjamini, Y. & Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), pp.1165–1188.

Bentley, D.R. & others, 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59.

Bieber, T., 2008. Atopic dermatitis. *New England Journal of Medicine*, 358(14), pp.1483–1494.

Bleazard, T., Lamb, J.A. & Griffiths-Jones, S., 2015. Bias in microRNA functional enrichment analysis. *Bioinformatics*, 31(10), pp.1592–1598.

Buyse, M. et al., 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*, 98(17), pp.1183–1192.

Campain, A. & Yang, Y.H., 2010. Comparison study of microarray meta-analysis methods. *BMC bioinformatics*, 11(1), p.408.

Carbonell, J. & others, 2012. A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome Med*, 4(8), p.62.

Carninci, P. et al., 2005. The transcriptional landscape of the mammalian genome. *Science*, 309(5740), pp.1559–1563.

Catalá-López, F. & Tobías, A., 2014. Metaanálisis de ensayos clínicos aleatorizados, heterogeneidad e intervalos de predicción. *Medicina Clínica*, 142(6), pp.270–274.

Catalá-López, F. & Tobías, A., 2013. Síntesis de la evidencia clínica y metaanálisis en red con comparaciones indirectas. *Medicina Clínica*, 140(4), pp.182–187.

Chen, M. et al., 2013. A powerful bayesian meta-analysis method to integrate multiple gene set enrichment studies. *Bioinformatics*, 29(7), pp.862–869.

Choi, J.K. et al., 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1), pp.i84–i90.

Church, G.M., 2006. Genomes for all. *Scientific American*, 294(1), pp.46–54.

Consortium, 1.G.P. & others, 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–1073.

Consortium, M. & others, 2010. The microArray quality control (mAQC)-iI study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8), pp.827–838.

Consortium, U. & others, 2014. Activities at the universal protein resource (uniProt). *Nucleic acids research*, 42(D1), pp.D191–D198.

Cook, R.D. & Weisberg, S., 1982. Residuals and influence in regression.

Cui, X., Churchill, G.A. & others, 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4), p.210.

DaVeiga, S.P., 2012. Epidemiology of atopic dermatitis: A review. In *Allergy and asthma proceedings*. OceanSide Publications, Inc, pp. 227–234.

DerSimonian, R. & Laird, N., 1986. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3), pp.177–188.

Dopazo, J., 2009. Formulating and testing hypotheses in functional genomics. *Artif Intell Med*, 45(2-3), pp.97–107.

Doxakis, E., 2010. Post-transcriptional regulation of alpha-synuclein expression by mir-7 and mir-153. *J. Biol. Chem.*, 285(17), pp.12726–12734.

Dubitzky, W., Granzow, M. & Berrar, D.P., 2007. *Fundamentals of data mining in genomics and proteomics*, Springer Science & Business Media.

Duval, S. & Tweedie, R., 2000a. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), pp.89–98.

Duval, S. & Tweedie, R., 2000b. Trim and fill: A simple funnel-plot–based

method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), pp.455–463.

D'haeseleer, P., 2005. How does gene expression clustering work? *Nature biotechnology*, 23(12), pp.1499–1501.

Evangelou, E. & Ioannidis, J.P., 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6), pp.379–389.

Flicek, P. et al., 2013. Ensembl 2014. *Nucleic acids research*, p.gkt1196.

Flicek, P. et al., 2014. Ensembl 2014. *Nucleic acids research*, 42(D1), pp.D749–D755.

Friedman, R.C. et al., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1), pp.92–105.

Fu, W. et al., 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431), pp.216–220.

Galbraith, R., 1988a. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in medicine*, 7(8), pp.889–894.

Galbraith, R., 1988b. Graphical display of estimates having differing standard errors. *Technometrics*, 30(3), pp.271–281.

Galbraith, R.F., 1994. Some applications of radial plots. *Journal of the American Statistical Association*, 89(428), pp.1232–1242.

García-Alonso, L. et al., 2012. Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments.

Nucleic acids research, 40(20), pp.e158–e158.

García-García, F. et al., 2016. Integrated gene set analysis for microRNA studies. *Bioinformatics*, p.btw334.

Gentleman, R.C. & others, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10), p.R80.

Glass, G.V., 1976. Primary, secondary, and meta-analysis of research. *Educational researcher*, pp.3–8.

Godard, P. & Eyll, J. van, 2015. Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. *Nucleic Acids Res.*, 43(7), pp.3490–3497.

González, I.F., Urrútia, G. & Alonso-Coello, P., 2011. Revisiones sistemáticas y metaanálisis: Bases conceptuales e interpretación. *Revista Española de Cardiología*, 64(8), pp.688–696.

Grützmann, R. et al., 2005. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 24(32), pp.5079–5088.

Gusev, Y., 2009. *MicroRNA profiling in cancer: A bioinformatics perspective*, Pan Stanford Publishing.

Hall, N., 2007. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9), pp.1518–1525.

He, L. & Hannon, G.J., 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5(7), pp.522–531.

Hedges, L.V., Gurevitch, J. & Curtis, P.S., 2008. The meta-analysis of response ratios in experimental ecology.

Herrero, J., Valencia, A. & Dopazo, J., 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2), pp.126–136.

Higgins, J. & Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11), pp.1539–1558.

Hong, F. & Breitling, R., 2008. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3), pp.374–382.

Huang, H.-C., Niu, Y. & Qin, L.-X., 2015. Differential expression analysis for rNA-seq: An overview of statistical methods and computational software. *Cancer informatics*, 14(Suppl 1), p.57.

Javitz, H.S. et al., 2002. The direct cost of care for psoriasis and psoriatic arthritis in the united states. *Journal of the American Academy of Dermatology*, 46(6), pp.850–860.

Joshi-Tope, G. & others, 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33(Database issue), pp.D428–432.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1), pp.27–30.

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, 8(2), p.e1002375.

Kodama, Y., Shumway, M. & Leinonen, R., 2012. The sequence read archive: Explosive growth of sequencing data. *Nucleic acids research*, 40(D1), pp.D54–D56.

Kozomara, A. & Griffiths-Jones, S., 2013. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, p.gkt1181.

Kumar, P., Henikoff, S. & Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the sIFT algorithm. *Nature protocols*, 4(7), pp.1073–1081.

Landrum, M.J. et al., 2014. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1), pp.D980–D985.

Lappalainen, I. et al., 2015. The european genome-phenome archive of human data consented for biomedical research. *Nature genetics*, 47(7), pp.692–695.

Lee, J.W. et al., 2005. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4), pp.869–885.

Lee, R.C., Feinbaum, R.L. & Ambros, V., 1993. The c. elegans heterochronic gene lin-4 encodes small rNAs with antisense complementarity to lin-14. *Cell*, 75(5), pp.843–854.

Lee, S.Y., Sohn, K.A. & Kim, J.H., 2012. MicroRNA-centric measurement improves functional enrichment analysis of co-expressed and differentially expressed microRNA clusters. *BMC Genomics*, 13 Suppl 7, p.S17.

Leidinger, P. et al., 2013. A blood based 12-miRNA signature of alzheimer disease patients. *Genome Biol*, 14(7), p.R78.

Li, N. et al., 2010. Whole genome dNA methylation analysis based on high throughput sequencing technology. *Methods*, 52(3), pp.203–212.

Lim, L.P. & others, 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027), pp.769–773.

Lockhart, D.J. et al., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13), pp.1675–1680.

Madeira, S.C. & Oliveira, A.L., 2004. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1), pp.24–45.

Marioni, J.C. et al., 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), pp.1509–1517.

McCain, J., 2006. The cancer genome atlas: new weapon in old war? *Biotechnol Healthc*, 3(2), pp.46–51.

McLendon, R. & others, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), pp.1061–1068.

Mi, G. & others, 2012. Length bias correction in gene ontology enrichment analysis using logistic regression. *PLoS ONE*, 7(10), p.e46128.

Minguez, P. et al., 2009. SNOW, a web-based tool for the statistical analysis

of protein–protein interaction networks. *Nucleic acids research*, 37(suppl 2), pp.W109–W114.

Montaner, D. & Dopazo, J., 2010. Multidimensional gene set analysis of genomic data. *PLoS ONE*, 5(4), p.e10348.

Montaner, D. et al., 2009. Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10, p.197.

Mootha & others, 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Gen.*, 34.

Mootha, V.K. et al., 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3), pp.267–273.

Morin, R.D. & others, 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, 18(4).

Mu, Z. et al., 2014. Molecular biology of atopic dermatitis. *Clinical Reviews in Allergy & Immunology*, pp.1–26.

Nagalakshmi, U. et al., 2008. The transcriptional landscape of the yeast genome defined by rNA sequencing. *Science*, 320(5881), pp.1344–1349.

Normand, S.-L.T., 1999. Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting. *Stat Med*, 18(3), pp.321–359.

Oh, S.J., Bae, D.S. & Suh, B.J., 2015. Synchronous triple primary cancers occurring in the stomach, kidney, and thyroid. *Annals of surgical treatment*

and research, 88(6), pp.345–348.

Pan, K.-H., Lih, C.-J. & Cohen, S.N., 2005. Effects of threshold choice on biological conclusions reached during analysis of gene expression by dNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25), pp.8961–8965.

Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4), pp.546–554.

Papapetrou, E.P. ando others, 2010. A genetic strategy for single and combinatorial analysis of miRNA function in mammalian hematopoietic stem cells. *Stem Cells*, 28(2).

Park, P.J., 2009. ChIP–seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10), pp.669–680.

Parkinson, H. et al., 2005. ArrayExpress—a public repository for microarray gene expression data at the eBI. *Nucleic acids research*, 33(suppl 1), pp.D553–D555.

Paule, R.C. & Mandel, J., 1982. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5), pp.377–385.

Poy, M.N. et al., 2004. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432(7014), pp.226–230.

R Core Team, 2016. *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Ramasamy, A. et al., 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, 5(9), p.e184.

Ramensky, V., Bork, P. & Sunyaev, S., 2002. Human non-synonymous sNPs: Server and survey. *Nucleic acids research*, 30(17), pp.3894–3900.

Rapp, S.R. et al., 1999. Psoriasis causes as much disability as other major medical diseases. *Journal of the American Academy of Dermatology*, 41(3), pp.401–407.

Rau, A., Marot, G. & Jaffrézic, F., 2014. Differential meta-analysis of rNA-seq data from multiple studies. *BMC bioinformatics*, 15(1), p.91.

Rhodes, D.R. et al., 2002. Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research*, 62(15), pp.4427–4433.

Robinson, M.D. & others, 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), pp.139–140.

Rodríguez, A.B.A. & Asfar, S.S., 2013. MicroRNAs circulantes:¿ nuevos biomarcadores en cáncer? *Eubacteria*, (32), pp.6–7.

Rothstein, H.R., Sutton, A.J. & Borenstein, M., 2006. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, John Wiley & Sons.

Salavert, F. et al., 2016. Actionable pathways: Interactive discovery of therapeutic targets using signaling pathway models. *Nucleic acids research*, p.gkw369.

Sartor, M.A. & others, 2009. LRpath: a logistic regression approach for iden-

tifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2), pp.211–217.

Schena, M., 1996. Genome analysis with gene expression microarrays. *Bioessays*, 18(5), pp.427–431.

Schmidt, F.L. & Hunter, J.E., 2014. *Methods of meta-analysis: Correcting error and bias in research findings*, Sage publications.

Schratt, G.M. et al., 2006. A brain-specific microRNA regulates dendritic spine development. *Nature*, 439(7074), pp.283–289.

Selbach, M. & others, 2008. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209), pp.58–63.

Shcherbata, H. et al., 2006. The microRNA pathway plays a regulatory role in stem cell division. *Cell Cycle*, 5(2), pp.172–175.

Shen, K. & Tseng, G.C., 2010. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10), pp.1316–1323.

Si, Y. et al., 2013. Model-based clustering for rNA-seq data. *Bioinformatics*, p.btt632.

Sidik, K. & Jonkman, J.N., 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*, 26(9), pp.1964–1981.

Sidik, K. & Jonkman, J.N., 2005. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2), pp.367–384.

Smyth, G.K., 2005. Limma: Linear models for microarray data. In *Bioinformatics and computational biology solutions using r and bioconductor*. Springer, pp. 397–420.

Spizzo, R. et al., 2010. MiR-145 participates with tP53 in a death-promoting regulatory loop and targets estrogen receptor-α in human breast cancer cells. *Cell Death & Differentiation*, 17(2), pp.246–254.

Spizzo, R. et al., 2009. RNA inhibition, microRNAs, and new therapeutic agents for cancer treatment. *Clinical Lymphoma and Myeloma*, 9, pp.S313–S318.

Stefani, G. & Slack, F.J., 2008. Small non-coding rNAs in animal development. *Nature reviews Molecular cell biology*, 9(3), pp.219–230.

Stenson, P.D. et al., 2012. The human gene mutation database (hGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics*, pp.1–13.

Sterne, J.A. & Egger, M., 2001. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of clinical epidemiology*, 54(10), pp.1046–1055.

Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15545–15550.

Suzuki, R. & Shimodaira, H., 2006. Pvclust: An r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), pp.1540–1542.

Tarazona, S. et al., 2012. NOIseq: A rNA-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, 17(B), pp.pp–18.

Viechtbauer, W., 2010. Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, 36(3), pp.1–48.

Wang, Y., Li, X. & Hu, H., 2011. Transcriptional regulation of co-expressed microRNA target genes. *Genomics*, 98(6), pp.445–452.

Xie, F. et al., 2014. High-throughput deep sequencing shows that microRNAs play important roles in switchgrass responses to drought and salinity stress. *Plant Biotechnology Journal*, 12(3), pp.354–366.

Zeggini, E. et al., 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5), pp.638–645.

Zhao, S. et al., 2014. Comparison of rNA-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1).