Métodos de Metaanálisis de Estudios Ómicos en Biomedicina

Jesús Gutiérrez Botella

Trabajo Final de Máster

Máster Universitario en Bioestadística

Tutor: Francisco García García
Tutor Académico: Antonio López Quílez

Facultad de Matemáticas
Universitat de València
Septiembre, 2020

"The human genome is a life written
in a book where every word has been written before.
A story endlessly rehearsed."

Johnny Rich. The Human Script: a Novel in 23 Chromosomes (2013)

RESUMEN

El desarrollo de las tecnologías de secuenciación masiva de ADN y de nuevas técnicas de análisis computacional de datos genómicos, ha potenciado un ingente aumento de información en las bases de datos de genomas y en los repositorios públicos de datos ómicos. En el campo de la biomedicina, se ha producido un significativo crecimiento de trabajos, cuyo objetivo es el estudio de los mecanismos moleculares de ciertas enfermedades con un enfoque genómico.

En concreto, en biomedicina, los estudios de expresión génica aportan una información muy valiosa, pues ayudan a comprender qué genes están implicados en la enfermedad por su variación en la expresión frente a los controles. Además, en ocasiones resulta de gran interés combinar la información que ofrecen varios estudios sobre la misma enfermedad, o sobre enfermedades parecidas, empleando técnicas de metaanálisis. Estas técnicas comprenden desde combinaciones sencillas de estudios por p-valores hasta modelos estadísticos más complejos que permiten modelizar también parámetros como la variabilidad intra e inter-estudios.

Por ello, en este trabajo se ha planteado analizar sets de datos de estudios de expresión génica de pacientes con diferentes grupos tumorales y con varios tipos de enfermedades de la piel, para su posterior metaanálisis a nivel de gen y a nivel de función mediante el uso de métodos para encontrar genes y funciones moleculares comunes y para comparar entre sí todos estos métodos.

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a Paco, mi tutor, toda su ayuda y su predisposición a ayudarme y a guiarme a lo largo de este trabajo. Ha sido un placer trabajar estos meses bajo su tutela.

También me gustaría agradecer a todos mis compañeros, ahora amigos, que he tenido la oportunidad de conocer en el máster. Me han apoyado siempre y han hecho de estos meses un tiempo muy especial.

Por último, a mi familia y amigos, que siempre han estado conmigo en cada paso y en cada decisión.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. La biomedicina

La **biomedicina** es una disciplina relativamente reciente que surge de la aplicación de los conocimientos de la genética y de la biología molecular modernas a los problemas médicos y de salud, y que se desarrolla con rapidez entre finales del siglo XX y el inicio de nuestro siglo.

La investigación biomédica tiene como característica principal la vocación de una traslación rápida del conocimiento generado en el laboratorio a los hospitales y los servicios sanitarios, lo que se denomina en estrategias *"bench to bedside"* (Curry, 2008).

Estas estrategias se utilizan en todas las fases clínicas de investigación de una enfermedad (Crabu, 2016): se intenta encontrar el mecanismo molecular de la patología, una manera de diagnosticarla (por marcadores moleculares, por diagnóstico genético, etc.), y un posible tratamiento que actúe sobre el mecanismo molecular defectuoso que lleva a desarrollar la enfermedad.

En la última década, la biomedicina se ha aplicado para investigar o tratar numerosas enfermedades de origen genético, como cánceres, enfermedades crónicas, minoritarias, etc.

Gracias a los últimos avances en las tecnologías de la información, ha habido un incremento notable de la información de estudios biomédicos almacenada en bases de datos públicas

(Costa, 2014). Además, el surgimiento de las tecnologías de secuenciación masiva de alto rendimiento (*arrays*, tecnologías NGS, etc.) el abaratamiento de su uso, y el desarrollo de las ciencias ómicas y la bioinformática han contribuido a que estos estudios, con datos genéticos, se hayan incrementado de manera exponencial.

1.2. El flujo de información genética en las células

Como se ha dicho anteriormente, los principales datos obtenidos en estudios biomédicos son datos de origen genético, ya sea de ácidos nucleicos, de proteínas o de metabolitos.

Para contextualizar, es necesario explicar el **dogma central de la biología molecular** (Figura 1.1). Este dogma describe cómo fluye la información desde los genes hasta las proteínas y los compuestos metabólicos, que son el producto final.

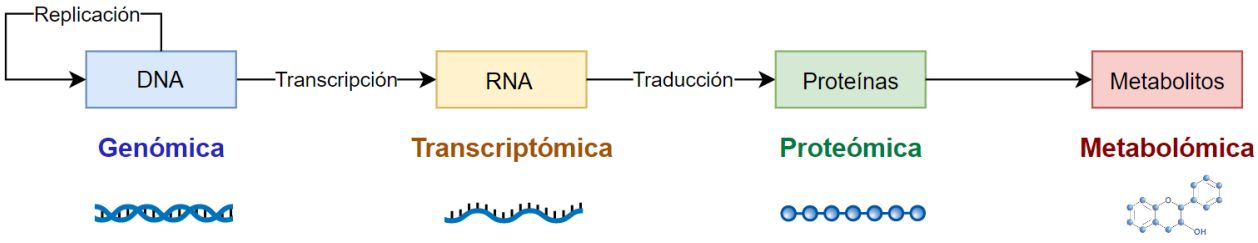

Figura 1.1 Dogma central de la biología molecular.

En las células eucariotas como las humanas, toda la información genética se almacena en el núcleo en forma de DNA (ácido desoxirribonucleico). El DNA es una macromolécula formada por la unión de nucleótidos en forma de cadena, que pueden ser de adenina (A), timina (T), citosina (C) y guanina (G), y que contiene toda la información acerca del fenotipo del organismo en pequeños fragmentos de DNA llamados **genes**.

La información de los genes se utiliza para fabricar proteínas en el citoplasma celular, por lo que se necesita una molécula que actúe como intermediaria y que pueda exportarse fuera del núcleo. Dicha molécula es el **RNA mensajero o RNAm** (ácido ribonucleico).

Por tanto, el primer paso para que un gen se exprese y origine una proteína es la síntesis del RNAm a partir del gen codificado en el genoma (DNA), y su exportación al citoplasma.

Este paso se denomina **transcripción** y genera transcritos, es decir, moléculas de RNA que representan genes. Estas moléculas de RNA son equivalentes a las moléculas de DNA en secuencia de nucleótidos, excepto por el detalle de que los nucleótidos de timina (T) cambian su composición química y son ahora de uracilo (U).

Una vez que la molécula de RNAm llega al citoplasma, es "traducida" a una proteína por unos pequeños orgánulos celulares llamados ribosomas. Estos ribosomas, en el proceso de la **traducción**, traducen la secuencia de nucleótidos a una secuencia de aminoácidos, subunidades por las que están formadas las proteínas. Este proceso de traducción hace uso de un "diccionario celular", llamado código genético, que contiene la información sobre a qué aminoácidos corresponden las secuencias de nucleótidos de RNAm.

Las **proteínas** en sí son las que representan el último nivel de información genética de los individuos, son las que realizan las funciones vitales (tienen actividad enzimática, son estructurales, algunas son hormonas, etc.). Las reacciones bioquímicas que tienen lugar en las células producen compuestos denominados **metabolitos**, que son sustancias que generan las enzimas durante el metabolismo.

Todo este flujo de información genética en la célula, desde el gen hasta su traducción a proteína, puede verse resumido en la Figura 1.2.

1.3. Datos ómicos en estudios biomédicos

En 1995 se completó la primera secuencia de un genoma completo de un organismo, *Haemophilus influenzae* (Fleischmann y col., 1995). Desde aquel hito histórico, se comenzaron a secuenciar los genomas de gran cantidad de organismos y se desarrollaron nuevas técnicas y tecnologías de secuenciación que abarataron los costes y el tiempo de análisis. Tanto es así, que en 2001 se finalizó el Proyecto Genoma Humano (PGH) (Schmutz y col., 2004), con la publicación de la secuencia del genoma humano completo.

En los últimos años, la genómica no ha parado de evolucionar y han aparecido un abanico de ciencias derivadas de ella: las ciencias "ómicas", que estudian diferentes aspectos de la organización a gran escala de la información genética de un organismo. El sufijo **–ómica** se utiliza en biología para referirse a la totalidad de genes, proteínas, metabolitos, y sus

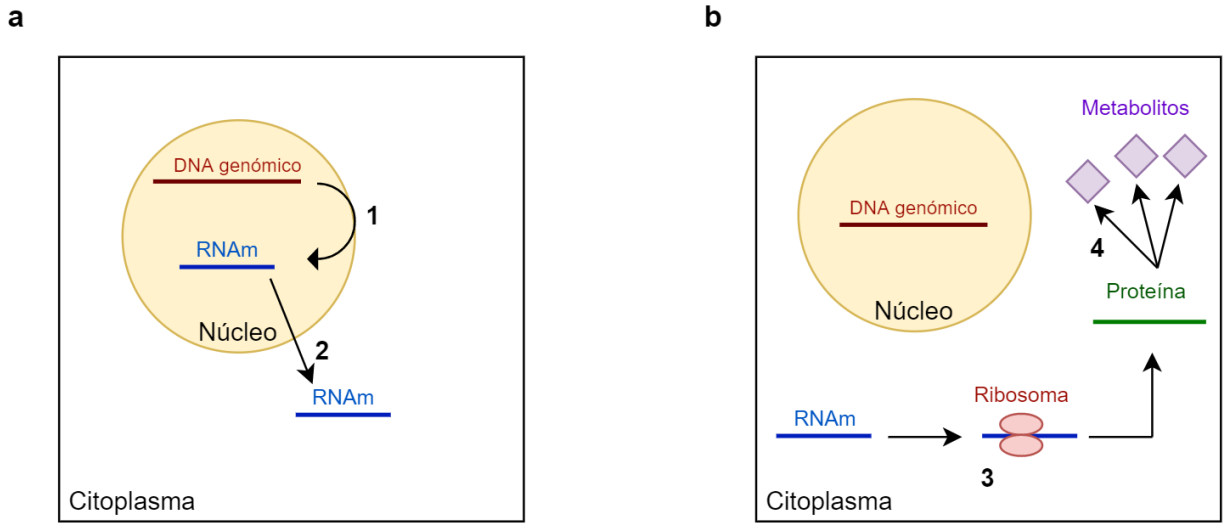

Figura 1.2 Flujo celular de información genética. (1) Los genes contenidos en el genoma son transcritos a RNA mensajero y (2) exportados del núcleo celular al citoplasma. (3) Aquí, son traducidos a proteína por los ribosomas. (4) Las proteínas actúan en las reacciones bioquímicas produciendo metabolitos.

relaciones entre ellos en un organismo.

Siguiendo el flujo de información genética que hemos descrito en el apartado anterior, podemos resumir las diferentes ómicas en genómica, transcriptómica, proteómica y metabolómica, entre otras.

1.3.1. Tecnologías de alto rendimiento

La extensión de las ómicas depende a su vez del desarrollo de **tecnologías de alto rendimiento** que permitan obtener datos masivos de las diferentes moléculas que se estudian.

El desarrollo de las nuevas tecnologías comerciales de alto rendimiento ha hecho que la obtención de información ómica de los organismos se abarate muy rápidamente. Dichas tecnologías difieren mucho en la técnica molecular que usan, pero todas permiten la secuenciación o la caracterización de material genético en grandes cantidades.

Las dos principales tecnologías de este tipo utilizadas en genómica y transcriptómica son los *microarrays* o micromatrices y las tecnologías de secuenciación masiva.

1.3.1.1. Microarrays

Un ***microarray*** es un pequeño dispositivo que contiene enganchados numerosos fragmentos de DNA a modo de cepillo. El DNA que contiene el array está en forma de cadenas simples que son capaces de unirse a otras moléculas de DNA o RNA complementarias, en este caso de la muestra del organismo, en un proceso denominado **hibridación**.

Así, es posible cuantificar la cantidad de material genético que se ha unido al *array*, y es posible detectar si el organismo contiene o no los genes del *microarray* en función de si ha habido o no hibridación. La detección, generalmente, se realiza a nivel de señales lumínicas o de fluorescencia.

1.3.1.2. Tecnologías de secuenciación masiva

Las **tecnologías de secuenciación masiva** (*Next Generation Sequencing*, NGS) son mucho más flexibles, ya que no se limitan a detectar o cuantificar el material genético que está en el *array*, sino que secuencian los nucleótidos de toda la muestra. La principal tecnología de este tipo es Illumina (Bentley y col., 2008), que permite la secuenciación del material genético en pequeños fragmentos de alrededor de 100 nucleótidos de longitud denominados "lecturas" o "*reads*".

En otras ómicas, como en metabolómica y en proteómica, no están presentes los nucleótidos. Para detectar metabolitos o proteínas se usan otras técnicas, como la **espectrometría de masas**, que identifica este tipo de compuestos por su relación masa/carga.

1.3.2. La genómica

Como tal, la genómica es la ciencia ómica más clásica y estudia aspectos que están relacionados con el genoma (DNA) de los organismos. Podríamos dividir la genómica en tres ramas fundamentales (Gutterson y col., 2004):

- **Genómica funcional.** Comprende la secuenciación de genomas, la predicción de sus genes y la asignación de funciones a los genes. Además, también engloba el estudio de

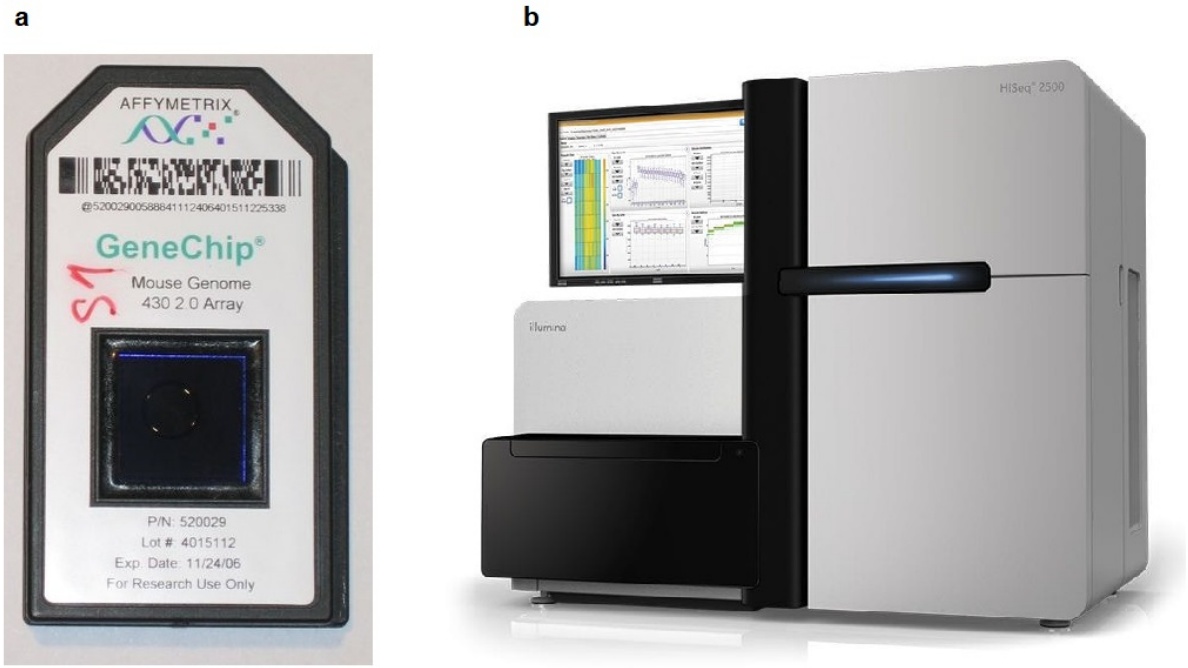

Figura 1.3 Principales tecnologías de alto rendimiento. Imágenes de (a) un *microarray* de Affymetrix para genoma de ratón, y de (b) un secuenciador HiSeq, de Illumina, uno de los más utilizados para la secuenciación de material genómico y transcriptómico. Fuentes: www.thermofisher.com/es/es/home/life-science/microarray-analysis.html y www.illumina.com/systems/sequencing-platforms.html.

regiones genómicas o mutaciones y su asignación a diversos fenotipos, como enfermedades, aspecto externo, metabolismo, etc.

- **Genómica estructural.** Tiene como objetivo el estudio de la estructura nucleotídica y tridimensional del genoma y de todos sus derivados, especialmente de las proteínas, basándose en su secuencia de aminoácidos. Multitud de enfermedades están causadas por defectos en estructuras tridimensionales de proteínas, por lo que tiene una aplicación importante en la biomedicina.
- **Genómica comparada.** Su finalidad es estudiar la evolución de las especies desde el punto de vista genómico, comparando sus genomas entre sí y evaluando las diferencias entre ellos.

1.3.3. La transcriptómica

El genoma de un individuo hace referencia a un material genético estático, presente en el núcleo de todas sus células somáticas, y que contiene toda su información genética.

Como se ha comentado anteriormente, son las proteínas las que verdaderamente van a ejercer las funciones que están codificadas en los genes. Es necesario copiar esa información en una molécula que no sea DNA, y que permita exportar la información fuera del núcleo para que se sinteticen las proteínas: el **RNA mensajero**, durante la transcripción.

En organismos complejos, cada fragmento de RNA contiene la información de un gen y se denomina también transcrito. El **transcriptoma** de un individuo hace referencia a la totalidad de transcritos que están expresándose en un momento y bajo unas condiciones determinadas (Wang y col., 2009).

Las células tienen mecanismos de regulación para controlar qué genes se transcriben o se expresan en cada momento: algunos de ellos se expresarán bajo unas condiciones y no bajo otras (por ejemplo, en muchas células cancerosas aparecen sobreexpresados numerosos genes que no lo están en células sanas, aunque en ambas sí formen parte del genoma).

Así, la **transcriptómica** estudia la expresión génica global de un individuo bajo unas condiciones determinadas. Dicha disciplina aporta mucha más información funcional que la genómica, ya que permite saber cuál es la respuesta celular ante un estímulo o un cambio de condición ambiental.

La mayoría de estudios de transcriptómica comprenden un paso de análisis de la expresión génica diferencial: un análisis estadístico que permite comparar qué genes se expresan entre dos o más condiciones, o incluso entre puntos temporales.

1.3.4. La proteómica y la metabolómica

La proteómica y la metabolómica suponen el estudio del último escalón funcional de los organismos, y aportan también información sobre la composición funcional de los individuos en un ambiente y bajo unas condiciones determinadas.

La **proteómica** se refiere a la caracterización a gran escala del conjunto de todas las

proteínas de un organismo, tejido o línea celular (Wasinger y col., 1995) a nivel de función y estructura tridimensional.

Por su parte, la **metabolómica** es una de las ómicas más recientes, y estudia el conjunto de compuestos metabólicos presentes en un organismo en una situación y en unas condiciones determinadas (Nielsen y col., 2005).

1.4. El trabajo con datos de transcriptómica

Aunque también existen estudios que trabajan con transcriptomas que utilizan ***microarrays*** como tecnología, estos estudios suelen basarse ya en la secuenciación masiva de RNA, lo que denominamos **RNA-Seq**.

Este tipo de análisis tiene, en general, unos objetivos principales: reconstruir, a partir de pequeños fragmentos de secuenciación que salen del secuenciador (lecturas), el transcriptoma de un organismo y estudiar la expresión génica a partir de él (Wang y col., 2009).

Los pasos principales a seguir en esta clase de estudios son (Li y col., 2011):

1. Extraer el RNA celular de las diferentes muestras.
2. Dividirlo en varios fragmentos de alrededor de 100 nucleótidos, denominados "lecturas".
3. Llevar dichas lecturas a secuenciar.
4. En caso de no disponer de un transcriptoma ya ensamblado y almacenado en una base de datos, utilizar las lecturas a modo de piezas de "puzle" para reconstruir el conjunto de transcritos del organismo y generar un transcriptoma *de novo*.
5. Cuantificar la expresión de cada gen. El nivel de expresión de cada gen se puede aproximar por la cantidad de lecturas que mapean, por similitud, contra cada gen.
6. Realizar los análisis pertinentes, según el diseño experimental, a nivel de gen o de función molecular.

El resultado de una cuantificación de la expresión génica es una matriz de conteos, en la que se obtiene, para cada transcrito y para cada muestra, el número de lecturas que han mapeado contra él. Esta matriz de conteos recibe el nombre de **matriz de expresión**, y será con la que se realicen los análisis posteriores (Figura 1.4).

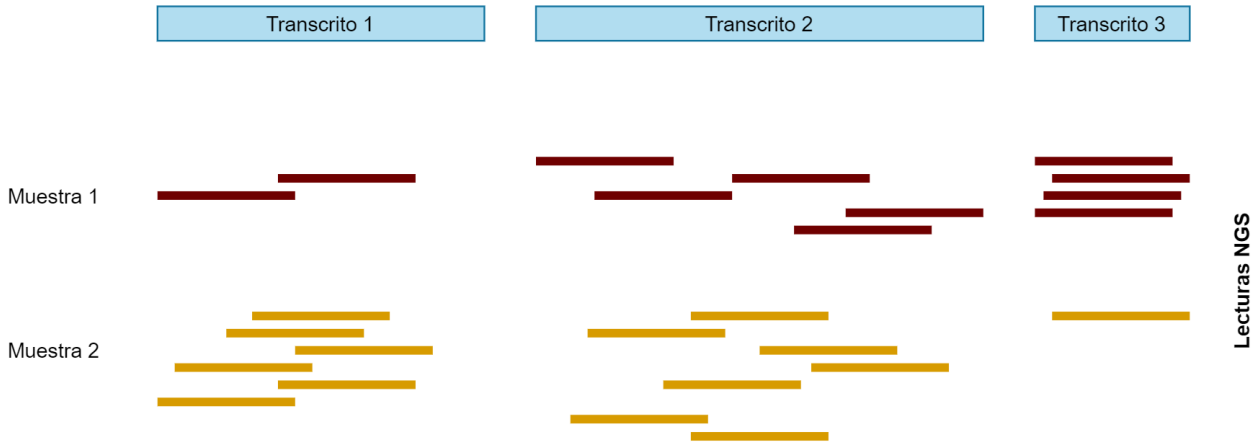

Figura 1.4 Construcción de la matriz de expresión en estudios de RNA-Seq. Una vez ensamblado el transcriptoma, todas las lecturas se mapean contra él, y se cuantifican las lecturas que caen sobre cada gen para cada muestra. En el ejemplo, la muestra 1 tiene 2 conteos para el primer transcrito, 5 para el segundo y 4 para el tercero; mientras que la segunda muestra 6, 7 y 1, respectivamente.

Los experimentos basados en *microarrays* construyen dicha matriz de expresión utilizando los genes que tienen en su matriz y contra los que hibridan los fragmentos de material genético. En este caso, la cuantificación de la expresión génica no se realiza por ”conteos” de lecturas que mapeen contra un determinado transcrito, sino por señales luminosas que indican la cantidad de ”hibridación” que ha habido contra cada transcrito del *array*.

1.4.1. Análisis diferencial de expresión génica

Una vez obtenida la matriz de expresión, uno de los objetivos más comunes es la búsqueda de **genes diferencialmente expresados**, es decir, genes que tienen diferencias en su nivel de expresión entre muestras, condiciones, tiempos u organismos (Soneson y col., 2013).

El hecho de que la construcción de la matriz de expresión se haga sobre conteos de lecturas en RNA-Seq, o sobre un análisis de imagen crudo en el caso de *microarrays*, hace que las muestras no sean comparables entre sí. Es necesario, por tanto, **normalizar** esta matriz.

En RNA-Seq, el objetivo principal de esta normalización es representar la abundancia “relativa” de las lecturas en cada gen. Pero existen muchos más problemas con este tipo de datos:

- La longitud de los genes es un factor decisivo a la hora de evaluar su nivel de expresión. Si un gen tiene el mismo nivel de expresión que otro, pero el doble de longitud, mapearán en el primero el doble de lecturas.
- Los genes con mucho nivel de expresión pueden provocar que el resto de genes vean sus conteos "enmascarados" por los primeros.

En *microarrays*, la tarea es más sencilla: la normalización consiste en eliminar el "ruido" de fondo detectado en la hibridación y escalar los valores de expresión de los diferentes *microarrays* a la misma escala. También se suelen emplear patrones de genes con concentración conocida o distribuciones de intensidad proporcionadas por el fabricante.

En RNA-Seq, se han propuesto numerosos métodos específicos de normalización, y se han implementado en algunos paquetes de Bioconductor, como *edgeR* o *DESeq2*, que utilizaremos en el presente trabajo, antes de proceder al test de expresión diferencial, como el TMM (*Trimmed Mean of M-values*), el RPKM (*Reads Per Kilobase Million*), etc.

La matriz de expresión descrita en el apartado anterior contiene datos de conteos, que pueden modelizarse generalmente por las distribuciones de Poisson o por una Binomial Negativa, que actualmente es la que más se utiliza al permitir un nivel de variabilidad mayor por incorporar un parámetro que controla la sobredispersión (Anders y col., 2010). Es sobre esta distribución sobre la que se realizarán los tests de expresión diferencial.

Los tests de expresión diferencial dan como resultado una lista de genes diferencialmente expresados entre condiciones, muestras, etc., con diversas métricas, de las que interesan:

- **Tamaño del efecto.** Cuantifica la relación entre la expresión del gen en una condición y en la otra. Se puede dar en forma de *fold-change*, de razón de *odds*, etc.
- **Estadístico de contraste** asociado a la expresión diferencial.
- ***P*-valor** asociado a dicho estadístico.
- ***P*-valor corregido.** Los estudios de este tipo suelen involucrar miles de genes, por lo que es necesario corregir el p-valor para comparaciones múltiples.

Los resultados suelen presentarse gráficamente en gráficas de volcán o en mapas de calor (*heatmaps*) (Figura 1.5).

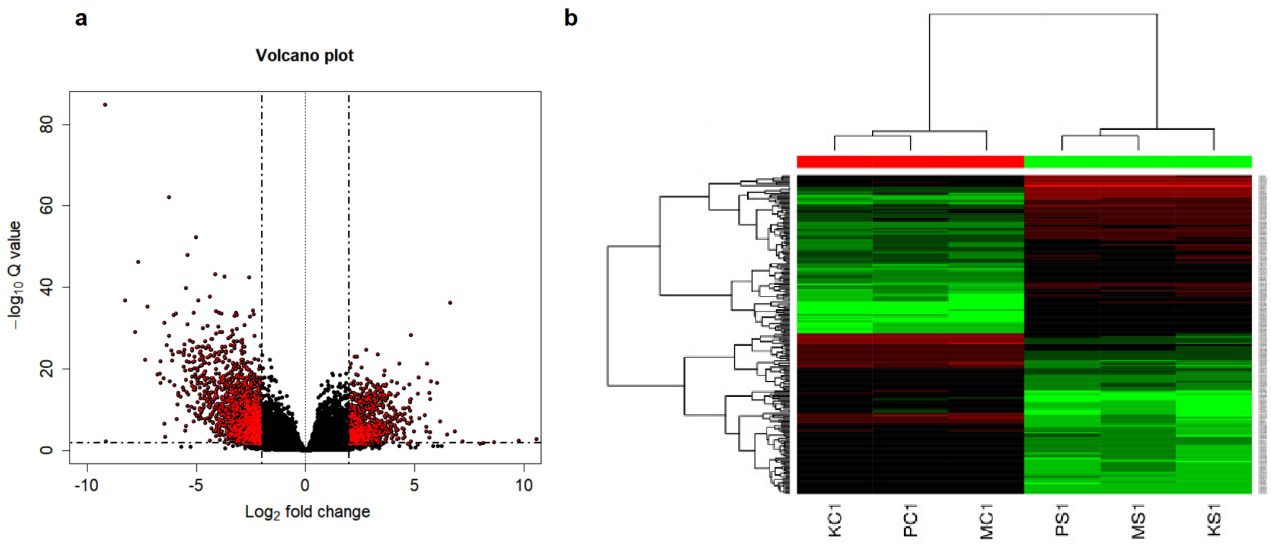

Figura 1.5 Representación de los resultados de un análisis de expresión diferencial. (a) *Volcano Plot*, donde cada punto representa un gen con su p-valor y su *fold-change* de la expresión diferencial. Los marcados en rojo son los considerados como diferencialmente expresados. (b) *Heatmap* con clustering jerárquico, donde se pueden ver los valores de expresión diferencial de cada gen en cada muestra en códigos de color: rojo implica sobreexpresión y verde infraexpresión.

1.4.2. Enriquecimientos funcionales

Los análisis de expresión diferencial proporcionan información sobre qué genes están involucrados en el proceso que se estudia. Sin embargo, no proporcionan información sobre la función molecular que realizan estos genes. Para incorporar esta información, es necesario incluir un paso de **anotación funcional**, es decir, asociarlos con funciones moleculares.

Estas anotaciones funcionales pueden hacerse a varios niveles y utilizando diferentes algoritmos bioinformáticos (con un alineamiento frente a secuencias de referencia, con predicción de la estructura proteica de los genes, etc.). Algunas de las anotaciones funcionales más importantes son:

- **Términos GO.** La *Gene Ontology* (GO) es una base de datos de funciones moleculares de los genes. Esta ontología provee un vocabulario que describe los genes y sus atributos, divididos en tres partes: (1) la función molecular del producto del gen, (2) el proceso biológico en el que participa dicho producto, y (3) la localización de dicho proceso en

un compartimento celular.
Cada término GO es una anotación, y consta de un ID único formado por identificadores alfanuméricos, y de una descripción (p.ej. el término GO con identificador GO:0007569 tiene de descripción *"cell aging"* y describe un proceso molecular).
Esta base de datos se revisa y amplía continuamente y actualmente ya cuenta con más de 38000 anotaciones GO diferentes (GOConsortium, 2004).

- **Rutas metabólicas KEGG.** El recurso KEGG (*Kyoto Encyclopedia of Genes and Genomes*) es un grupo de bases de datos centradas en contener datos de reacciones y rutas metabólicas, de enzimas, y datos de genomas para varios organismos.
  Las bases de datos KEGG sirven para entender funciones biológicas más generales, a nivel de organismo o a nivel celular, utilizando información genómica y molecular (Kanehisa y col., 2000).

Una vez los genes están anotados con estos términos GO o en rutas y reacciones metabólicas con KEGG, es posible determinar enriquecimientos significativos de estas anotaciones entre los genes diferencialmente expresados y los demás.

Este paso se denomina enriquecimiento funcional, y consiste en un test que puede ser de varios tipos (test exacto de Fisher o χ^2, entre otros) para detectar las funciones moleculares enriquecidas entre dos listas, generalmente, una lista de test, los genes diferencialmente expresados; y una de referencia, con todos los demás.

Hay dos tipos principales de enriquecimientos funcionales (Tarca y col., 2013):

1) Análisis de sobrerrepresentación (*Over Representation Analysis*, ORA)

Esta clase de métodos tienen que ver con análisis estadísticos sencillos basados en tablas de contingencia, donde se comprueba la asociación entre los genes diferencialmente expresados y su pertenencia a una categoría (por ejemplo, a un set de genes anotados con un mismo GO, pertenecientes a una misma ruta metabólica, etc.).

Algunas distribuciones típicas que se utilizan para los ORA son la hipergeométrica, la chi-cuadrado (χ^2), etc.

La principal limitación de los ORA es que sólo se testean aquellos genes que superen el

corte, generalmente arbitrario, para considerarse diferencialmente expresados (determinado p-valor y tamaño del efecto).

2) Análisis de grupos de genes (*Gene Set Analysis*, GSA)

Los GSA buscan resolver el principal inconveniente de los ORA, eliminando el paso de seleccionar los genes significativos y utilizando todos los genes del set de datos.

Esta clase de métodos asocian una puntuación para cada anotación (GO, KEGG, etc.) usando el nivel de expresión de todos los genes que la constituyen y, a partir de estas puntuaciones, se realiza el análisis para comprobar las anotaciones enriquecidas.

Hay muchos métodos englobados dentro del GSA, pero destaca principalmente el GSEA (*Gene Set Enrichment Analysis*) (Subramanian y col., 2005).

Todos los pasos que se siguen comúnmente en un experimento con datos de RNA-Seq aparecen resumidos en la Figura 1.6.

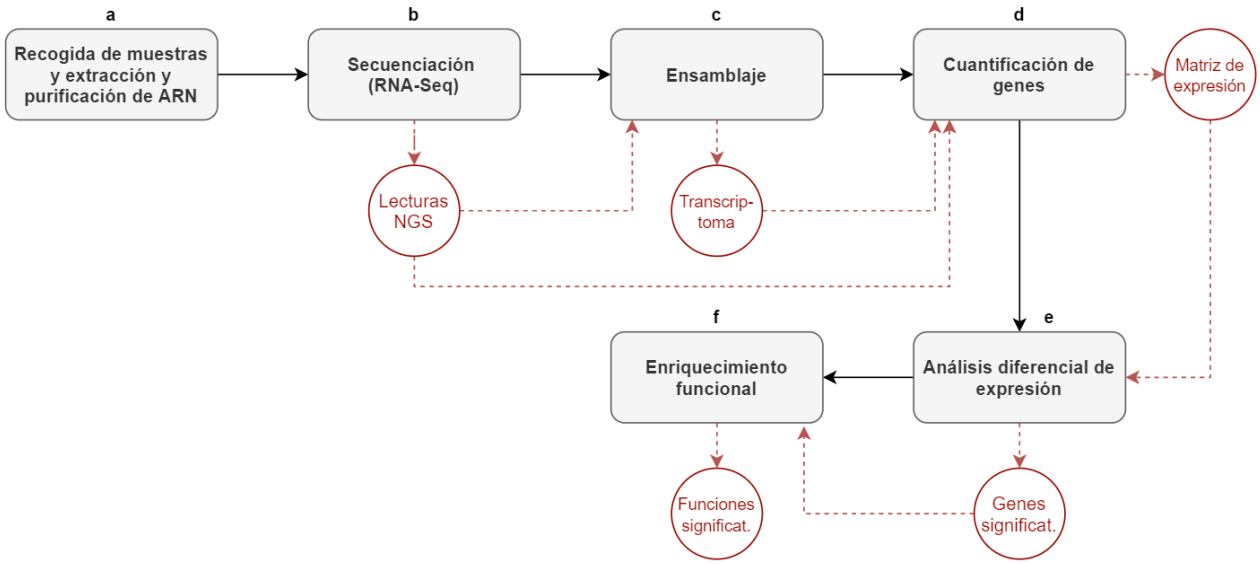

Figura 1.6 Esquema del proceso completo de un análisis de RNA-Seq. Los pasos principales son la extracción y secuenciación de RNA (a-b), el ensamblaje de un transcriptoma con las lecturas generadas (c), el mapeo para construir una matriz de expresión (d), el análisis diferencial de expresión entre condiciones experimentales (e) y un enriquecimiento funcional (f).

1.5. El metaanálisis

El **metaanálisis** es un conjunto de herramientas estadísticas que permiten sintetizar la información de varios estudios con datos u objetivos similares. En general, los metaanálisis requieren la recopilación del tamaño del efecto y una estimación del error de dicho efecto.

Esta síntesis generalmente resulta en una mejor estimación del efecto del tratamiento, y permite controlar las fuentes de variación entre estudios.

El metaanálisis sigue las siguientes fases:

1. **Formulación de objetivos.** Se define qué tipo de estudios se quieren metaanalizar y para qué.
2. **Cuantificación de los efectos.** Se miden los efectos que se evalúan. En experimentos de expresión diferencial, mediremos el nivel de cambio en la expresión expresión y los p-valores asociados, ambos en escala continua.
3. **Recopilación de los estudios.** Se definen los criterios de selección de los estudios que se incluirán en el metaanálisis (calidad de los estudios, metodología, tipo de datos, etc.) y se recogen los que cumplen con los requisitos. La recogida de estudios en caso de expresión génica se realiza desde diferentes repositorios como GEO (*Gene Expression Omnibus*), ArrayExpress o TCGA (*The Cancer Genome Atlas*).
4. **Combinación de los estudios.** Existen diferentes técnicas estadísticas para combinar los efectos o los p-valores, generalmente dependiendo de la heterogeneidad que presentan los estudios.
   a. Combinación de p-valores o de rangos: representan métodos sencillos de combinación que no tienen en cuenta la variabilidad o heterogeneidad entre estudios.
   b. Modelo de Efectos Fijos: cuando se asume que el efecto producido por la exposición es constante de estudio en estudio. Es el más usual cuando se estudia toda la población de artículos.
   c. Modelo de Efectos Aleatorios: se asume una distribución al azar a lo largo de distintos estudios. Cuando sólo se estudia un estrato de la población de artículos. Este último método estará orientado cuando exista heterogeneidad.

5. **Análisis de heterogeneidad.** Este análisis permite valorar hasta qué punto los resultados que provienen de diferentes estudios pueden ser resumidos en una única medida.

6. **Análisis de sensibilidad.** El análisis de sensibilidad pretende estudiar la influencia de cada uno de los estudios en la estimación global del efecto y, por lo tanto, la robustez o estabilidad de la medida final obtenida.

7. **Cálculo de la significación estadística o tamaño del efecto combinado.**

8. **Obtención de conclusiones.**

Los metaanálisis, sin embargo, tienen algunas limitaciones. Entre ellas, podemos destacar que cuando los estudios sean muy heterogéneos, los resultados pueden ser muy difíciles de interpretar. Además, puede existir un sesgo de publicación: sólo se publican aquellos estudios que muestran resultados significativos acordes a la hipótesis de los estudios, y generalmente los resultados negativos se suelen omitir.

1.5.1. Metaanálisis de estudios de transcriptómica

El metaanálisis en estudios de expresión génica sirve para combinar estudios cuyo resultado sean genes diferencialmente expresados o funciones moleculares enriquecidas entre condiciones similares. En nuestro caso, serán pacientes con varias enfermedades relacionadas contra controles.

El metaanálisis se puede realizar a dos niveles:

- **Metaanálisis a nivel de gen.** El objetivo es combinar la información de los genes diferencialmente expresados entre varios estudios.
- **Metaanálisis a nivel de función.** En este tipo de metaanálisis, los genes deben anotarse funcionalmente y se debe realizar un enriquecimiento funcional de cada estudio. El metaanálisis se realiza sobre las funciones moleculares enriquecidas para obtener las más importantes en la combinación de los estudios.

Capítulo 2

Objetivos

El **objetivo principal** de este trabajo es revisar y evaluar los principales métodos de metaanálisis de datos ómicos, que permitan conocer las estrategias de análisis más adecuadas a cada escenario de integración de estudios biomédicos.

Para llevarlo a cabo, se alcanzarán los siguientes **objetivos específicos**:

- Revisar y comparar los métodos de metaanálisis a nivel de gen.
- Revisar los métodos de enriquecimiento funcional.
- Revisar y comparar los métodos de metaanálisis a nivel de función.
- Evaluar conjuntamente todos los métodos descritos.
- Aplicar estos métodos a dos sets de datos de estudios de transcriptómica: un set de datos de tumores y otro de enfermedades dermatológicas.

Capítulo 3

Material y Metodología

3.1. Sets de datos

Todos los métodos que se describen a continuación se aplicarán a dos sets de datos diferentes de transcriptómica por ser el tipo de ómica más utilizada actualmente en caracterización funcional: un set de datos de tumores y otro de enfermedades dermatológicas. Todos contienen datos de expresión génica de varias enfermedades relacionadas entre sí y provienen de estudios diferentes. Cabe destacar que las técnicas de metaanálisis que se describirán a continuación pueden aplicarse sobre otros datos ómicos como metabolómica o proteómica.

En ambos estudios, los nombres de los genes vienen de la base de datos de genes humanos de Ensembl[1], que ya se encuentran funcionalmente anotados y caracterizados. El proceso de anotación funcional, es decir, la asignación de los términos GO a cada gen se llevó a cabo desde la base de datos de BioMart[2].

3.1.1. Expresión génica en tumores

Según el Instituto Nacional del Cáncer (NCI), el cáncer es un conjunto de enfermedades generalmente multifactoriales en las que unas células tumorales se dividen sin control y

[1] www.ensembl.org
[2] www.ensembl.org/biomart/martview/c9433082bc619934adb82cfc5ba1fe76

pueden invadir tejidos cercanos. El conocimiento de los patrones comunes de expresión génica en este tipo de alteraciones es muy útil en Biomedicina, ya que puede desvelar los mecanismos moleculares comunes que comparten estas enfermedades.

El proyecto TCGA (*The Cancer Genome Atlas*) es un proyecto iniciado en 2005 y que tiene como finalidad, a través de la genómica, caracterizar los cambios moleculares responsables de la aparición del cáncer[3].

Desde este proyecto se han seleccionado 17 estudios de expresión génica en diferentes tipos de tumores, de estudios diferentes (Tabla 3.1). Se pretende metaanalizar todos los estudios para encontrar patrones de expresión diferencial similares en todos ellos.

El set de datos consiste en la matriz de expresión de los conteos crudos de todos los genes y los metadatos de cada estudio, que indican el diseño experimental. En todos los estudios se plantea un diseño pareado, en bloques, de tal manera que para cada paciente se recoge una muestra de tejido sano ("control") y una muestra del tejido canceroso ("caso").

3.1.2. Expresión génica en enfermedades dermatológicas

La psoriasis y la dermatitis son dos enfermedades dermatológicas muy comunes con signos clínicos que suelen ser muy molestos en los pacientes que las padecen: enrojecimientos, molestias en la piel o alternancia de brotes y remisiones, entre otros síntomas.

Como en el caso anterior, conocer los mecanismos genéticos y moleculares que están detras de cada una de las enfermedades y qué relación tienen a escala bioquímica puede ayudar a desarrollar mecanismos de detección y nuevos tratamientos que sean más precisos, dentro del campo de la Biomedicina.

Este set de datos de expresión se obtuvo de estudios que emplearon *microarrays* para los análisis (Tabla 3.2) y se extrajeron de la base de datos *Gene Expression Omnibus*, GEO (Edgar y col., 2002). Sin embargo, en este caso, el punto de partida para cada estudio es el resultado del análisis diferencial de expresión: para cada gen, si está o no diferencialmente expresado por su *fold-change* y por su p-valor corregido.

El diseño experimental en este caso es simple, con datos de casos y de controles.

[3] www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Estudio	Tipo de cáncer	Muestras
BLCA	Carcinoma de vejiga	38
BRCA	Carcinoma invasivo de mama	216
CHOL	Colangiocarcinoma (cáncer de las vías biliares)	18
COAD	Adenocarcinoma de colon	82
ESCA	Carcinoma esofágico	16
HNSC	Carcinoma de células escamosas de cabeza y cuello	86
KICH	Carcinoma renal de células cromófobas	46
KIRC	Carcinoma renal de células claras	144
KIRP	Carcinoma renal de células papilares	62
LIHC	Carcinoma de hígado	100
LUAD	Adenocarcinoma de pulmón	112
LUSC	Carcinoma de células escamosas de pulmón	98
PRAD	Adenocarcinoma de próstata	102
READ	Adenocarcinoma de recto	18
STAD	Adenocarcinoma de estómago	54
THCA	Carcinoma de tiroides	116
UCEC	Carcinoma endometrial del cuerpo uterino	46

Tabla 3.1 Descripción del conjunto de estudios sobre cáncer del TCGA.

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 U133A Plus*
GSE41662	PS	*Affymetrix Human Genome U133A Plus 2.0*
GSE41663	PS	*Affymetrix Human Genome U133A Plus 2.0*
GSE34248	PS	*Affymetrix Human Genome U133A Plus 2.0*
GSE13355	PS	*Affymetrix Human Genome U133A Plus 2.0*
GSE30999	PS	*Affymetrix Human Genome U133A 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	DA	*Affymetrix Human Genome U133 Plus 2.0*

Tabla 3.2 Descripción del conjunto de estudios sobre enfermedades dermatológicas de GEO. PS: Psoriasis · DA: Dermatitis Atópica.

3.2. Preprocesado de los datos

Desde los datos crudos de conteos de los estudios del TCGA, y partiendo de la expresión diferencial de los estudios de enfermedades dermatológicas, se realizarán los pasos contenidos en la Figura 3.1.

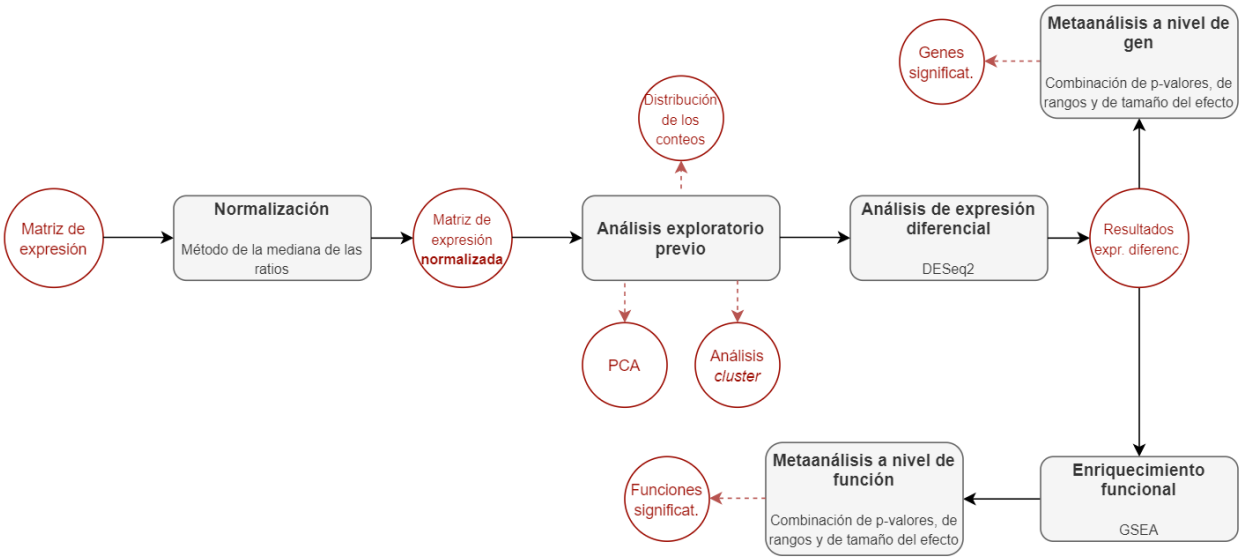

Figura 3.1 Metodología general de metaanálisis de estudios de transcriptómica.

Tanto la normalización como el análisis de expresión diferencial siguen diferentes estrategias según si los datos son procedentes de RNA-Seq o de *microarrays*. Como este preprocesado solo es necesario para el set de datos del TCGA, cuya tecnología es RNA-Seq, solo se describirán estos pasos de preprocesado para datos de este tipo.

El primer paso del preprocesado es el tratamiento de los datos crudos obtenidos por RNA-Seq para construir una matriz de expresión. Como se ha comentado en el apartado introductorio, estos pasos comprenden el mapeo de las lecturas contra un transcriptoma de referencia y la cuantificación de las lecturas que mapean contra cada transcrito.

La matriz resultante con su diseño experimental (es decir, a qué grupo pertenece cada muestra), será el punto de inicio del análisis posterior. Dicho análisis comprenderá la normalización de los datos de conteo, un análisis exploratorio de los sets de datos y el análisis de expresión diferencial. Todos estos pasos se llevarán a cabo con las librerías de R base y con el paquete *DESeq2*, de Bioconductor (Love y col., 2014).

3.2.1. Modelo para datos de conteo de RNA-Seq

La matriz de expresión es una matriz de conteo K, que tiene como filas los diferentes genes i y como columnas las muestras j. Entonces, cada K_{ij} representa el número de lecturas que han mapeado contra el gen i en la muestra j.

Los datos de conteo K_{ij} se describen con un modelo lineal generalizado (GLM) de la familia binomial negativa, con media μ_{ij} y con un parámetro de dispersión α_i:

$$K_{ij} \sim \mathrm{NB}\left(\mu_{ij},\ \alpha_i\right) \tag{3.1}$$

La media μ_{ij} se toma como una cantidad q_{ij}, que es proporcional a la cantidad de lecturas de RNA que mapean contra el gen en la muestra, siempre escalado por un factor de normalización por muestra s_j:

$$\mu_{ij} = s_j\, q_{ij} \tag{3.2}$$

3.2.1.1. Cálculo de los factores de normalización

Los experimentos de RNA-Seq llevan asociados una enorme variabilidad entre las diferentes muestras. Uno de los principales problemas es la diferencia entre la profundidad de secuenciación en cada una de ellas, es decir, que la cantidad de lecturas que han sido correctamente mapeadas en cada una de las muestras puede ser muy diferente y eso hace que no se puedan comparar entre sí.

Aunque existan otros sesgos técnicos que pueden tenerse en cuenta para la normalización (proporción de GC, longitud media de los genes, etc.), tener en cuenta la profundidad de secuenciación suele ser suficiente y resulta más sencillo.

Para estimar los factores de normalización s_j se utiliza el método de la **mediana de las ratios**. En primer lugar, se crea una muestra de "pseudo-referencia" para cada gen i, K_i^R, igual a la media geométrica de los conteos de ese gen en todas las m muestras. Después, se calcula para cada gen en cada muestra la ratio entre el conteo del gen i en la muestra j (K_{ij}) y el valor de K_i^R, siempre que para el gen i $K_i^R \neq 0$.

Para cada muestra j, se obtiene su s_j como la mediana de los valores de las ratios obtenidos para todos los genes dentro de la muestra:

$$s_j = \underset{i:\, K_i^R \neq 0}{\text{mediana}} \frac{K_{ij}}{K_i^R}, \text{ siendo } K_i^R = \left(\prod_{j=1}^{m} K_{ij} \right)^{\frac{1}{m}} \tag{3.3}$$

3.2.2. Análisis exploratorio

En primer lugar y antes de realizar el análisis diferencial de expresión, es necesario realizar un análisis exploratorio sobre los datos de cada estudio. Para cada uno, realizaremos un análisis de componentes principales (PCA), un *clustering* jerárquico y un *boxplot* sobre la distribución de los conteos normalizados de cada muestra.

3.2.2.1. Distribución de los conteos

Como hemos dicho en el apartado anterior, para hacer las muestras comparables entre sí es necesario tener una cantidad de lecturas mapeadas (profundidad de secuenciación) similar entre muestras. Esto se consigue gracias a la normalización descrita anteriormente.

Es necesario comprobar que la distribución de conteos por gen sea similar entre las muestras una vez se ha normalizado la matriz de expresión. Por ello, se recurre a un *boxplot* donde se puede comprobar que la distribución por muestra sea similar.

3.2.2.2. Análisis de Componentes Principales

Un **Análisis de Componentes Principales (PCA)** es un método no supervisado de reducción de la dimensión de un set de datos, y se utiliza para visualizar, a modo exploratorio, en un conjunto reducido de variables la máxima cantidad de información (James y col., 2013).

Imaginemos que queremos representar n observaciones de p características $X_1, X_2, ..., X_p$. Cada una de estas observaciones está en un espacio p-dimensional, pero no todas estas dimensiones son igualmente relevantes. El PCA intenta reducir este espacio en menos dimensiones, haciendo que éstas sean lo más "relevantes" posible. Esta relevancia se mide por la varianza

que cada dimensión es capaz de explicar de los datos originales.

Dichas nuevas dimensiones o variables se denominan **componentes principales**. La primera componente principal de este set $X_1, X_2, ..., X_p$ es la combinación lineal normalizada de las p variables

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p \tag{3.4}$$

con la mayor varianza, entendiendo por normalización $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. El vector ϕ_1 definido por $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$ es el vector de la componente principal 1, y sus direcciones son las de máxima varianza de los datos.

En un set de datos $\mathbf{X}$, de $n \times p$, las variables se estandarizan para que las columnas tengan media 0 restando la media de la columna a cada dato. Ahora se busca la combinación lineal de los valores de las variables de la forma

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip} \tag{3.5}$$

que tenga la mayor varianza muestral sujeto a la restricción anterior $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. En otras palabras, el vector de *loadings* de la primera componente principal resuelve el problema de optimización siguiente, que es simplemente maximizar la varianza de los n valores de z_{i1}:

$$\text{máx} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1}x_{ij} \right)^2 \right\} \text{ sujeto a } \sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{3.6}$$

El vector $z_{11}, z_{21}, ..., z_{n1}$ es el vector de *scores* de la primera componente principal.

Después de calcular la primera componente principal Z_1, el objetivo es encontrar la segunda componente principal Z_2. Esta segunda componente es la combinación lineal de las variables con máxima varianza **que estén incorreladas** con Z_1. Esto quiere decir que el cálculo es igual pero se añade además la restricción de que el vector de loadings ϕ_2 sea ortogonal a la dirección de ϕ_1.

Además, es posible calcular la varianza de los datos que explica cada componente principal. La varianza total del set de datos si las variables han sido centradas en 0 es la siguiente:

$$\sum_{j=1}^{p} \mathrm{Var}\left(X_j\right) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \tag{3.7}$$

Y la varianza explicada por la componente m-ésima es la siguiente:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2 \tag{3.8}$$

Por tanto, la **proporción de varianza explicada por la componente m-ésima** ($\mathrm{PVE_m}$) es:

$$PVE_m = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} \tag{3.9}$$

Los resultados del PCA se utilizarán para representar en el espacio de las dos primeras componentes principales, las de mayor varianza, las observaciones y comprobar cómo se agrupan según si proceden de un caso o de un control.

3.2.2.3. *Clustering* jerárquico

El análisis *cluster* se refiere a un conjunto de técnicas para encontrar subgrupos o *clusters* dentro de un set de datos según su similitud en cuanto a las observaciones. En un set de datos de n observaciones con p variables, el análisis *cluster* clasifica las n observaciones en grupos más o menos homogéneos en cuanto a los valores de sus variables (James y col., 2013).

El *clustering* jerárquico presenta la clasificación en forma de **diagrama de árbol**, que se construye iniciándose por las hojas, que se van combinando hasta el final.

Para este análisis, partimos de un set de datos $n \times p$. El siguiente paso es el cálculo de una **matriz de disimilitudes o de distancias**, de $n \times n$, que contiene la distancia entre cada par de muestras. Generalmente, se calcula la distancia euclídea entre cada par de puntos:

$$d\left(x, y\right) = \sqrt{\sum_{v=1}^{p} \left(x_v - y_v\right)^2} \tag{3.10}$$

Empezando desde el final del dendograma, el algoritmo considera cada punto n como un *cluster* independiente. Los dos *clusters* más similares entre sí se fusionan en un solo grupo, teniendo ahora $n-1$ grupos. En la siguiente etapa, los dos *clusters* más similares se juntan de nuevo, teniendo ahora $n-2$. El algoritmo repite este proceso hasta quedarse con un único *cluster*, de forma iterativa.

El concepto de disimilitud entre dos puntos se determina en la matriz anterior, pero no la distancia entre dos grupos con más de un punto dentro de cada uno. Es necesario un método de aglomeración para definir la disimilitud entre dos grupos de observaciones. Hay muchos métodos de este tipo (*complete, average, single, etc.*), pero todos intentan determinar una medida para este tipo de situaciones.

El resultado de la clasificación se puede visualizar, como hemos dicho en la primera parte, en forma de dendograma y en forma de mapa de calor, donde se representa en un dendograma cómo ha sido la clasificación y en el mapa las similitudes entre muestras.

3.2.3. Análisis de expresión diferencial

El análisis de expresión diferencial se lleva a cabo ajustando un GLM con un *link* logaritmo en base 2. Con el modelo anterior:

$$\begin{gathered} K_{ij} \sim \mathrm{NB}\left(\mu_{ij},\ \alpha_i\right); \qquad \mu_{ij} = s_j\, q_{ij} \\ \log_2 q_{ij} = \textstyle\sum_r x_{jr}\beta_{ir} \end{gathered} \tag{3.11}$$

La **matriz de diseño** viene representada por x_{jr}, donde indicaremos si la muestra pertenece al grupo de los casos o a los controles, o si existe diseño por bloques; y los coeficientes del GLM vienen dados por los parámetros β_{ir}. Este GLM se ajusta para cada gen y se obtiene una estimación del *fold-change* en base logarítmica (*logFC*) y de su error estándar.

Este *logFC* es el tamaño del efecto, esto es, la cantidad de cambio de expresión que existe entre la referencia y el tratamiento (en nuestro caso, entre las muestras caso y las controles):

$$\mathrm{logFC} = \log_2 \tfrac{E_{caso}}{E_{control}} \tag{3.12}$$

En un análisis comparativo comprobamos la **hipótesis nula** de que no existen cambios en la expresión génica entre las dos condiciones, es decir, que el *logFC* entre ambas es 0.

Para ello, cada *logFC* estimado se divide entre su error estándar, obteniendo un estadístico z y se aplica un test de Wald para obtener el p-valor asociado a la expresión diferencial.

Por tanto, desde un análisis de expresión diferencial obtendremos la lista de genes con un *fold-change* como medida del tamaño del efecto, su desviación estándar, el estadístico de contraste y un p-valor y un p-valor corregido. Para considerar un gen como **diferencialmente expresado** deberá tener un *fold-change* mayor que 2 (sobreexpresado) o menor que -2 (infraexpresado), y un p-valor ajustado menor a 0,05.

La visualización de los resultados del análisis de expresión diferencial a modo general se realizará a través de un *volcano plot* para cada estudio, de acuerdo con la Figura 1.5 (a).

3.2.3.1. Ajuste de *p*-valores

Este tipo de experimentos de transcriptómica generalmente involucran un número enorme de genes, por lo que hay que aplicar algún tipo de ajuste a los p-valores crudos obtenidos en el test de expresión diferencial. Estos p-valores son corregidos utilizando el procedimiento de *Benjamini-Hochberg* para controlar la tasa de falsos descubrimientos (**FDR**) a un nivel Q. El Q para este tipo de tests se fija en 0,05 (Benjamini y col., 1995).

El primer paso, es ordenar los p-valores en orden ascendiente, desde un p-valor pequeño al más grande: $P_1, P_2, ..., P_m$ y se les asigna un ranking: el p-valor más pequeño tiene un orden $i = 1$, el siguiente $i = 2$, el mayor $i = m$.

Después se compara cada p-valor individual con el valor crítico de Benjamini-Hochberg, $\frac{i}{m}Q$, donde i es el orden del p-valor en la lista, m el número total de tests y Q la tasa de falsos descubrimientos que se establezca como criterio.

El p-valor más grande que sea $p \leq \frac{i}{m}Q$ se declara como significativo, y todos los p-valores más pequeños también se declaran como significativos, incluso los que no sean menores que el valor crítico de Benjamini-Hochberg. Los p-valores corregidos (o q-valores), se calculan como $q = \frac{p_m}{i}Q$.

3.3. Enriquecimiento funcional

Los **métodos GSA** para enriquecimiento funcional, o métodos de Análisis de Grupos de Genes, consideran experimentos con datos de perfiles de expresión génica a nivel de todo el genoma. Las muestras generalmente pertenecen a dos clases, A o B (en nuestro caso, controles y casos). Los genes se ordenan de acuerdo con estas clases con cualquier métrica con sentido biológico. Esta métrica puede tener en cuenta el tamaño del efecto, el p-valor o cualquier otra propiedad: una combinación de las anteriores, el estadístico de contraste de la expresión diferencial, etc.

Con esta métrica se construye una lista ordenada, x, de todos los genes en el experimento.

Dado un set de genes c, por ejemplo, los genes que están anotados con un mismo término GO o con una misma ruta KEGG, el objetivo de este análisis es determinar si hay relación entre la pertenencia a una clase u otra según la distribución de c en x para cada set de genes.

Existen numerosas metodologías para llevar a cabo un análisis de enriquecimiento GSA, siendo la más común el GSEA (Subramanian y col., 2005), pero la aproximación que se utilizará aquí está basada en **modelos de regresión logística** (Montaner y col., 2010).

La base, como se ha descrito anteriormente, son los resultados de la expresión diferencial. Para cada set de genes c (genes anotados con el mismo GO), el modelo empleado es el siguiente:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \tag{3.13}$$

La variable dependiente, y, se define como 1 para los genes que se encuentren en c y como 0 para cualquier otro gen. La variable explicativa x es una métrica de la expresión diferencial, en este caso se empleará el estadístico de contraste.

Así, π representa la proporción de genes anotados con un GO específico (pertenecen a c) a un determinado valor de x.

Por tanto, $\pi/(1-\pi)$ representa el *odd* para un gen con valor x de su estadístico de contraste de la expresión diferencial sea miembro de esta categoría c. Si el *log(odds)* se incrementa o disminuye con x, la conclusión es que el término GO que estamos considerando está

asociado con la expresión diferencial y, por tanto, se puede considerar como **enriquecido**.

En el modelo, α representa el intercepto y β la pendiente, que se interpreta como los cambios en los logaritmos de los *odds* para cada unidad de crecimiento del estadístico de contraste, x. Entonces:

- Si $\beta > 0$, el término GO de interés está "enriquecido" o "sobrerrepresentado" entre los genes sobreexpresados.
- Si $\beta < 0$, el término GO está "infrarrepresentado" entre los genes sobreexpresados.
- Si $\beta = 0$, el término GO no guarda relación con la expresión diferencial.

Para determinar, en cada c, el valor de β se calcula el p-valor asociado a la hipótesis nula de que $\beta = 0$ frente a la alternativa de $\beta \neq 0$ con un **test de Wald**, cuyo estadístico de contraste W es $W = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}}\right)^2$, utilizando como $\hat{\beta}$ el estimador máximo verosímil de β, y como $s_{\hat{\beta}}$ el error estándar de $\hat{\beta}$.

Para $\beta = 0$, el estadístico W sigue una distribución χ^2 con un grado de libertad.

Los p-valores obtenidos para cada término GO c se ajustan después para controlar la tasa de falsos descubrimientos (FDR), según la metodología de Benjamini-Hochberg descrita anteriormente.

Para realizar el paso de enriquecimiento funcional, es necesario realizar la anotación de los genes con términos GO, recurriendo a BioMart. También es necesario obtener la jerarquía de los GOs y sus descripciones correspondientes. Para todo esto utilizaremos el paquete de Bioconductor *go.db*[4].

El análisis de enriquecimiento, una vez anotados los genes correctamente, se llevará a cabo utilizando el paquete de Bioconductor *mdgsa* (Montaner y col., 2010).

En la tabla final de resultados de enriquecimiento sólo se tendrán en cuenta los términos GOs relacionados con **procesos biológicos (BP)** para reducir los resultados, y para cada GO se obtendrá una medida del tamaño del efecto (logaritmo del *odds-ratio*, LOR) con su error estándar, un p-valor y un p-valor corregido.

[4]Carlson M (2019). GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.8.2

Los términos GO que superen el corte de p-valor corregido $<$0,05 seran considerados como enriquecidos, ya sea infra o sobrerrepresentados (LOR menor o mayor que 0, respectivamente).

3.4. Metaanálisis

Los **metaanálisis** comprenden una gran variedad de técnicas que permiten integrar la información de varios estudios en un solo análisis.

En este caso, partiremos de los resultados obtenidos en el análisis de expresión diferencial (metaanálisis a nivel de gen): tamaño del efecto ($logFC$), error estándar del tamaño del efecto ($SE(logFC)$) y p-valor; o los obtenidos en el enriquecimiento funcional (metaanálisis a nivel de función): tamaño del efecto (LOR), error estándar ($SE(LOR)$) y p-valor. Algunos de estos métodos son:

- **Métodos de combinación de *p*-valores.** Para cada gen o función molecular, estos métodos combinan los p-valores de la expresión diferencial o del enriquecimiento funcional. La combinación de p-valores da lugar a una nueva medida de la significación de cada gen o función, un ***p*-valor combinado**, que puede obtenerse por diversos métodos (Rau y col., 2015).

- **Métodos de combinación del tamaño del efecto.** Estos métodos suelen ser los más apropiados para realizar un metaanálisis, ya que consideran modelos más complejos para obtener un **tamaño del efecto combinado** de cada gen o función molecular. Estos modelos pueden resumirse en los de efectos fijos, en el caso de que los estudios no muestren apenas variabilidad, y en los de efectos aleatorios, si se tienen estudios más heterogéneos (Choi y col., 2003).

- **Métodos de combinación de rangos.** Estos métodos combinan también los tamaños del efecto usando un método más sencillo. Una vez obtenido el tamaño del efecto tras la expresión diferencial o el enriquecimiento funcional, los genes o funciones moleculares son ordenados según este tamaño del efecto y se les asigna una puntuación (su posición en el ranking). Dicha puntuación es la que posteriormente se combina para realizar el metaanálisis (Kolde y col., 2012).

Para realizar los metaanálisis se utilizaron los paquetes *metaMA* (Marot, Foulley y col., 2009) y *metafor* (Viechtbauer, 2010) para los métodos de combinación del tamaño del efecto, *metaRNASeq* (Rau y col., 2015) para la de p-valores y *RankProd* (Hong y col., 2006) para la combinación de rangos.

3.4.1. Métodos de combinación de p-valores

Como hemos descrito anteriormente, los resultados de una expresión diferencial o de un enriquecimiento funcional de un único estudio llevan asociados un p-valor por gen o término GO. Por tanto, de un resultado de s estudios, tendremos una lista p_{gs}, es decir, un p-valor crudo para cada gen o término GO g en cada estudio s.

Para los metaanálisis, vamos a considerar dos métodos de combinación de p-valores: la normal inversa y el método de combinación de Fisher.

Estos dos procedimientos asumen que, bajo la hipótesis nula, cada vector de p-valores se distribuye de manera uniforme.

3.4.1.1. Método de la normal inversa o ponderado

Para cada gen o término GO g, se define el estadístico N_g como

$$N_g = \sum_{s=1}^{S} w_s \phi^{-1} \left(1 - p_{gs}\right), \tag{3.14}$$

donde p_{gs} es el p-valor crudo obtenido para un gen o GO g en el análisis de expresión diferencial o enriquecimiento funcional del estudio s. ϕ es la función de distribución acumulada de la distribución normal estándar y w_s un vector de pesos.

Este vector de pesos proporciona un peso para cada estudio en el cálculo del estadístico de contraste, de tal manera que se puede dar más o menos peso a un estudio dependiendo de sus características. Aunque hay muchas posibilidades, es razonable ponderar los estudios según el número de muestras que contengan según la metodología de (Marot y Mayer, 2009):

$$w_s = \sqrt{\frac{\sum_c R_{cs}}{\sum_g \sum_c R_{cg}}}, \tag{3.15}$$

donde $\sum_c R_{cs}$ es el número de muestras en el estudio s. Esto hace que los estudios con un mayor número de muestras biológicas tengan un mayor peso a la hora del cálculo del p-valor combinado. Bajo la hipótesis nula, el estadístico N_g sigue una distribución normal estándar. Por ello, se puede realizar un test sobre la cola de la derecha.

3.4.1.2. Método de Fisher o no ponderado

Para este método, el estadístico F_g para cada gen o término GO g se define como:

$$F_g = -2 \sum_{s=1}^{S} \ln\left(p_{gs}\right), \tag{3.16}$$

con la principal desventaja de que no se puede proporcionar pesos a los estudios cuando se utiliza este método de metaanálisis.

El estadístico F_g sigue, bajo la hipótesis nula, una distribución χ^2 con $2S$ grados de libertad, siendo s el número de estudios.

El resultado de este tipo de metaanálisis es una lista de genes o términos GO con el p-valor resultado de combinar todos los p-valores de todos los estudios.

3.4.1.3. Consideraciones para los métodos de combinación de *p*-valores

Los métodos previamente descritos de combinación de p-valores tienen una serie de problemas en los casos de metaanálisis a nivel de gen, es decir, cuando se trata con datos de experimentos de RNA-Seq (Rau y col., 2015).

La principal asunción para poder aplicar estos métodos es que los p-valores de todos los genes del análisis diferencial de expresión de los estudios deben estar uniformemente distribuidos bajo la hipótesis nula de los dos estadísticos descritos N_g y F_g.

Esta hipótesis no siempre se satisface con datos de este tipo. Es frecuente observar una

acumulación de p-valores cercanos a 1 debido a genes con bajo número de conteos (muy poco expresados en todas las muestras), por lo que debe haber un paso de filtrado de los genes con menos expresión. Este filtrado se realiza por defecto con el paquete *DESeq2*.

Además, y teniendo en cuenta que la técnica de combinación de p-valores identifica genes significativos utilizando únicamente el p-valor y no el tamaño del efecto, es necesario identificar genes con patrones de expresión conflictivos (esto es, genes infraexpresados en una condición en unos estudios y sobreexpresados en otros). Estos genes deben ser eliminados también del análisis o, al menos, su patrón debe ser estudiado a posteriori. En el presente trabajo, se eliminarán aquellos genes que presenten comportamientos de este tipo en más del 20 % de los estudios. También se tendrá en cuenta esta última consideración en los metaanálisis a nivel de función.

Ambos métodos requieren también una corrección por comparaciones múltiples del p-valor combinado descrita por Benjamini-Hochberg.

3.4.2. Métodos de combinación del tamaño del efecto

Como hemos visto en apartados anteriores, la principal métrica para demostrar expresión diferencial o enriquecimiento es una medida del tamaño del efecto: $logFC$ para genes y LOR para funciones, entre la condición de referencia y la de contraste.

Para este tipo de metaanálisis denotamos con μ la media general de este tamaño del efecto, que estamos interesados en combinar entre estudios. Los efectos observados en los k estudios independientes se denotan por $y_i = 1, 2, 3, ..., k$.

El modelo general es el siguiente:

$$\begin{aligned} y_i &= \theta_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0,\ s_i^2\right) \\ \theta_i &= \mu + \delta_i, \quad \delta_i \sim N\left(0,\ \tau^2\right) \end{aligned} \tag{3.17}$$

- La variación entre estudios viene dada por τ^2.
- El error de muestreo en el estudio i-ésimo viene dado por s_i^2.
- μ, por tanto, denota el tamaño de la expresión diferencial para cada gen, o el tamaño

del efecto del enriquecimiento para cada término GO, en todos los estudios.

Un **modelo de efectos fijos (FEM)** asume que las diferencias observadas entre los tamaños del efecto suceden únicamente por la variabilidad inherente al muestreo, y por tanto considera que $\tau^2 = 0$. En consecuencia, el modelo para un FEM se quedaría únicamente como $y_i \sim N\left(\mu,\ s_i^2\right)$.

Por su parte, un **modelo de efectos aleatorios (REM)** considera que cada tamaño del efecto parte de una distribución con una media específica de cada estudio θ_i y con una varianza s_i^2. Además, cada θ_i viene de una distribución con una media general μ y una varianza τ^2. Por tanto, así se obtiene como modelo principal $y_i \sim N\left(\theta_i,\ s_i^2\right)$ y $\theta_i \sim N\left(\mu,\ \tau^2\right)$.

En el metaanálisis por combinación del tamaño del efecto, más complejo que el de combinación de p-valores, el resultado obtenido contiene una medida del tamaño del efecto media μ con su intervalo de confianza (IC) al 95 %, el estadístico τ^2 y un p-valor y un p-valor corregido. Se considerarán como genes o funciones significativas aquellas con un efecto cuyo IC no incluya al 0, y cuyo p-valor corregido sea menor a 0,05.

3.4.2.1. Análisis de heterogeneidad, sensibilidad y estudios influyentes

Estos metaanálisis combinan el efecto medido, es decir, la magnitud de la expresión diferencial de genes o del enriquecimiento de funciones moleculares, entre todos los estudios. Es necesario realizar un **análisis de heterogeneidad** de los estudios para decidir el análisis más apropiado. Un FEM es un modelo mucho más sencillo tanto a la hora de estimar como a la hora de interpretar los resultados, pero no siempre es el adecuado.

Para determinar qué tipo de modelo es apropiado para combinar una serie de estudios se puede aplicar un test para evaluar la homogeneidad de los efectos entre los estudios: el **test de Cochran** (Cochran, 1954). Esto es equivalente a comprobar la hipótesis nula de que $\tau^2 = 0$.

El estadístico de contraste es Q, y se calcula como:

$$Q = \sum w_i \left(y_i - \hat{\mu}\right)^2, \tag{3.18}$$

donde $w_i = s_i^{-2}$ y $\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$.

Bajo la hipótesis nula, Q sigue una distribución χ^2_{k-1} siendo k el número de estudios. El rechazo de esta hipótesis nula significa un valor alto de Q, lo que supone un alto nivel de heterogeneidad entre estudios. Esto es indicativo de que el modelo más apropiado es un modelo de efectos aleatorios (REM). En caso de quedarse con la hipótesis nula, no se puede rechazar que los estudios sean homogéneos y se puede aplicar uno de efectos fijos (FEM).

El test de Cochran suele acompañarse de otros métodos, generalmente gráficos, que ayudan a comprobar la presencia de heterogeneidad y de posibles sesgos en los estudios. Estos gráficos reciben el nombre de **gráficos de embudo**.

Estos gráficos representan en el eje X la magnitud del efecto frente a una medida de precisión en el eje Y. Cada punto representa un estudio. Cuando no existan sesgos y los estudios sean homogéneos, se espera que los puntos se distribuyan en forma de embudo y se sitúen siempre dentro de la región de confianza.

Entre los estadísticos para medir la heterogeneidad, hasta ahora se ha descrito el τ^2. Sin embargo, existen otras métricas que ayudan a cuantificar esta heterogeneidad (Higgins y col., 2002). Algunas son I^2 y H^2, detalladas a continuación:

$$H^2 = \frac{Q}{k-1} \qquad I^2 = \frac{H^2}{H^2 - 1}, \tag{3.19}$$

donde Q representa el estadístico de Cochran y k el número de estudios.

- H^2 representa el cociente entre la variabilidad total y la variabilidad en el muestreo: cuando τ^2 sea 0, H^2 será 1.
- I^2 estima la relación entre la variabilidad entre estudios y el total de la variabilidad.

También es necesario verificar si existen **observaciones influyentes**. En algunos estudios, la magnitud del efecto estimado y de la variabilidad puede producir una gran influencia en el metaanálisis. Para evaluar estas observaciones, es necesario excluir estos estudios del análisis y observar cambios en el modelo ajustado para valorar su influencia. Esto se realiza con diversas métricas: residuos estandarizados, distancias de Cook, estimadores de τ^2, ...

Se utiliza una estrategia *leave-one-out* de validación cruzada para el cálculo de estas estadísticas, excluyendo cada vez un estudio i del metaanálisis y calculando, en concreto, las siguientes métricas (Viechtbauer y Cheung, 2010):

- **Residuos estandarizados.** Este valor indica en cuántas desviaciones estándares cambia la media estimada del tamaño del efecto al excluir el estudio i-ésimo para el ajuste del modelo.
- **DFFITS** Comparte significado con el anterior, pero en una escala relativa.
- **Distancia de Cook.** Es la distancia de Mahalanobis entre el set completo de valores predichos en el modelo incluyendo al estudio i-ésimo y sin incluirlo.
- **Ratio de las covarianzas.** Representa el determinante de la matriz de varianzas-covarianzas de las estimaciones de los parámetros incluyendo el estudio i-ésimo, dividido por el determinante sin incluir el estudio. Un valor inferior a 1 indica que la eliminación del estudio i hace que la estimación de los coeficientes del modelo sea más precisa, y viceversa.
- **Estimación de τ^2.** Es la cantidad estimada de heterogeneidad residual medida con τ^2 en el modelo sin el estudio i-ésimo.
- **Estimación de Q.** Es la cantidad estimada de heterogeneidad residual medida con el estadístico de Cochran Q en el modelo sin el estudio i-ésimo.
- **DFBETAS.** Indica cuántas desviaciones estándares cambian los coeficientes estimados del modelo después de excluir el estudio i-ésimo del modelo.
- **Weights.** Comparte significado con el anterior, pero en una escala relativa.

En general y bajo el supuesto de que no haya estudios influyentes, la representación gráfica de dichas métricas habiendo excluido un estudio cada vez no debería presentar ninguna observación atípica. Si la exclusión de un mismo estudio influye en gran medida para todos los genes o todas las anotaciones funcionales, debería excluirse del análisis al ser considerado dicho estudio como influyente.

El **análisis de sensibilidad** sigue la misma estrategia que el anterior: se utiliza una estrategia de validación cruzada por *leave-one-out* para comprobar si, al excluir un estu-

dio i, cambia la dirección del tamaño del efecto, su magnitud o su significación en genes diferencialmente expresados o funciones moleculares enriquecidas. En concreto:

- **Tamaño del efecto.** Se comprobará cómo influye la eliminación del estudio i-ésimo en relación al tamaño del efecto estimado comparándolo con el estimado con todos los estudios. Para ello, se realizará un test t para comparar, para cada gen o función molecular, todos los valores estimados del tamaño del efecto por *leave-one-out* con el del set de datos completo.
- **Significación.** Se comprobará cómo cambian los p-valores de significación de genes o funciones moleculares del metaanálisis cuando se excluye un estudio. Para ello, y siguiendo la estrategia del *leave-one-out*, se contará el número de veces que los genes o funciones detectados como significativos siguen teniendo un $p < 0{,}05$ tras eliminar uno a uno los estudios.

3.4.3. Combinación de rangos

Aunque en este caso particular utilicemos el $logFC$ o los LOR como métricas para este método de metaanálisis, se puede utilizar cualquier métrica relacionada con la expresión génica que tenga sentido biológico. Al utilizar estas métricas, se puede considerar esta parte también como una combinación del tamaño del efecto, pero con una estrategia más sencilla que la descrita anteriormente, utilizando un test no paramétrico (Breitlinga y col., 2004).

Una vez calculada la expresión diferencial de todos los genes o el enriquecimiento de los términos GO g en todos los estudios i, se construye una **lista ordenada (rangos)**, de 1 a g, según los $logFC$ o los LOR: un 1 para el gen o GO con un valor más alto de tamaño del efecto, un 2 para el segundo, etc.

Aunque hay diversas métricas de combinación de rangos, la más utilizada es el producto. Para cada gen o función molecular g en $i = 1, 2, ..., k$ estudios se calcula el producto RP_g como la media geométrica de los rangos:

$$RP_g = \prod_{i=1}^{k} (r_{ig})^{\frac{1}{k}} \tag{3.20}$$

En algunos casos, sobre todo para sets de datos pequeños, esta métrica suele ser suficientemente interpretativa por sí sola. Los genes y las funciones moleculares con un RP muy pequeño son los mejores candidatos para ser los sobreexpresados o enriquecidos, y viceversa.

Es necesario para sets de datos más grandes, realizar un análisis de significación estadística. Para este cálculo, se utiliza una estrategia de **estimación basada en permutaciones**.

Para cada gen o término GO g se calcula el valor de su RP_g experimental. Además, se calcula un número l de valores de RP en experimentos aleatorios, con el mismo número k de estudios y g de genes o funciones. Cada uno de estos experimentos consiste en realizar permutaciones de los genes o GOs en el set de datos, y se calcula con la fórmula detallada arriba.

Una vez obtenidos los RP simulados, se cuenta cuántos RP están por debajo o por encima del RP experimental obtenido, y se estima de la probabilidad de observar un RP mayor que los RP obtenidos de con permutaciones, obteniéndose un p-valor experimental. Es posible después obtener un p-valor corregido según el procedimiento de Benjamini-Hochberg para controlar la tasa de falsos descubrimientos.

3.5. Evaluación de los métodos descritos

1. Los **metaanálisis a nivel de gen** se evaluarán para los métodos de combinación de p-valores, de tamaño del efecto (los FEM y los REM), y la combinación de rangos. Se compararán los resultados de los tres métodos con el resultado de aplicar, a todos los estudios de cada set de datos, una intersección de los genes significativos.

2. A nivel de **metaanálisis de funciones**, una vez obtenido el GSA para cada estudio, se compararán los métodos de combinación de p-valores, de tamaño del efecto (los FEM y los REM), y la combinación de rangos con también una intersección de las funciones enriquecidas en cada estudio individual.

3. Para el metaanálisis funcional, se evaluarán **dos estrategias diferentes**: *(a)* llevar a cabo un GSA individual de cada estudio y después un metaanálisis funcional, tal y como está descrito en el escenario 2; y *(b)* llevar a cabo un metaanálisis a nivel de gen y, sobre este resultado, aplicar un GSA.

Capítulo 4

Resultados

4.1. Set de datos de tumores

4.1.1. Preprocesado

El preprocesado de estos datos comprende un análisis exploratorio de los datos generados por RNA-Seq y un análisis de expresión génica diferencial entre enfermos y controles del set de datos de tumores del TCGA (The Cancer Genome Atlas).

4.1.1.1. Análisis exploratorio

Para cada uno de los 17 estudios, se realizó un PCA, un *clustering* jerárquico y un análisis de la distribución de los conteos (boxplot), para todas las muestras.

Lo que esperamos obtener con este análisis es una agrupación, tanto en el PCA como en el análisis *cluster*, de las muestras de individuos enfermos y, por otra parte, una agrupación de las muestras de individuos sanos. Además, los conteos deben estar normalizados. Se observarán estos gráficos para dos estudios, BLCA y CHOL, a modo de ejemplo.

En cuanto a la distribución de los conteos, podemos observar que las distribuciones entre las muestras son más o menos equivalentes y que apenas aparecen observaciones atípicas en ningún set de datos (Figura 4.1).

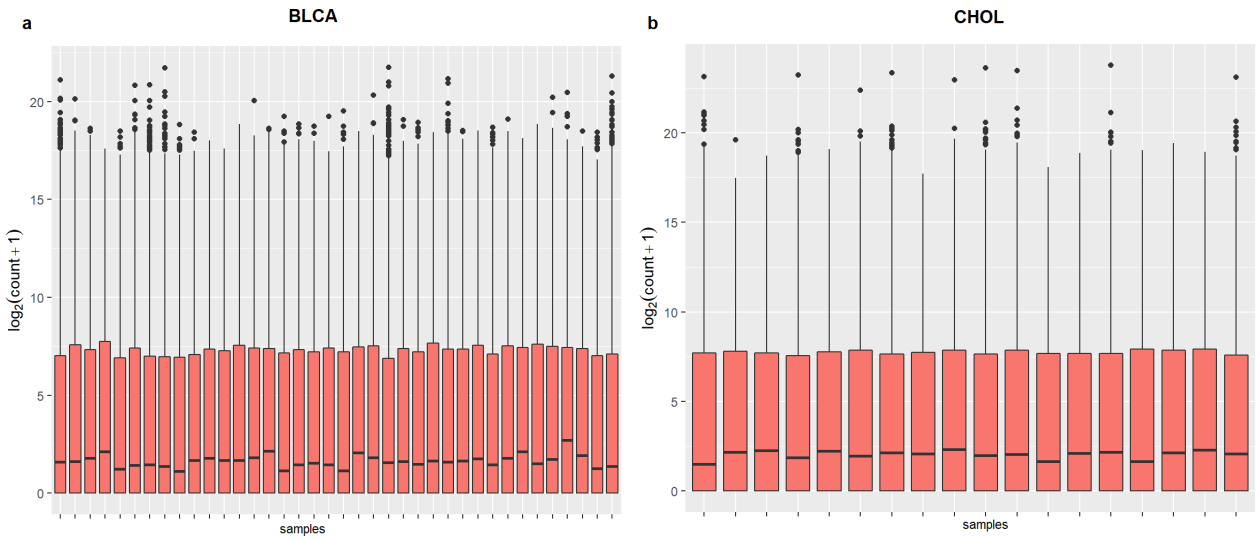

Figura 4.1 Distribución de los conteos normalizados para los estudios (a) BLCA y (b) CHOL del TCGA.

En cuanto al Análisis de Componentes Principales, los resultados son consistentes en casi todos los estudios: las muestras procedentes de los controles difieren mucho de las tumorales.

En el caso concreto de estos dos estudios, el BLCA muestra una división no tan clara de las muestras en las dos primeras dimensiones (Figura 4.2 a). El CHOL muestra una separación mucho más clara, por lo que podríamos decir que las diferencias en la expresión génica aquí son mucho mayores (Figura 4.2 b).

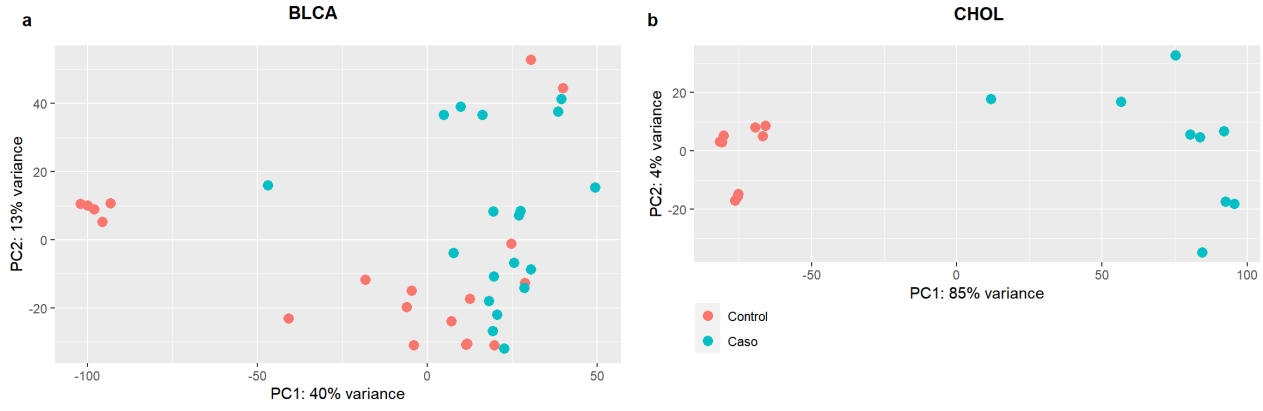

Figura 4.2 Análisis de Componentes Principales para los estudios (a) BLCA y (b) CHOL. Las muestras rojas indican muestras procedentes de un control, y las azules corresponden a las procedentes de un caso.

El análisis *cluster* aporta unos resultados similares a los observados en el PCA. Mientras que en el estudio CHOL los controles forman un grupo separado de las muestras cancerosas, en el BLCA esto no es así y la separación en grupos no es clara (Figura 4.3).

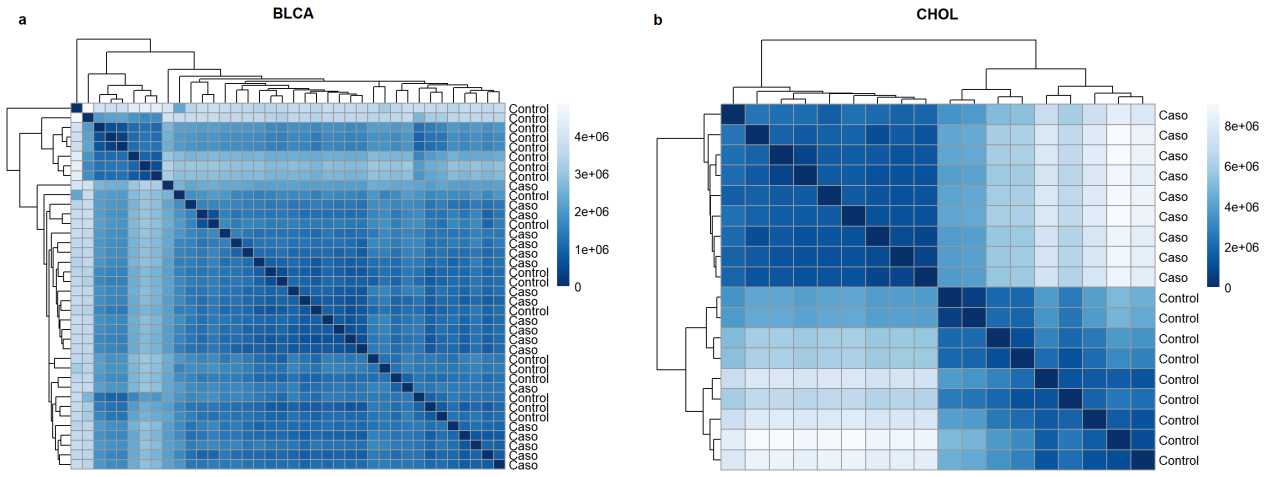

Figura 4.3 Análisis *cluster* para los estudios (a) BLCA y (b) CHOL.

4.1.1.2. Análisis de expresión diferencial

Con el análisis de expresión diferencial se pueden obtener los genes sobre e infraexpresados entre controles y muestras de tejidos cancerosos.

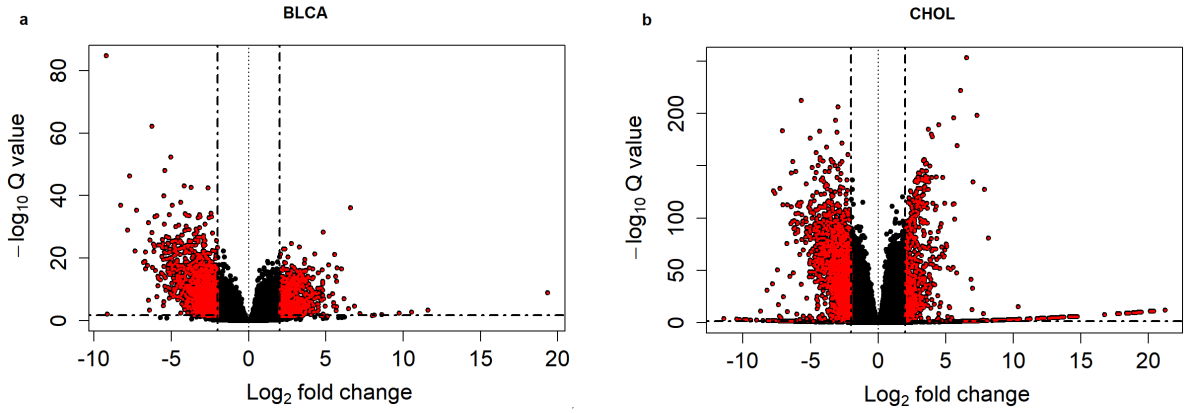

Figura 4.4 *Volcano plot* de los resultados de la expresión diferencial de los estudios (a) BLCA y (b) CHOL. En rojo se representan los genes diferencialmente expresados.

Estudio	Up	Down	Estudio	Up	Down
BLCA	2881	3871	LIHC	2099	3645
BRCA	2729	5723	LUAD	3621	4407
CHOL	4742	4915	LUSC	5169	6628
COAD	4243	4979	PRAD	1583	2693
ESCA	1376	1966	READ	2741	4276
HNSC	3045	3391	STAD	2840	3266
KICH	3911	7216	THCA	2381	3600
KIRC	4527	5301	UCEC	4124	5081
KIRP	2484	5440			

Tabla 4.1 Genes diferencialmente expresados en los estudios del TCGA. En la columna *up* aparecen representados los genes sobreexpresados, y en la columna *down* los infraexpresados, en las muestras de tumores respecto de los controles.

El *volcano plot* representa la cantidad de genes diferencialmente expresados entre condiciones, generalmente marcados en rojo, de acuerdo a su tamaño del efecto y a su significación estadística (Figura 4.4).

En la Tabla 4.1 se detalla la cantidad de genes que se han encontrado como diferencialmente expresados en los 17 estudios del TCGA.

4.1.2. Metaanálisis a nivel de gen

Como se ha descrito en el capítulo anterior, el metaanálisis a nivel de gen se realiza mediante dos métodos diferentes: (1) combinación de p-valores, ponderado y sin ponderar; y (2) combinación de tamaño del efecto, con modelos de efectos fijos y aleatorios.

Estos métodos se evaluarán junto con los resultados de la intersección de los análisis de expresión diferencial de cada estudio, es decir, se seleccionarán aquellos genes sobreexpresados e infraexpresados en todos los estudios.

La intersección da como resultado 50 genes diferencialmente expresados, 23 infraexpresados y 27 sobreexpresados.

Genes sobreexpresados			Genes infraexpresados		
Gen	LFC medio	P Combi	Gen	LFC medio	P Combi
ENSG00000170373	4.841	0.000	ENSG00000196616	-5.634	0.000
ENSG00000029559	4.454	0.000	ENSG00000034971	-4.866	0.000
ENSG00000139800	3.688	0.000	ENSG00000164530	-4.705	0.000
ENSG00000249550	3.495	0.000	ENSG00000181234	-4.364	0.000
ENSG00000099953	3.394	0.000	ENSG00000168079	-4.348	0.000
ENSG00000060718	3.341	0.000	ENSG00000121871	-4.168	0.000
ENSG00000122133	3.299	0.000	ENSG00000161649	-4.153	0.000
ENSG00000101057	3.287	0.000	ENSG00000184601	-4.134	0.000
ENSG00000175063	3.245	0.000	ENSG00000112936	-4.097	0.000
ENSG00000178776	3.227	0.000	ENSG00000184905	-4.096	0.000

Tabla 4.2 Top 10 de genes diferencialmente expresados del set de datos del TCGA ordenados por *log fold-change* según el metaanálisis con métodos de combinación de *p*-valores. LFC medio: media del *log fold-change* entre todos los estudios. · P Combi: *p*-valor combinado y ajustado para comparaciones múltiples.

4.1.2.1. Métodos de combinación de *p*-valores

En este set de datos, los métodos de combinación de *p*-valores dan como resultado 2302 genes diferencialmente expresados, 1596 infraexpresados y 706 sobreexpresados. Ambos métodos, el de la normal inversa y el de Fisher, dan exactamente los mismos resultados, que se resumen en la Tabla 4.2.

Todos los genes detectados como diferencialmente expresados por la intersección entre estudios también se detectaron como tal con la combinación de *p*-valores, aunque este último método permitió detectar más de 2000 genes adicionales que no podrían haber sido detectados sin técnicas de metaanálisis.

4.1.2.2. Métodos de combinación del tamaño del efecto

Dentro de estos métodos se evaluarán dos modelos principales: uno de efectos fijos (FEM) y otro de efectos aleatorios (REM), ambos detallados en la Metodología. La evaluación

también será conjunta a la intersección de genes que se ha obtenido anteriormente.

Con el FEM se obtienen 4932 genes diferencialmente expresados, 3509 infraexpresados y 1423 sobreexpresados. Por su parte, el REM detecta un total de 4091 genes con expresión diferencial entre casos y controles, de los cuales 2762 están infraexpresados y 1329 sobreexpresados. Los resultados principales, el top 10 de genes sobreexpresados y el 10 de infraexpresados, de cada modelo aparecen resumidos en la Tabla 4.3.

La manera de representar los resultados de este tipo de metaanálisis suele ser en un **diagrama de bosque**, en el que se representa el LFC de cada estudio y su IC al 95 %. En la Figura 4.5 aparece el diagrama de bosque para el gen ENSG00000251026, presentado en la Tabla 4.3 para el metaanálisis con un REM.

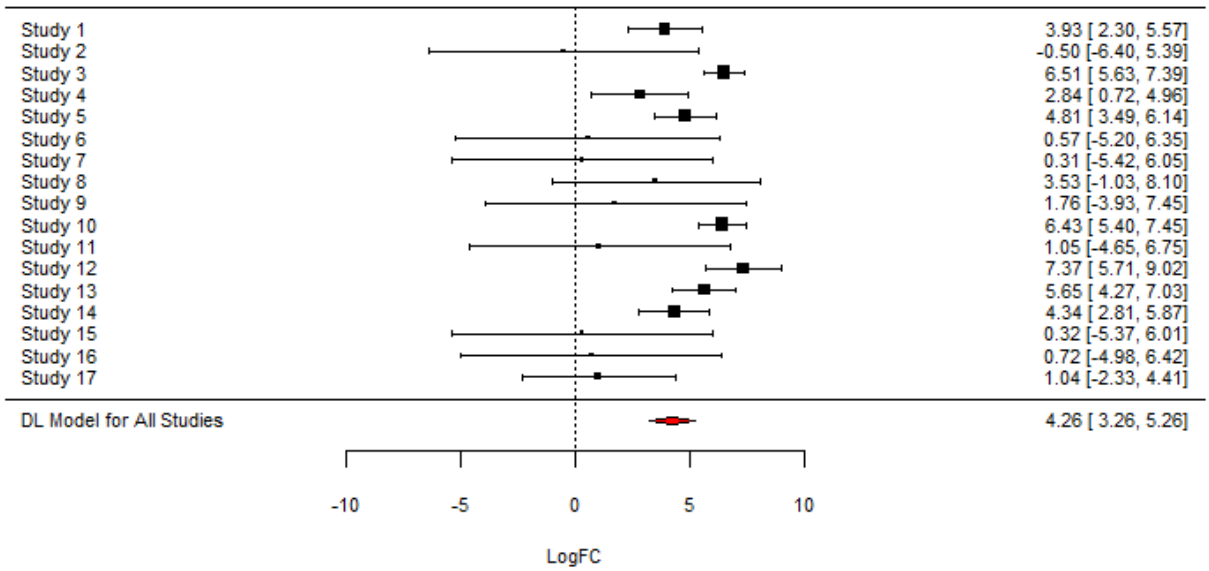

Figura 4.5 Diagrama de bosque para el gen ENSG00000251026 en el set de datos del TCGA con un Modelo de Efectos Aleatorios (REM).

Análisis de heterogeneidad, sensibilidad y estudios influyentes

El análisis de heterogeneidad se realizará a través de las métricas descritas en el capítulo previo y a través de diversos diagramas. Emplearemos el metaanálisis con un REM como base para presentar los resultados del análisis de heterogeneidad.

Gen	LFC	IC 95 %	τ^2	Q	QP	P	P Ajust
REM. Top 5 Genes Infraexpresados							
ENSG00000196616	-5.60	[-6.61, -4.60]	4.264	584.92	4×10^{-114}	6×10^{-25}	7×10^{-24}
ENSG00000034971	-4.95	[-6.56, -3.34]	10.86	1189.1	3×10^{-243}	2×10^{-9}	3×10^{-8}
ENSG00000164530	-4.73	[-5.39, -4.06]	1.795	240.97	4×10^{-42}	2×10^{-43}	3×10^{-39}
ENSG00000168079	-4.35	[-5.11, -3.59]	2.323	316.36	1×10^{-57}	2×10^{-29}	2×10^{-26}
ENSG00000181234	-4.34	[-5.03, -3.65]	1.885	243.76	1×10^{-42}	6×10^{-35}	2×10^{-31}
REM. Top 5 Genes Sobreexpresados							
ENSG00000170373	4.81	[3.34, 6.29]	9.107	564.02	9×10^{-110}	1×10^{-10}	3×10^{-9}
ENSG00000029559	4.49	[3.76, 5.22]	1.945	140.74	5×10^{-22}	2×10^{-33}	5×10^{-30}
ENSG00000251026	4.26	[3.26, 5.26]	2.212	52.95	8×10^{-6}	5×10^{-17}	5×10^{-15}
ENSG00000139800	3.92	[2.51, 5.32]	7.599	330.66	1×10^{-60}	4×10^{-8}	6×10^{-7}
ENSG00000272328	3.73	[2.84, 4.62]	1.615	48.94	3×10^{-5}	2×10^{-16}	2×10^{-14}
FEM. Top 5 Genes Infraexpresados							
ENSG00000198398	-9.04	[-9.52, -8.57]		151.96	3×10^{-24}	2×10^{-304}	4×10^{-302}
ENSG00000279460	-7.03	[-7.52, -6.53]		143.65	1×10^{-22}	7×10^{-170}	3×10^{-168}
ENSG00000263761	-6.78	[-7.41, -6.15]		192.74	2×10^{-32}	1×10^{-99}	2×10^{-98}
ENSG00000161992	-6.23	[-6.60, -5.87]		616.87	6×10^{-121}	2×10^{-246}	2×10^{-244}
ENSG00000215231	-6.07	[-6.55, -5.59]		65.66	6×10^{-8}	6×10^{-134}	2×10^{-132}
FEM. Top 5 Genes Sobreexpresados							
ENSG00000168619	7.69	[7.00, 8.38]		135.69	5×10^{-21}	7×10^{-107}	1×10^{-105}
ENSG00000230432	5.43	[4.92, 5.95]		99.94	4×10^{-14}	1×10^{-95}	2×10^{-94}
ENSG00000251026	5.38	[4.93, 5.82]		52.95	8×10^{-6}	1×10^{-124}	3×10^{-123}
ENSG00000170373	5.15	[4.91, 5.40]		564.02	1×10^{-109}	0	0
ENSG00000250874	5.05	[4.46, 5.64]		55.50	3×10^{-6}	3×10^{-63}	3×10^{-62}

Tabla 4.3 Top 10 de genes diferencialmente expresados del set de datos del TCGA según el Modelo de Efectos Aleatorios (REM) y de Efectos Fijos (FEM) ordenados por *log fold-change*. LFC e IC 95 %: estimación del *log fold-change* combinado e intervalo de confianza al 95 %. · τ^2: medida de heterogeneidad. · Q y QP: estadístico de contraste y p-valor del test de Cochran. · P y P Ajust: p-valor de la expresión diferencial y p-valor corregido para comparaciones múltiples.

Para los 5 genes sobreexpresados y los 5 infraexpresados del REM presentados en la Tabla 4.3, las métricas para estudiar la heterogeneidad se presentan en la Tabla 4.4. Por ejemplo, para el gen ENSG00000196616, el valor del I^2 indica que la relación entre la variabilidad que presentan los estudios y la variabilidad total es del 97.26 %. El valor del H^2 es de 36.56, lo que representa el cociente entre la variabilidad total y la variabilidad en el muestreo. Esto es indicativo de que los estudios son muy heterogéneos y que el REM es el modelo adecuado.

Las gráficas de embudo para el estudio de la heterogeneidad, en este caso obtenidas para el gen ENSG00000251026, aparecen representadas en la Figura 4.6. Estas gráficas indican que, para este gen, hay tres estudios fuera de la región de confianza. Si esto se repitiese de manera generalizada para todos los genes, habría que plantearse la exclusión de estos estudios del análisis. En nuestro set de datos, esto no sucede.

Los resultados del análisis de sensibilidad y de estudios influyentes aparecen detallados para el top 10 de genes en la Tabla 4.5. Además, las métricas de estudios influyentes en forma gráfica para el gen ENSG00000251026 aparecen representadas en la Figura 4.7.

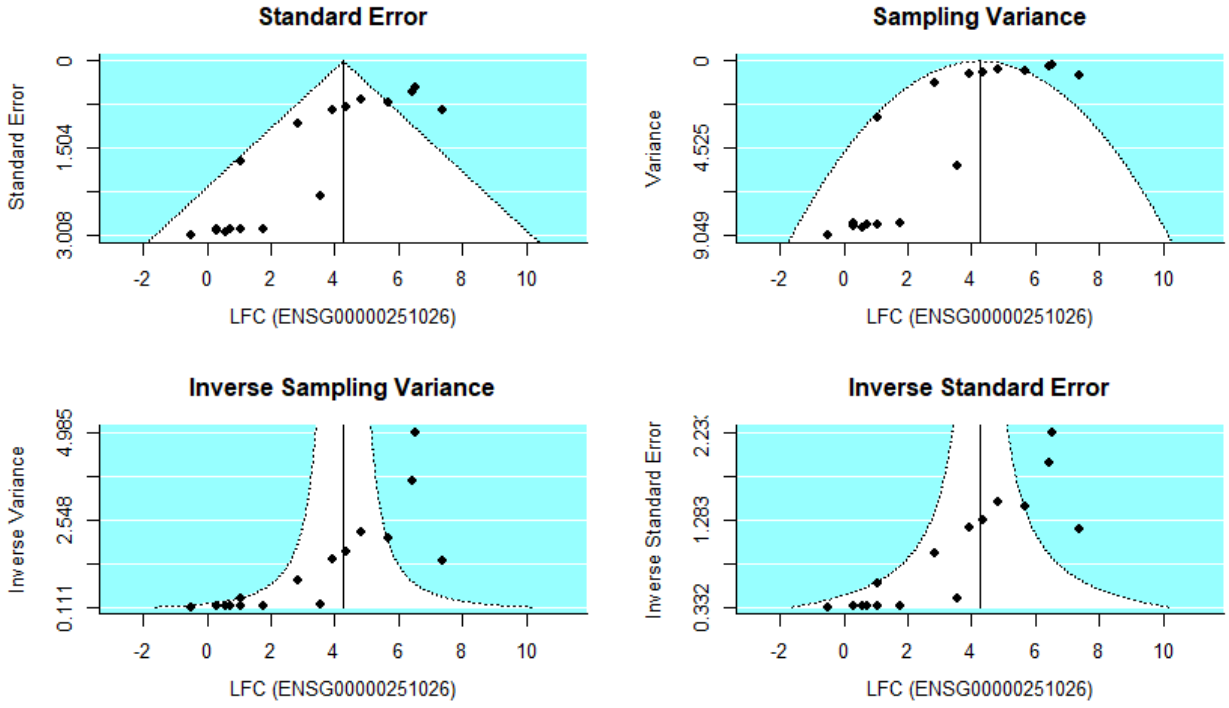

Figura 4.6 Diagramas de embudo para el gen ENSG00000251026 en el set de datos del TCGA con un Modelo de Efectos Aleatorios (REM). En todos ellos tres estudios quedan fuera de la región de confianza y por tanto se consideran heterogéneos.

Gen	τ^2	I^2 (%)	H^2
ENSG00000196616	4.264	97.26	36.56
ENSG00000034971	10.86	98.65	74.32
ENSG00000164530	1.795	93.36	15.06
ENSG00000168079	2.323	94.94	19.77
ENSG00000181234	1.885	93.44	15.24
ENSG00000170373	9.107	97.16	35.25
ENSG00000029559	1.945	88.63	8.80
ENSG00000251026	2.212	69.78	3.31
ENSG00000139800	7.599	95.16	20.67
ENSG00000272328	1.615	67.31	3.06

Tabla 4.4 Métricas de heterogeneidad para el Modelo de Efectos Aleatorios (REM) del metaanálisis a nivel de gen con el set de datos del TCGA.

	E. Influyentes		Sensibilidad	
Gen	**Signo LFC**	**Núm. influ.**	**Tam. Efecto**	**Significación**
ENSG00000196616	17	0	0.9999	17
ENSG00000034971	16	0	0.9788	17
ENSG00000164530	17	1 (BLCA)	0.9932	17
ENSG00000168079	17	0	0.9994	17
ENSG00000181234	17	0	0.9960	17
ENSG00000170373	16	0	0.9934	17
ENSG00000029559	17	0	0.9875	17
ENSG00000251026	16	0	0.8551	17
ENSG00000139800	15	0	0.9587	17
ENSG00000272328	17	0	0.8585	17

Tabla 4.5 Análisis de estudios influyentes y sensibilidad para el Modelo de Efectos Aleatorios (REM) del metaanálisis a nivel de gen con el set de datos del TCGA. Signo LFC: número de estudios con el mismo signo del LFC que el estimado. · Núm. influ: número de estudios influyentes. · Tam. Efecto: p-valor del test t sobre los LFC obtenidos por *leave-one-out* comparado con el LFC estimado con el set de datos completo. · Significación: Número de estudios con un p-valor $< 0{,}05$ para ese gen.

El **análisis de estudios influyentes** dio como resultado que para 2481 genes del metaanálisis (un 55.35 % del total de los diferencialmente expresados) no se detectaron estudios influyentes. Para el resto de genes, se detectó al menos algún estudio influyente. Los estudios que más se detectaron como influyentes fueron KICH y KIRC, detectados como influyentes únicamente para 317 y 318 genes, respectivamente (un 7.07 % y un 7.10 %). Por tanto, y teniendo esto en cuenta, no se debería eliminar ninguno del metaanálisis.

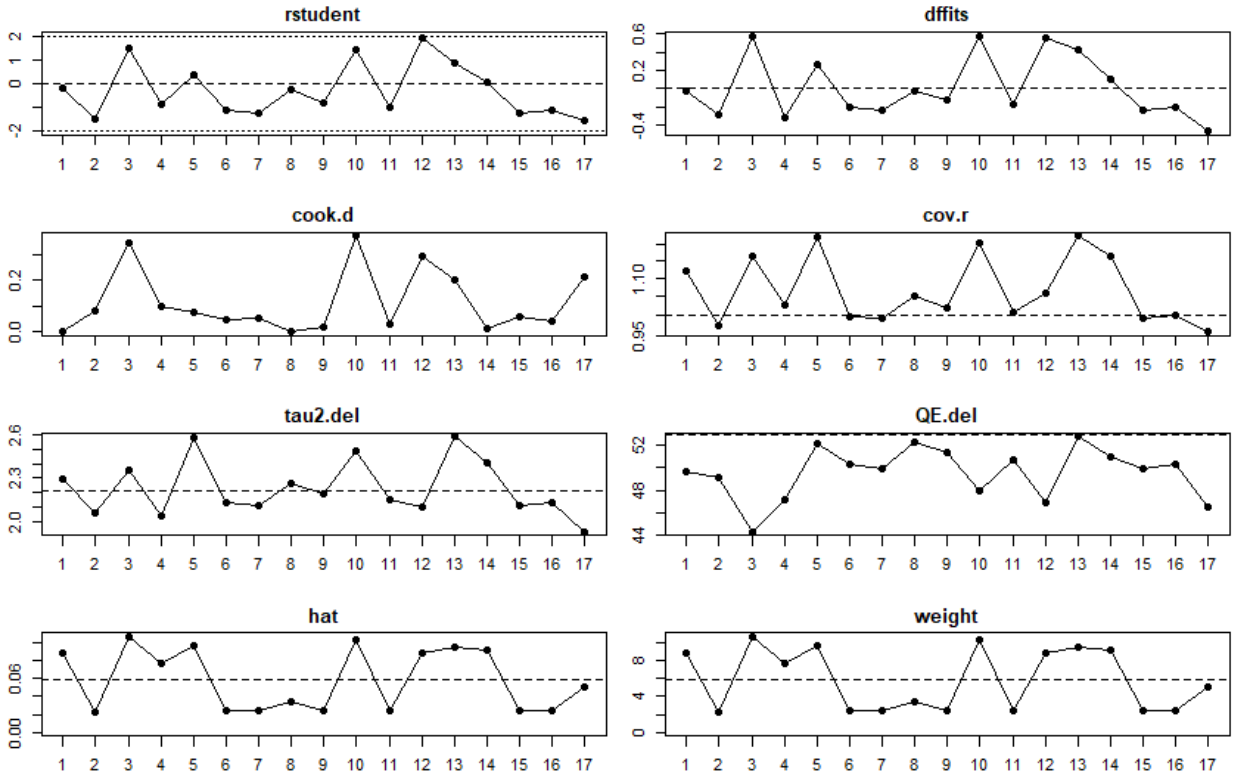

Figura 4.7 Análisis de estudios influyentes en el modelo REM del metaanálisis a nivel de gen del set de datos de TCGA para el gen ENSG00000251026. rstudent y dffits: residuos estandarizados. · Cook.d: distancia de Cook. · cov.r: ratio de las covarianzas. · tau2.del: medida de τ^2 al excluir un estudio. · QE.del: medida del estadístico Q al excluir un estudio. · weight y hat: número de desviaciones estándar que cambian los coeficientes al excluir un estudio.

En cuanto al **análisis de sensibilidad**, el LFC estimado en el metaanálisis y todos los estimados para cada gen excluyendo un estudio cada vez (por *leave-one-out*) son similares para el 99.91 % de los estudios: sólo 4 genes mostraron un p-valor $< 0{,}05$. Además, para el 92.90 % de los genes diferencialmente expresados, los p-valores de la expresión diferencial de los estudios individuales fueron todos significativos.

Gen	Rank	LFC medio	P Ajust
ENSG00000196616	1-Down	-5.634	7×10^{-27}
ENSG00000034971	2-Down	-4.866	3×10^{-24}
ENSG00000164530	3-Down	-4.705	1×10^{-22}
ENSG00000168079	4-Down	-4.348	2×10^{-20}
ENSG00000181234	5-Down	-4.364	2×10^{-20}
ENSG00000029559	1-Up	4.454	9×10^{-25}
ENSG00000170373	2-Up	4.841	9×10^{-25}
ENSG00000197587	3-Up	3.465	4×10^{-18}
ENSG00000101057	4-Up	3.287	6×10^{-18}
ENSG00000139800	5-Up	3.688	9×10^{-18}

Tabla 4.6 Top de genes diferencialmente expresados del set de datos del TCGA según el método de combinación de rangos. Rank: lugar que ocupa el gen en el ranking, *up* en el de genes sobreexpresados, y *down* en el de infraexpresados, al realizarse el test bidireccionalmente. · LFC medio: logaritmo del *fold-change* medio. · P Ajust: p-valor ajustado por comparaciones múltiples por FDR (Benjamini-Hochberg).

4.1.2.3. Combinación de rangos

La combinación de rangos se realiza bidireccionalmente para el mismo set de datos: una clasificando al principio de la lista los genes con mayor LFC, y otra al contrario, es decir, clasificando al principio de la lista los de menor LFC (los infraexpresados).

Este método ha dado como resultado un total de 4164 genes diferencialmente expresados: 2941 han sido detectados como infraexpresados (LFC medio < 1 y FDR $< 0{,}05$), y 1223 se han detectado como sobreexpresados (LFC medio > 1 y FDR $< 0{,}05$). Los principales resultados para este tipo de metaanálisis aparecen en la Tabla 4.6.

4.1.2.4. Comparativa de todos los métodos

La comparativa de los diferentes métodos de metaanálisis a nivel de gen descritos anteriormente se realiza utilizando gráficos de Venn, donde se pueden observar las intersecciones de los genes significativos entre todos los métodos. Las comparativas que se realizarán serán:

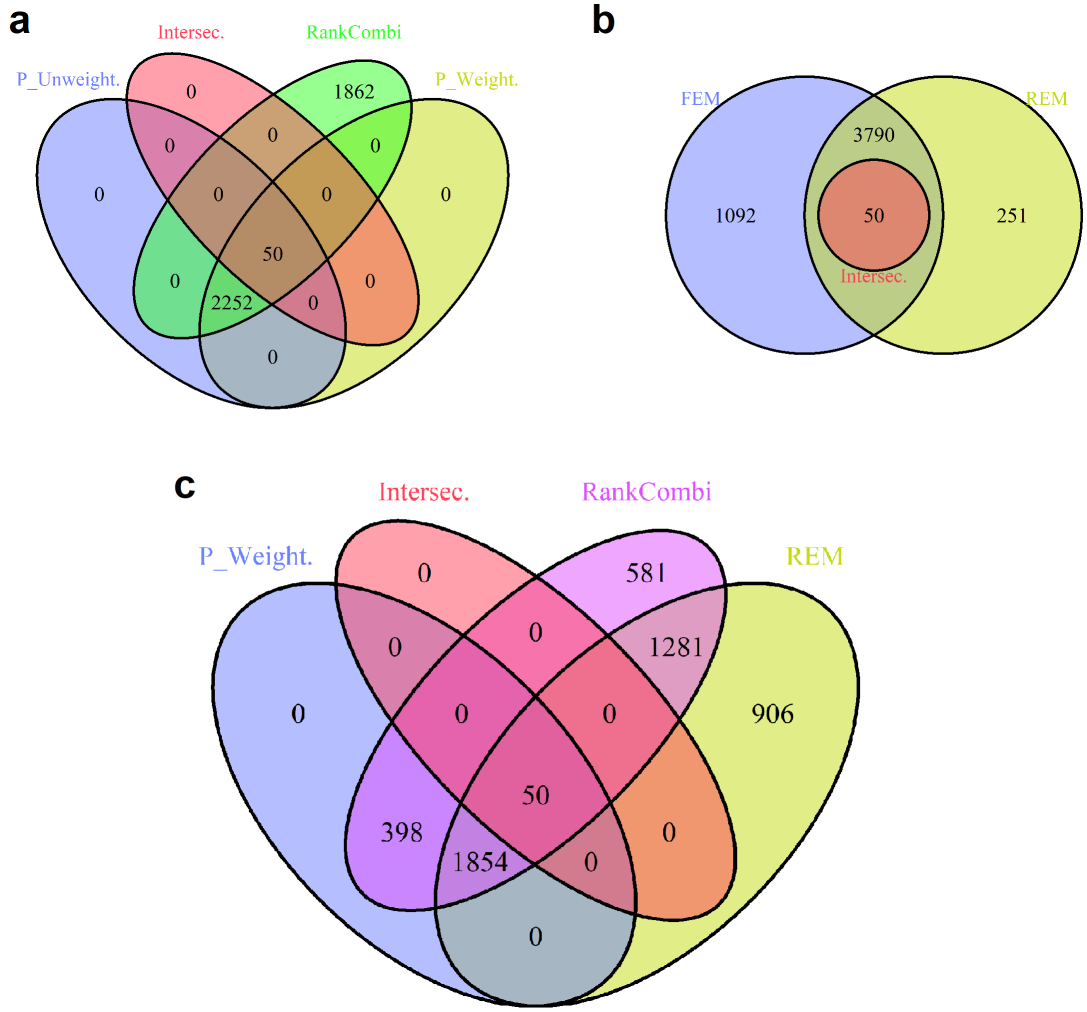

Figura 4.8 Comparativa de resultados del metaanálisis a nivel de gen del set de datos de TCGA. PUnweight.: combinación de p-valores sin ponderar. · PWeight.: combinación de p-valores ponderada. · FEM: Modelo de Efectos Fijos. · REM: Modelo de Efectos Aleatorios. · RankCombi: Combinación de rangos. · Intersecc.: intersección de genes diferencialmente expresados en todos los estudios.

- Los dos métodos de combinación de p-valores junto al método de combinación de rangos y la intersección de genes significativos en todos los estudios, como métodos más sencillos de metaanálisis (Figura 4.8 a).
- Los dos métodos de combinación de tamaño del efecto (modelos de efectos fijos y de efectos aleatorios) y la intersección de genes significativos en todos los estudios, como modelos complejos de metaanálisis (Figura 4.8 b).

Estudio	Top	Bottom	Estudio	Top	Bottom
BLCA	44	18	LIHC	29	22
BRCA	61	2	LUAD	48	13
CHOL	17	9	LUSC	50	25
COAD	17	45	PRAD	9	50
ESCA	66	3	READ	20	10
HNSC	20	13	STAD	43	4
KICH	0	130	THCA	33	1
KIRC	23	6	UCEC	31	10
KIRP	16	29			

Tabla 4.7 Términos GO enriquecidos en los 17 estudios del TCGA. En la columna *top* aparecen representados los términos GO sobrerrepresentados en los tejidos cancerosos, y en la columna *bottom* los infrarrepresentados.

- Un método de cada grupo, para una comparación completa: un modelo de efectos aleatorios, una combinación ponderada de p-valores, la intersección y la combinación de rangos (Figura 4.8 c).

El diagrama se ha construido haciendo uso de la librería de R *VennDiagram* (Chen y col., 2011).

4.1.3. Enriquecimiento funcional

A partir de los resultados de los análisis de expresión diferencial de cada uno de los 17 estudios del TCGA, se realiza un paso intermedio de **enriquecimiento funcional** de términos GO, previo al metaanálisis de funciones. Este tipo de estudios revelan los términos GO enriquecidos. El número de términos GO enriquecidos, sobrerrepresentados (*top*) e infrarrepresentados (*bottom*), aparecen detallados en la Tabla 4.7.

Como muestra, se representan los resultados del enriquecimiento funcional del estudio BLCA (Tabla 4.8): los tres principales términos GO infrarrepresentados y sobrerrepresentados, ordenados por el logaritmo de su *odds-ratio* (LOR) y sus métricas más importantes.

GO	Descripción	LOR	SE	P Ajust
GO:0035584	calcium-mediated signaling using intracellular calcium source	-2.073	0.218	2×10^{-18}
GO:0032355	response to estradiol	-1.777	0.316	8×10^{-6}
GO:0001837	epithelial to mesenchymal transition	-1.641	0.288	5×10^{-6}
GO:0007052	mitotic spindle organization	2.747	0.189	5×10^{-44}
GO:0031145	anaphase-promoting complex-dependent catabolic process	2.179	0.184	4×10^{-29}
GO:1901990	regulation of mitotic cell cycle phase transition	2.179	0.184	4×10^{-29}

Tabla 4.8 Top de términos GO enriquecidos en el estudio BLCA del set de datos del TCGA. · GO y descripción: término GO y función molecular que representa. · LOR: logaritmo del *odds-ratio* para el enriquecimiento entre muestras de pacientes enfermos y sanos. · SE: error estándar del LOR. · P Ajust: p-valor del enriquecimiento ajustado por comparaciones múltiples por FDR según la metodología de Benjamini-Hochberg.

4.1.4. Metaanálisis a nivel de función

El metaanálisis a nivel de función se realizará con las mismas técnicas que las detalladas para el metaanálisis a nivel de gen (combinación de p-valores, de tamaño del efecto y de rangos), pero utilizando como ítem las funciones moleculares representadas por los términos GO. Además, se incorporará un método de metaanálisis adicional: se realizará un análisis de enriquecimiento funcional sobre los resultados del metaanálisis a nivel de gen.

La intersección de términos GO significativos en todos los estudios a nivel funcional dio como resultado únicamente 8 términos GO detectados como enriquecidos, estando todos sobrerrepresentados en las muestras de pacientes enfermos respecto a las muestras de individuos sanos.

Por su parte, realizar un análisis de enriquecimiento funcional sobre los resultados del metaanálisis de gen con un Modelo de Efectos Aleatorios (el más apropiado por la heterogeneidad que presentan los estudios) dio como resultado 129 términos GO enriquecidos: 62 infrarrepresentados y 67 sobrerrepresentados.

GO	Descripción	LOR	P Fish.	P N.Inv.
GO:0042178	xenobiotic catabolic process	-1.112	0.000	0.000
GO:1901687	glutathione derivative biosynthetic process	-1.060	0.000	0.000
GO:0070527	platelet aggregation	-0.935	2×10^{-5}	9×10^{-3}
GO:0031145	anaphase-promoting complex-dependent catabolic process	1.410	0.000	0.000
GO:1901990	regulation of mitotic cell cycle phase transition	1.410	0.000	0.000
GO:0007059	chromosome segregation	1.343	2×10^{-12}	5×10^{-11}

Tabla 4.9 Top de términos GO enriquecidos en el set de datos del TCGA según el metaanálisis con métodos de combinación de *p*-valores ordenados por su logaritmo del *odds-ratio*. · LOR: logaritmo del *odds-ratio* para el enriquecimiento entre muestras de pacientes enfermos y sanos. · P Fish. y P N.Inv.: *p*-valor combinado según la metodología de Fisher (no ponderado) y la de la normal inversa (ponderado), respectivamente, ajustado para comparaciones múltiples.

4.1.4.1. Métodos de combinación de *p*-valores

Las dos metodologías de metaanálisis por combinación de *p*-valores (Fisher o sin ponderar, y normal inversa o ponderado) arrojan unos resultados muy similares.

El método de combinación de *p*-valores ponderado detecta como enriquecidos 82 términos GO: 44 sobrerrepresentados y 38 infrarrepresentados. El método no ponderado detecta 83, 46 sobrerrepresentados y 37 infrarrepresentados. En la Tabla 4.9 se representa un resumen de los resultados obtenidos por ambos métodos.

4.1.4.2. Métodos de combinación del tamaño del efecto

Como anteriormente, se evaluarán los dos tipos de modelos descritos: un modelo de efectos fijos (FEM) y otro de efectos aleatorios (REM).

El modelo REM da como resultado la detección de 292 términos GO enriquecidos, 148 sobrerrepresentados en pacientes cancerosos y 144 infrarrepresentados. El FEM, por su parte, detecta 426 términos GO enriquecidos, 208 y 218 sobre e infrarrepresentados, respectivamente. Los principales resultados de este método aparecen resumidos en la Tabla 4.10.

GO	Descripción	LOR	Q	QP	P Ajust
REM. Top 5 Términos GO infrarrepresentados					
GO:0055119	relaxation of cardiac muscle	-1.15	47.28	6×10^{-5}	3×10^{-11}
GO:0042178	xenobiotic catabolic process	-1.14	155.8	6×10^{-25}	4×10^{-6}
GO:0043408	regulation of MAPK cascade	-1.11	32.48	9×10^{-3}	4×10^{-13}
GO:1901687	glutathione derivative biosynthetic process	-1.02	18.46	3×10^{-1}	1×10^{-23}
GO:0070527	platelet aggregation	-0.99	36.64	2×10^{-3}	1×10^{-10}
REM. Top 5 Términos GO sobrerrepresentados					
GO:0007052	mitotic spindle organization	2.14	65.95	5×10^{-8}	2×10^{-65}
GO:0031145	anaphase-promoting complex-dependent [...]	1.52	65.10	7×10^{-8}	1×10^{-26}
GO:1901990	regulation of mitotic cell cycle phase [...]	1.52	65.10	7×10^{-8}	1×10^{-26}
GO:0038061	NIK/NF-kappaB signaling	1.22	86.94	9×10^{-12}	2×10^{-12}
GO:0000077	DNA damage checkpoint	1.21	33.27	7×10^{-3}	7×10^{-15}
FEM. Top 5 Términos GO infrarrepresentados					
GO:0042178	xenobiotic catabolic process	-1.70	155.8	6×10^{-25}	3×10^{-112}
GO:0055119	relaxation of cardiac muscle	-1.34	47.27	6×10^{-5}	5×10^{-45}
GO:0043408	regulation of MAPK cascade	-1.25	32.48	9×10^{-3}	2×10^{-33}
GO:0035584	calcium-mediated signaling using [...]	-1.21	70.79	7×10^{-9}	9×10^{-40}
GO:0032355	response to estradiol	-1.11	35.84	3×10^{-3}	5×10^{-28}
FEM. Top 5 Términos GO sobrerrepresentados					
GO:0007052	mitotic spindle organization	2.40	65.95	5×10^{-8}	0.000
GO:0031145	anaphase-promoting complex-dependent [...]	1.73	65.10	7×10^{-8}	2×10^{-158}
GO:1901990	regulation of mitotic cell cycle phase [...]	1.73	65.10	7×10^{-8}	2×10^{-158}
GO:0038061	NIK/NF-kappaB signaling	1.47	86.94	9×10^{-12}	1×10^{-103}
GO:0010972	negative regulation of G2/M transition [...]	1.43	86.42	1×10^{-11}	9×10^{-91}

Tabla 4.10 Top 10 de términos GO enriquecidos del set de datos del TCGA según el Modelo de Efectos Aleatorios (REM) y de Efectos Fijos (FEM) ordenados por *log odds-ratio*. LOR: estimación del *log odds-ratio* de todos los estudios. · Q y QP: estadístico de contraste y p-valor del test de Cochran. · P Ajust: p-valor del enriquecimiento corregido para comparaciones múltiples.

Como antes, es posible obtener **diagramas de bosque** en el que se representa el LOR de cada análisis de enriquecimiento individual y el LOR final estimado para un metaanálisis. En la Figura 4.9 se representa este diagrama en un REM para el término GO:0043408, *regulación de cascadas de MAPK*, reacciones bioquímicas comunes en procesos cancerosos (Fang y col., 2005).

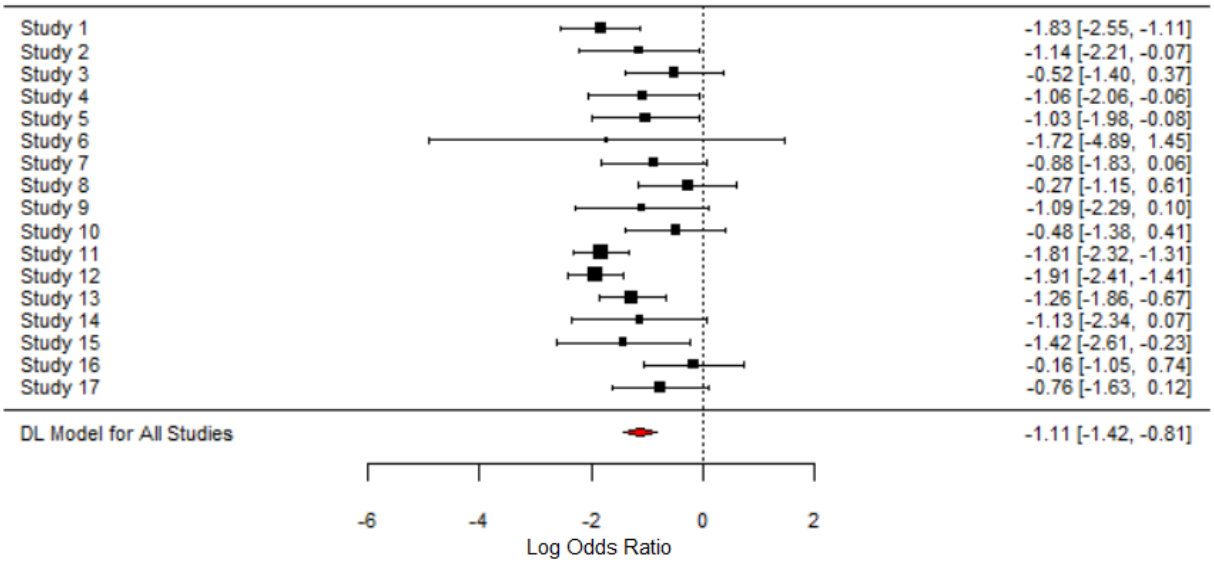

Figura 4.9 Diagrama de bosque para el término GO:0043408, *"regulación de cascadas de MAPK"*, en el set de datos del TCGA con un Modelo de Efectos Aleatorios (REM).

Análisis de heterogeneidad, sensibilidad y estudios influyentes

Esta validación del modelo se realizará sobre el modelo de efectos aleatorios (REM), al ser el más adecuado por ser los estudios bastante heterogéneos.

Las métricas de **heterogeneidad** de los 10 términos GO presentados como enriquecidos en la Tabla 4.10 aparecen detalladas en la Tabla 4.11. Además, se presenta en la Figura 4.10, a modo de ejemplo, las gráficas de embudo para el término anterior, GO:0043408. En estas gráficas podemos ver que solo dos estudios parecen salir de la región de confianza y presentar un extra de heterogeneidad. Sin embargo, no parece que esto se repita sistemáticamente para todos los términos GO, por lo que no hay problemas por la presencia de estudios demasiado heterogéneos.

Térm. GO	τ^2	$\mathbf{I^2}$ **(%)**	$\mathbf{H^2}$
GO:0055119	0.301	66.16	2.955
GO:0042178	0.903	89.73	9.740
GO:0043408	0.186	50.74	2.030
GO:1901687	0.023	13.33	1.154
GO:0070527	0.214	56.32	2.290
GO:0007052	0.160	75.74	4.122
GO:0031145	0.219	75.42	4.069
GO:1901990	0.219	75.42	4.069
GO:0038061	0.346	81.59	5.434
GO:0000077	0.177	51.91	2.079

Tabla 4.11 Métricas de heterogeneidad para el Modelo de Efectos Aleatorios del metaanálisis a nivel funcional con el set de datos del TCGA.

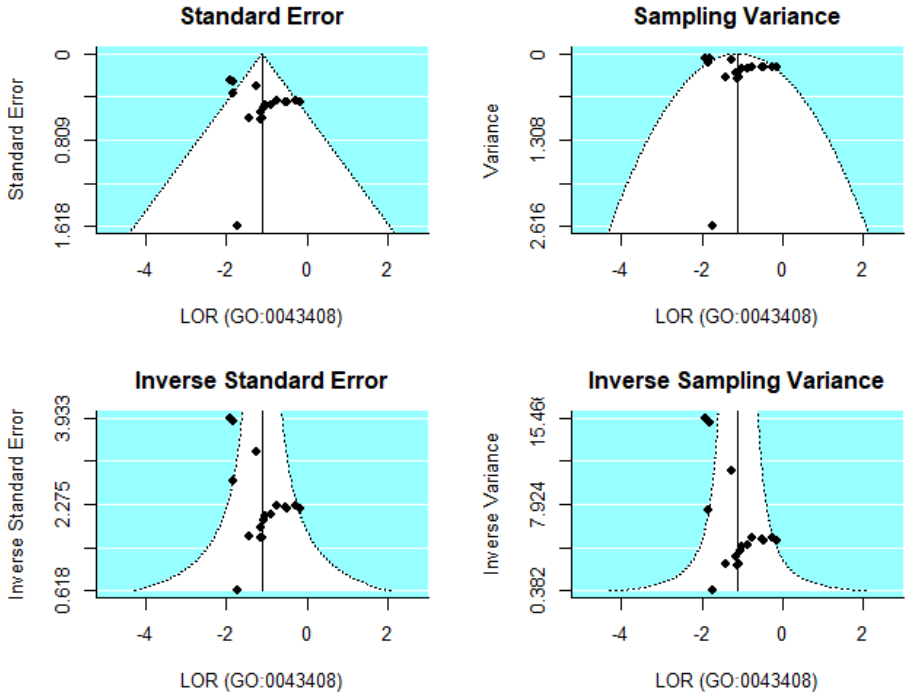

Figura 4.10 Diagramas de embudo para el término GO:0043408 en el set de datos del TCGA con un Modelo de Efectos Aleatorios (REM). Dos estudios quedan fuera de la región de confianza y se toman como heterogéneos.

	E. Influyentes		Sensibilidad	
Térm. GO	**Signo LFC**	**Núm. influ.**	**Tam. Efecto**	**Significación**
GO:0055119	15	1 (COAD)	0.9617	17
GO:0042178	17	1 (STAD)	0.7210	17
GO:0043408	17	0	0.9969	17
GO:1901687	17	0	0.9632	17
GO:0070527	16	0	0.9843	17
GO:0007052	17	1 (KIRP)	0.9088	17
GO:0031145	16	1 (KIRP)	0.9670	17
GO:1901990	16	1 (KIRP)	0.9670	17
GO:0038061	15	1 (KIRP)	0.9425	17
GO:0000077	16	1 (KIRP)	0.9768	17

Tabla 4.12 Análisis de estudios influyentes y sensibilidad para el Modelo de Efectos Aleatorios del metaanálisis a nivel funcional con el set de datos del TCGA. Signo LOR: número de estudios con el mismo signo del LOR que el estimado. · Núm. influ: número de estudios influyentes. · Tam. Efecto: p-valor del test t sobre los LOR obtenidos por *leave-one-out* comparado con el LOR estimado con todo el set de datos. · Significación: Número de estudios con un p-valor $< 0{,}05$ para ese GO.

El **análisis de estudios influyentes** ha detectado que para 209 términos GO (un 71.6 %) no hay estudios influyentes. Para 48 de estos términos, un 16.8 %, el estudio KIRP aparece como influyente. Podría ser necesario excluirlo del metaanálisis al estar detectado como *outlier* para muchas funciones moleculares. El resto de los estudios se detectan como influyentes para menos de un 2 % de los términos GO.

El **análisis de sensibilidad** dio como resultado que los LOR estimados en el metaanálisis excluyendo un estudio cada vez y el LOR final eran similares en todos los casos, al mostrar todos un p-valor para el test $t > 0{,}05$. Además, para el 99 % de los términos GO enriquecidos, los p-valores del enriquecimiento fueron significativos para todos los estudios.

El resultado detallado de ambos análisis aparece detallado para los primeros términos GO en la Tabla 4.12. Además, el análisis de estudios influyentes en forma gráfica para el GO:0043408 aparece representado en la Figura 4.11.

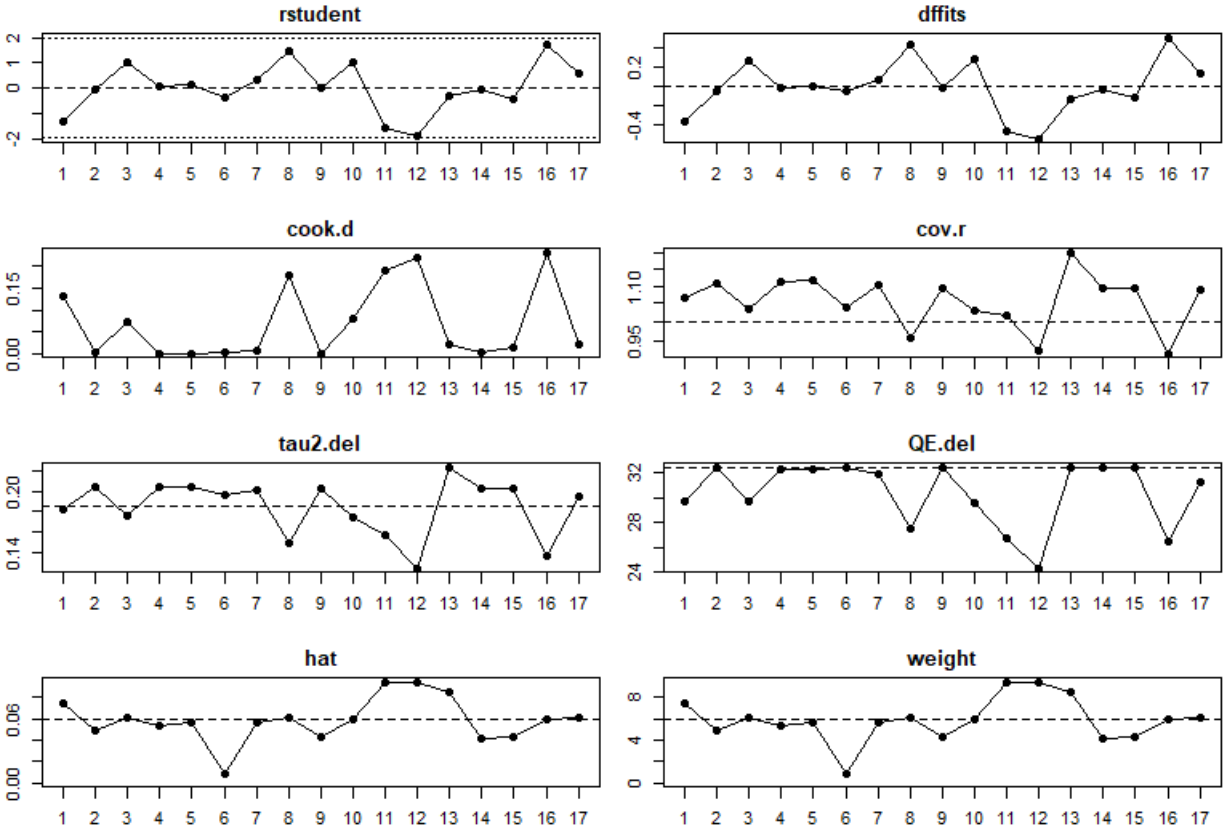

Figura 4.11 Análisis de estudios influyentes en el modelo REM del metaanálisis a nivel funcional del set de datos de TCGA para el término GO:0043408. rstudent y dffits: residuos estandarizados. · Cook.d: distancia de Cook. · cov.r: ratio de las covarianzas. · tau2.del: medida de τ^2 al excluir un estudio. · QE.del: medida del estadístico Q al excluir un estudio. · weight y hat: número de desviaciones estándar que cambian los coeficientes al excluir un estudio.

Según la Tabla 4.12, parece que para los principales términos GO detectados como sobrerrepresentados en tejidos cancerosos el estudio KIRP constituyen una observación influyente. Una observación detallada de estos términos reveló que estos términos GO tienden a estar también sobrerrepresentados en este estudio pero en mucha menor medida.

Incluso, en algunos casos donde el término GO aparece como sobrerrepresentado en el resto de estudios, aquí en el estudio KIRP como infrarrepresentado. Podría ser debido a problemas con el set de datos o en el enriquecimiento funcional o, si se busca una explicación biológica, a que este tipo de cáncer se rige por mecanismos moleculares muy diferentes al resto.

GO	Descripción	Rank	LOR medio	P Ajust
GO:0055119	relaxation of cardiac muscle	1-Bottom	-1.091	9×10^{-13}
GO:0043408	regulation of MAPK cascade	2-Bottom	-1.087	1×10^{-11}
GO:0042178	xenobiotic catabolic process	3-Bottom	-1.111	1×10^{-11}
GO:1901687	glutathione derivative biosynthetic process	4-Bottom	-1.060	3×10^{-11}
GO:0070527	platelet aggregation	5-Bottom	-0.934	2×10^{-8}
GO:0007052	mitotic spindle organization	1-Top	1.941	2×10^{-32}
GO:0031145	anaphase-promoting complex- [...]	2-Top	1.409	2×10^{-17}
GO:1901990	regulation of mitotic cell cycle phase [...]	3-Top	1.409	5×10^{-16}
GO:0007059	chromosome segregation	4-Top	1.342	6×10^{-14}
GO:0010972	negative regulation of G2/M transition [...]	5-Top	1.144	3×10^{-12}

Tabla 4.13 Top de términos GO enriquecidos del set de datos del TCGA según el método de combinación de rangos. Rank: lugar que ocupa el GO en el ranking, *top* en el de términos GO sobrerrepresentados, y *bottom* en el de infrarrepresentados, al realizarse el test bidireccionalmente. · LOR medio: logaritmo del *odds-ratio*. · P Ajust: p-valor ajustado por comparaciones múltiples por FDR (Benjamini-Hochberg).

4.1.2.3. Combinación de rangos

En la Tabla 4.13 aparecen los principales resultados de metaanálisis según el método de combinación de rangos. En total, este método ha dado como resultado 159 términos GO enriquecidos: 74 sobrerrepresentados en las muestras tumorales respecto a las muestras de tejidos sanos ($FDR < 0{,}05$ y $LOR > 0$), y 85 infrarrepresentados ($FDR < 0{,}05$ y $LOR < 0$).

4.1.4.4. Comparativa de todos los métodos

Como anteriormente, la comparativa se realizará utilizando gráficos de Venn para comprobar las intersecciones de términos GO significativos:

- Los dos métodos de combinación de p-valores junto al método de combinación de rangos y la intersección de términos GO significativos (Figura 4.12 a).
- Los dos métodos de combinación de tamaño del efecto (REM y FEM), el GSA desde el metaanálisis a nivel de gen y la intersección de genes significativos (Figura 4.12 b).

- Un método de cada grupo: un modelo de efectos aleatorios, una combinación ponderada de p-valores, el GSA y la combinación de rangos (Figura 4.12 c).

Además, los términos GO enriquecidos detectados por todos los métodos aparecen representados en un esquema en la Figura 4.13. Este esquema fue realizado con OmicsBox[1].

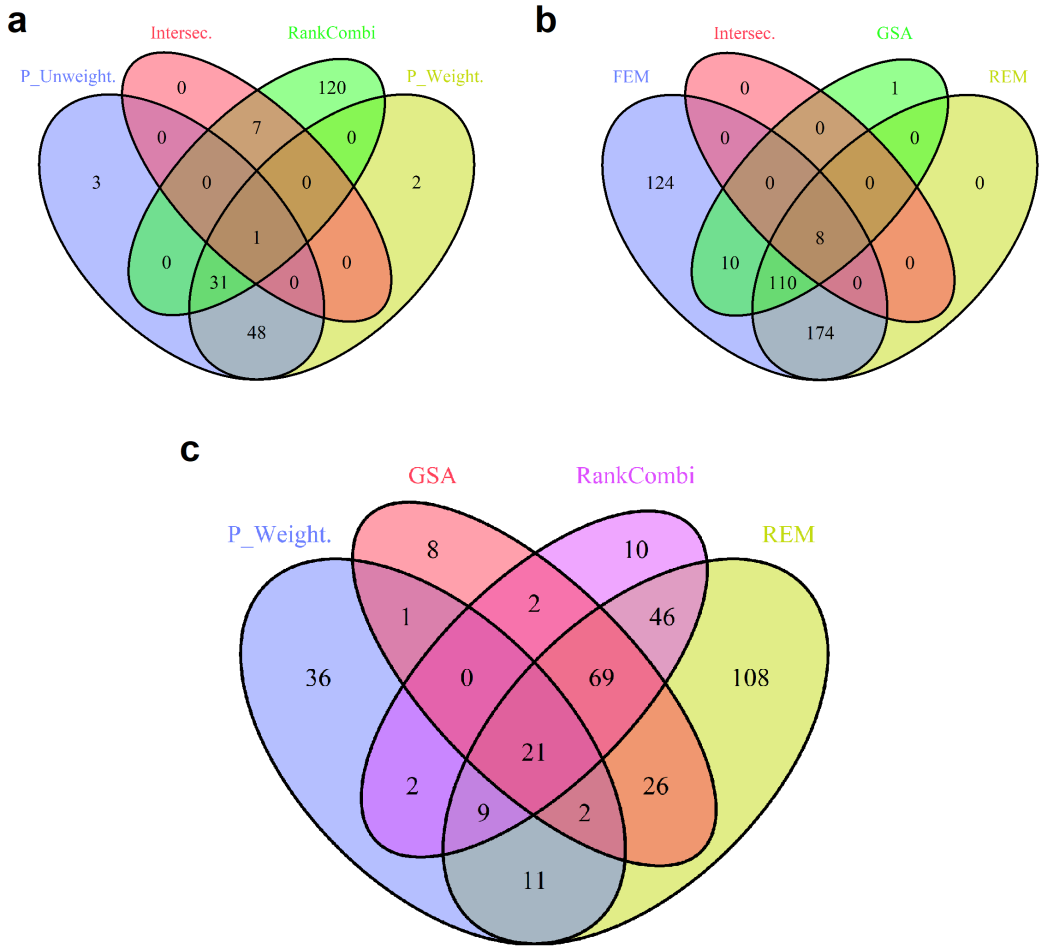

Figura 4.12 Comparativa de resultados del metaanálisis a nivel de función del set de datos de TCGA. PUnweight.: combinación de p-valores sin ponderar. · PWeight.: combinación de p-valores ponderada. · FEM: Modelo de Efectos Fijos. · REM: Modelo de Efectos Aleatorios. · RankCombi: Combinación de rangos. · Intersecc.: intersección de genes diferencialmente expresados en todos los estudios. · GSA: Enriquecimiento funcional desde el metaanálisis a nivel de gen.

[1]OmicsBox - Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, https://www.biobam.com/omicsbox

Figura 4.13 Términos GO enriquecidos (en azul los infrarrepresentados y en rojo los sobrerrepresentados, coloreados en intensidad según su FDR) en el set de datos del TCGA.

4.2. Set de datos de enfermedades dermatológicas

La principal diferencia de este set de datos de transcriptómica en enfermedades dermatológicas con el anterior es que la plataforma utilizada es el *microarray*, frente al RNA-Seq del anterior. La metodología utilizada para la expresión diferencial es diferente, pero no la del metaanálisis.

El punto de partida son directamente los resultados de la expresión diferencial, por lo que no es necesario realizar este paso ni un preprocesado de los datos crudos. Sin embargo, sí es necesario realizar un pequeño análisis exploratorio sobre esta expresión diferencial.

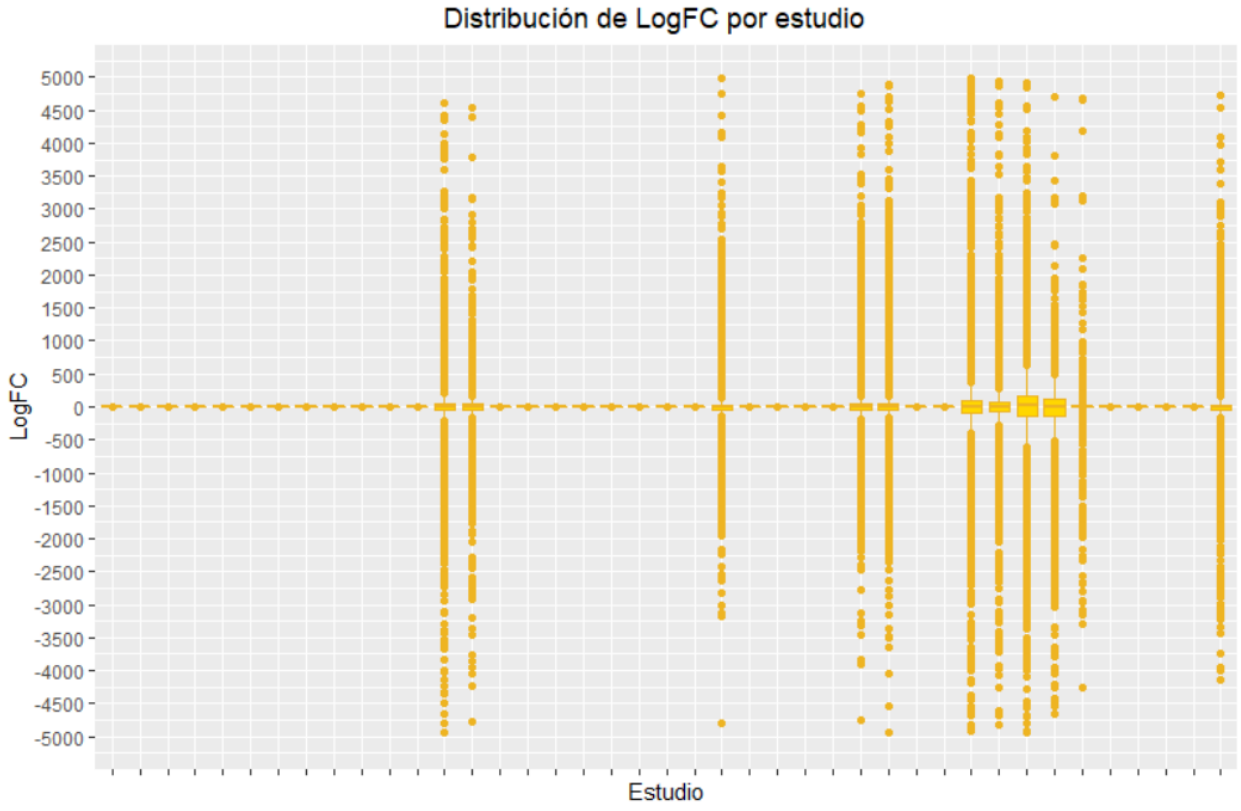

Figura 4.14 Distribución de LFC por estudio de los análisis de expresión diferencial del set de datos de enfermedades dermatológicas.

Un vistazo a la distribución de los LogFC de los genes en cada estudio (Figura 4.14) informa de que hay estudios cuyos valores son muy extremos. En general, estos valores están escala logarítmica para permitir que las diferencias entre genes con un alto nivel de expresión en una condición y muy bajo en la otra sean cuantificables. Tras descargar estos

Genes sobreexpresados			Genes infraexpresados		
Gen	LFC medio	P Combi	Gen	LFC medio	P Combi
ENSG00000206073	3.381	0.000	ENSG00000156076	-1.842	0.000
ENSG00000198074	2.747	0.000	ENSG00000156284	-1.392	0.000
ENSG00000057149	2.343	0.000	ENSG00000127324	-1.263	0.000
ENSG00000186832	2.223	0.000	ENSG00000106809	-1.166	0.000
ENSG00000012223	2.057	0.000	ENSG00000144891	-1.107	0.000

Tabla 4.14 Top 5 de genes diferencialmente expresados del set de datos de enfermedades dermatológicas ordenados por *log fold-change* según el metaanálisis con métodos de combinación de *p*-valores. LFC medio: media del *log fold-change* entre todos los estudios. · P Combi: *p*-valor combinado y ajustado por FDR.

sets de datos conflictivos de GEO y realizar el análisis de la expresión génica diferencial, se llegó a la conclusión de que el problema con ellos era el pre-procesado. Por tanto, y dado que los datos de partida eran ya los resultados de expresión diferencial, los 11 estudios con valores extremos se descartaron: los metaanálisis se realizarán con los 30 restantes.

4.2.1. Metaanálisis a nivel de gen

4.2.1.1. Métodos de combinación de *p*-valores

Ambos métodos de combinación de *p*-valores dan el mismo resultado: hay 40 genes diferencialmente expresados, 31 genes sobreexpresados en los las muestras de pacientes, y 9 infraexpresados (Tabla 4.14).

4.2.1.2. Métodos de combinación del tamaño del efecto

Como con el set de datos anterior, también se evaluarán los dos tipos de modelos: efectos fijos (FEM) y efectos aleatorios (REM).

El FEM detecta un total de 10 genes diferencialmente expresados, 8 sobreexpresados y 2 infraexpresados. El REM es capaz de detectar casi ocho veces más de genes, hasta un total de 79, estando 61 sobreexpresados y 18 infraexpresados (Tabla 4.15).

Gen	LFC	IC 95 %	τ^2	Q	QP	P	P Ajust
REM. Top 5 Genes Infraexpresados							
ENSG00000156076	-1.61	[-1.85, -1.38]	0.340	2963.9	0.000	8×10^{-41}	1×10^{-38}
ENSG00000178776	-1.53	[-2.17, -0.89]	1.929	1479.4	1×10^{-303}	2×10^{-6}	9×10^{-6}
ENSG00000174808	-1.53	[-1.75, -1.30]	0.317	3721.9	0.000	7×10^{-40}	8×10^{-38}
ENSG00000125571	-1.49	[-1.73, -1.25]	0.354	2659.8	0.000	3×10^{-34}	3×10^{-32}
ENSG00000138294	-1.38	[-1.55, -1.38]	0.140	2106.2	0.000	6×10^{-60}	3×10^{-57}
REM. Top 5 Genes Sobreexpresados							
ENSG00000214822	3.83	[2.86, 4.80]	0.341	3.264	7×10^{-2}	1×10^{-14}	1×10^{-13}
ENSG00000214856	3.66	[2.76, 4.56]	0.253	2.484	1×10^{-1}	1×10^{-15}	2×10^{-14}
ENSG00000206073	3.27	[2.84, 3.71]	1.241	10465	0.000	3×10^{-49}	8×10^{-47}
ENSG00000163220	3.01	[2.48, 3.55]	1.912	7155.4	0.000	1×10^{-28}	7×10^{-27}
ENSG00000124102	2.79	[2.33, 3.15]	1.121	27901	0.000	3×10^{-39}	4×10^{-37}
FEM. Top 5 Genes Infraexpresados							
ENSG00000188984	-1.21	[-1.33, -1.09]		131.00	2×10^{-30}	3×10^{-89}	2×10^{-87}
ENSG00000178776	-1.00	[-1.06, -0.93]		1479.4	1×10^{-303}	8×10^{-209}	4×10^{-206}
FEM. Top 5 Genes Sobreexpresados							
ENSG00000214822	3.87	[3.34, 4.41]		3.264	7×10^{-2}	4×10^{-46}	9×10^{-45}
ENSG00000214856	3.59	[3.04, 4.14]		2.484	1×10^{-1}	4×10^{-37}	6×10^{-36}
ENSG00000183760	2.17	[1.99, 2.35]		252.09	2×10^{-55}	1×10^{-120}	1×10^{-118}
ENSG00000214810	1.73	[1.50, 1.96]		6.693	9×10^{-3}	8×10^{-48}	2×10^{-46}
ENSG00000227471	1.56	[1.40, 1.72]		2.070	3×10^{-6}	1×10^{-80}	5×10^{-79}

Tabla 4.15 Top 10 de genes diferencialmente expresados del set de datos de enfermedades dermatológicas según el Modelo de Efectos Aleatorios (REM) y de Efectos Fijos (FEM) ordenados por *log fold-change*. LFC e IC 95 %: estimación del *log fold-change* combinado e intervalo de confianza al 95 %. · τ^2: medida de heterogeneidad. · Q y QP: estadístico de contraste y p-valor del test de Cochran. · P y P Ajust: p-valor de la expresión diferencial y p-valor corregido para comparaciones múltiples.

Análisis de heterogeneidad, sensibilidad y estudios influyentes

Además de las métricas de heterogeneidad obtenidas en apartados anteriores (τ^2, H^2 e I^2), los gráficos de embudo son los que aportan más información acerca de la heterogeneidad que presentan los estudios. Como muestra, se han obtenido los gráficos de embudo para el gen ENSG00000178776 (Figura 4.15).

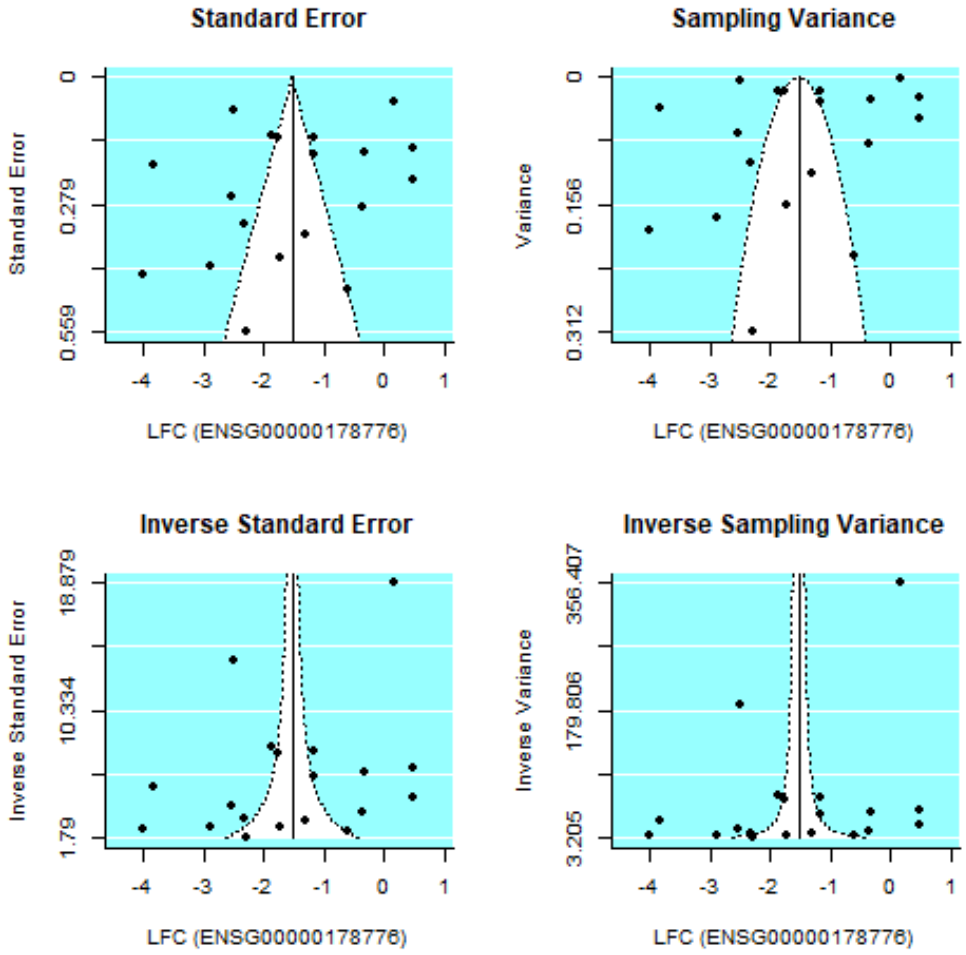

Figura 4.15 Diagramas de embudo para el gen ENSG00000178776 en el set de datos de enfermedades dermatológicas con un Modelo de Efectos Aleatorios (REM).

Estos gráficos muestran que los estudios son muy heterogéneos, quedando muchos de ellos fuera de la región de confianza de todos los gráficos. Esto se observa de manera sistemática para todos los genes del set de datos, por lo que podemos concluir que estamos delante de un conjunto de estudios muy heterogéneo.

	E. Influyentes		Sensibilidad	
Gen	**Signo LFC**	**Núm. influ.**	**Tam. Efecto**	**Significación**
ENSG00000156076	23/30	12/30	0.9072	30/30
ENSG00000178776	26/30	0/30	0.9904	30/30
ENSG00000174808	27/30	8/30	0.9177	30/30
ENSG00000125571	24/30	11/30	0.8054	30/30
ENSG00000138294	23/30	14/30	0.7210	30/30
ENSG00000214822	2/2	2/2	0.9732	2/2
ENSG00000214856	2/2	2/2	0.9342	2/2
ENSG00000206073	25/30	11/30	0.9943	30/30
ENSG00000163220	24/29	5/29	0.9790	29/29
ENSG00000124102	24/29	7/29	0.9515	29/29

Tabla 4.16 Análisis de estudios influyentes y sensibilidad para el Modelo de Efectos Aleatorios (REM) del metaanálisis a nivel de gen con el set de datos de enfermedades dermatológicas. Signo LFC: número de estudios con el mismo signo del LFC que el estimado y número de estudios sin valores faltantes para ese gen. · Núm. influ: número de estudios influyentes. · Tam. Efecto: p-valor del test t sobre los LFC obtenidos por *leave-one-out* comparado con el LFC estimado con el set de datos completo. · Significación: Número de estudios con un p-valor $< 0{,}05$ para ese gen.

En la Tabla 4.16 se pone de manifiesto que, aunque el análisis de sensibilidad da unos buenos resultados, existen muchos estudios que se detectan como influyentes según las métricas que se han descrito anteriormente (distancia de Cook, etc). Esto puede ser debido al efecto de la variabilidad que se observa en la estimación del LFC en cada estudio.

En el gráfico de bosque para el gen anterior se observan grandes diferencias en la variabilidad del tamaño del efecto estimado, no debidas a la enfermedad en estudio (Figura 4.16). Esto puede provocar que tantos estudios hayan sido declarados como influyentes.

4.2.1.3. Combinación de rangos

El método de combinación de rangos da como resultado 126 genes diferencialmente expresados: 89 sobreexpresados y 37 con infraexpresión (Tabla 4.17).

Gen	Rank	LFC medio	P Ajust
ENSG00000102891	1-Down	-1.018	2×10^{-42}
ENSG00000174808	2-Down	-1.597	4×10^{-37}
ENSG00000178776	3-Down	-1.550	3×10^{-35}
ENSG00000125571	4-Down	-1.702	4×10^{-35}
ENSG00000156076	5-Down	-1.842	4×10^{-35}
ENSG00000214856	1-Up	3.714	2×10^{-70}
ENSG00000214822	2-Up	3.811	6×10^{-65}
ENSG00000163220	3-Up	3.071	5×10^{-55}
ENSG00000206073	4-Up	3.381	9×10^{-51}
ENSG00000143556	5-Up	2.103	1×10^{-50}

Tabla 4.17 Top de genes diferencialmente expresados del set de datos de enfermedades dermatológicas según el método de combinación de rangos. Rank: lugar que ocupa el gen en el ranking, *up* en el de genes sobreexpresados, y *down* en el de infraexpresados, al realizarse el test bidireccionalmente. · LFC medio: logaritmo del *fold-change* medio. · P Ajust: p-valor ajustado por FDR (Benjamini-Hochberg).

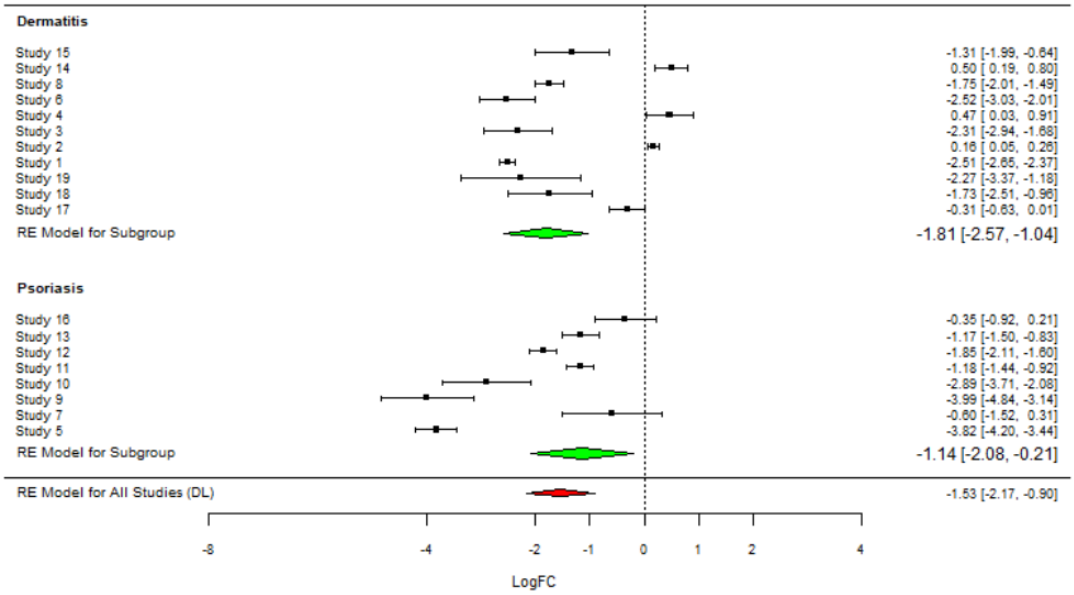

Figura 4.16 Diagrama de bosque para el gen ENSG00000178776 en el set de datos de enfermedades dermatológicas con un Modelo de Efectos Aleatorios (REM). Se representa un metaanálisis individual para cada enfermedad y el metaanálisis global.

4.2.1.4. Comparativa de todos los métodos

La comparativa (Figura 4.17) se volverá a realizar a tres niveles, igual que la que se realizó con el set de datos anterior.

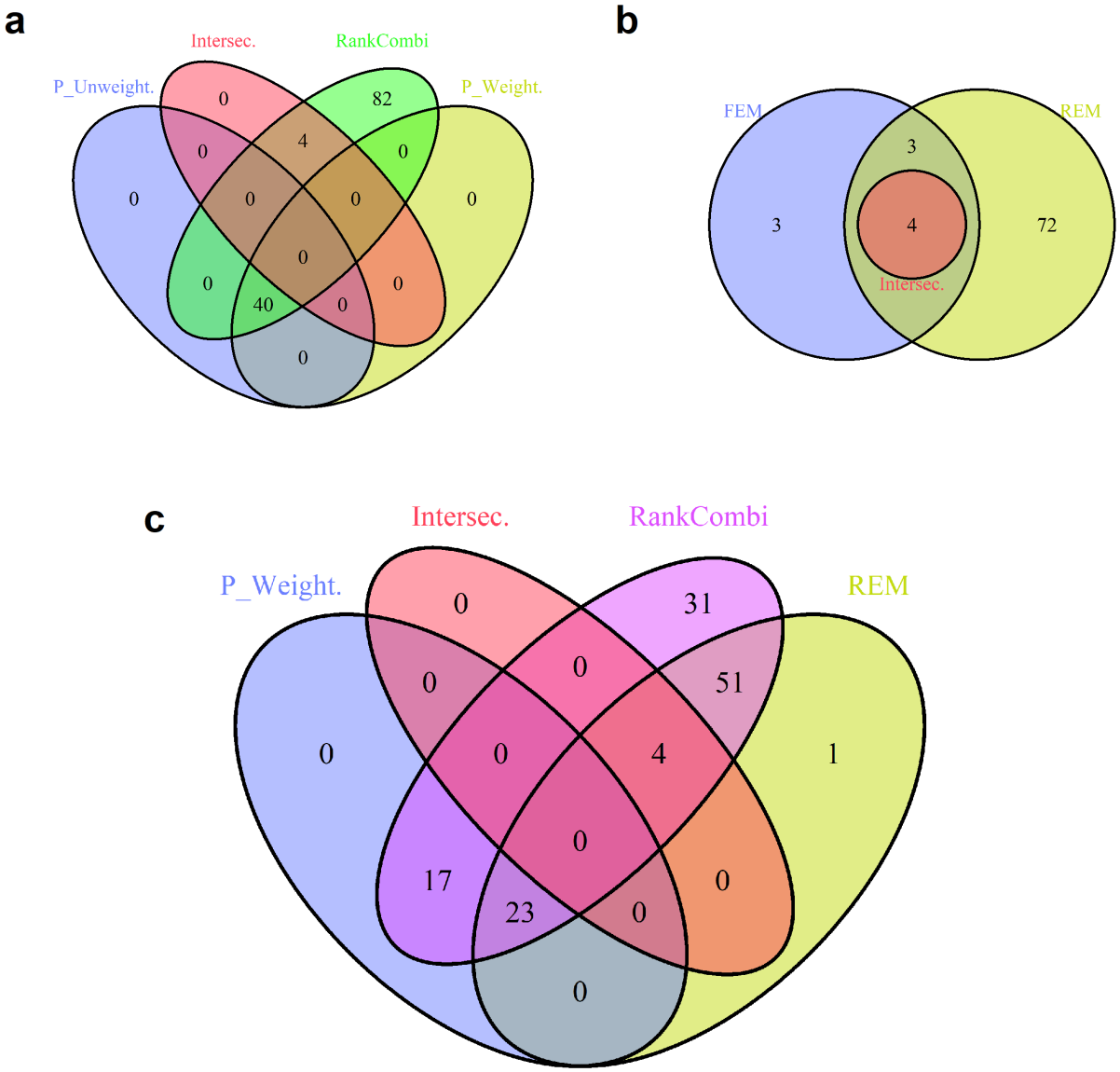

Figura 4.17 Comparativa de resultados del metaanálisis a nivel de función del set de datos de enfermedades dermatológicas. PUnweight.: combinación de p-valores sin ponderar. · PWeight.: combinación de p-valores ponderada. · FEM: Modelo de Efectos Fijos. · REM: Modelo de Efectos Aleatorios. · RankCombi: Combinación de rangos. · Intersecc.: intersección de genes diferencialmente expresados en todos los estudios. · GSA: Enriquecimiento funcional desde el metaanálisis a nivel de gen.

Estudio	Top	Bottom	Estudio	Top	Bottom
GSE11903	6	1	GSE31835(1)	2	19
GSE12511	10	5	GSE31835(2)	0	0
GSE13355(1)	23	5	GSE32407	2	1
GSE13355(2)	3	1	GSE32924(1)	2	1
GSE14905(1)	17	8	GSE32924(1)	4	5
GSE14905(2)	4	3	GSE36482(1)	1	3
GSE16161(1)	0	0	GSE36482(2)	0	1
GSE16161(2)	1	0	GSE36482(2)	0	20
GSE18686	18	3	GSE40263	3	58
GSE26866	26	0	GSE52471	8	1
GSE26952(1)	1	0	GSE53431	30	4
GSE26952(2)	2	0	GSE6281(1)	4	0
GSE27887	26	6	GSE6281(2)	1	0
GSE30768	3	3	GSE6281(3)	9	0
GSE30999	16	3	GSE6281(4)	8	1

Tabla 4.18 Términos GO enriquecidos en los 30 estudios de enfermedades dermatológicas. En la columna *top* aparecen representados los términos GO sobrerrepresentados en las muestras de pacientes, y en la columna *bottom* los infrarrepresentados.

4.2.2. Enriquecimiento funcional

El número de términos GO enriquecidos para este set de datos, sobrerrepresentados (*top*) e infrarrepresentados (*bottom*), aparecen detallados en la Tabla 4.18.

4.2.3. Metaanálisis a nivel de función

La intersección de términos GO a lo largo de todos los estudios no dio como resultado ninguna función molecular significativa. Realizar un análisis de enriquecimiento funcional sobre los resultados del metaanálisis de gen con un REM dio como resultado 108 términos GO enriquecidos: 29 infrarrepresentados y 79 sobrerrepresentados.

El metaanálisis funcional de este set de datos sigue las mismas técnicas que los demás. Al no haberse encontrado problemas ni comportamientos extraños en los datos (estudios demasiado heterogéneos, muchos estudios influyentes, p-valores o tamaños del efecto demasiado extremos, por ejemplo) sólo se enunciarán los resultados principales en cada apartado.

4.2.3.1. Métodos de combinación de p-valores

El método de combinación de p-valores ponderado dio como resultado 82 términos GO enriquecidos, 53 sobrerrepresentados y 29 infrarrepresentados en los pacientes. El método no ponderado detectó 83 términos GO enriquecidos, 55 y 28 sobrerrepresentados e infrarrepresentados, respectivamente.

4.2.3.2. Métodos de combinación del tamaño del efecto

El Modelo de Efectos Fijos (FEM) detectó 321 términos GO enriquecidos, 194 sobrerrepresentados y 127 infrarrepresentados. Con el Modelo de Efectos Aleatorios (REM) los términos GO enriquecidos fueron también 321: 195 sobrerrepresentados y 126 infrarrepresentados.

Análisis de heterogeneidad, sensibilidad y estudios influyentes

Para 182 términos GO (un 74 % del total) no existe ningún estudio influyente. Para el resto, ningún estudio destaca especialmente por detectase como influyente para un gran número de anotaciones GO, por lo que no hay problemas de influencia.

En el análisis de sensibilidad, sólo para un estudio todos los tests t dieron como resultado un $p > 0{,}05$. Además, para todos los términos GO enriquecidos, los p-valores del enriquecimiento fueron significativos para todos los estudios.

4.2.3.3. Combinación de rangos

El método de combinación de rangos da como resultado 231 términos GO enriquecidos: 113 sobrerrepresentados y 118 infrarrepresentados en los pacientes frente a los controles.

4.2.3.4. Comparativa de todos los métodos

La comparativa de todos los métodos en diagramas de Venn aparece en la Figura 4.18, y el resultado esquemático de los términos GO enriquecidos detectados por todos los métodos se ilustra en la Figura 4.19.

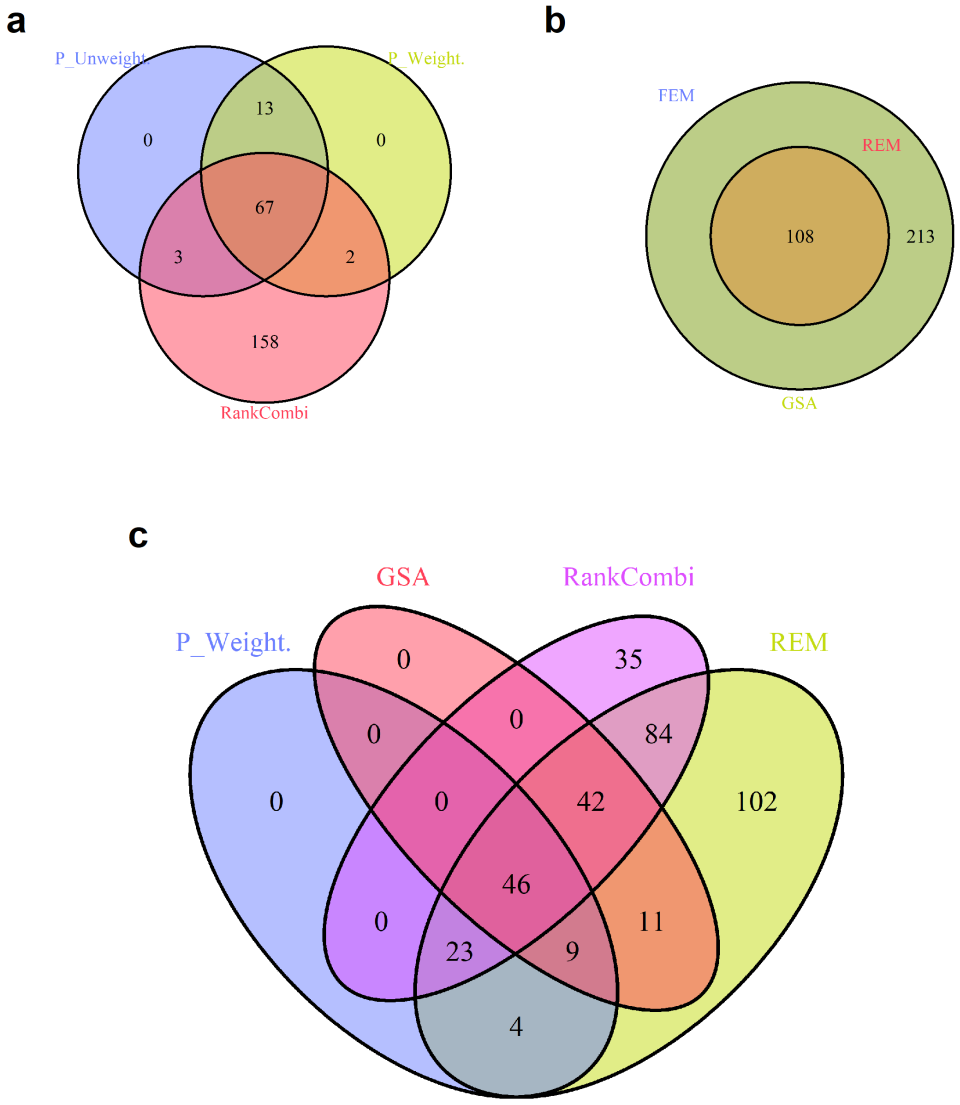

Figura 4.18 Comparativa de resultados del metaanálisis a nivel de función del set de datos de enfermedades dermatológicas. PUnweight.: combinación de p-valores sin ponderar. · PWeight.: combinación de p-valores ponderada. · FEM: Modelo de Efectos Fijos. · REM: Modelo de Efectos Aleatorios. · RankCombi: Combinación de rangos. · Intersecc.: intersección de genes diferencialmente expresados en todos los estudios. · GSA: Enriquecimiento funcional desde el metaanálisis a nivel de gen.

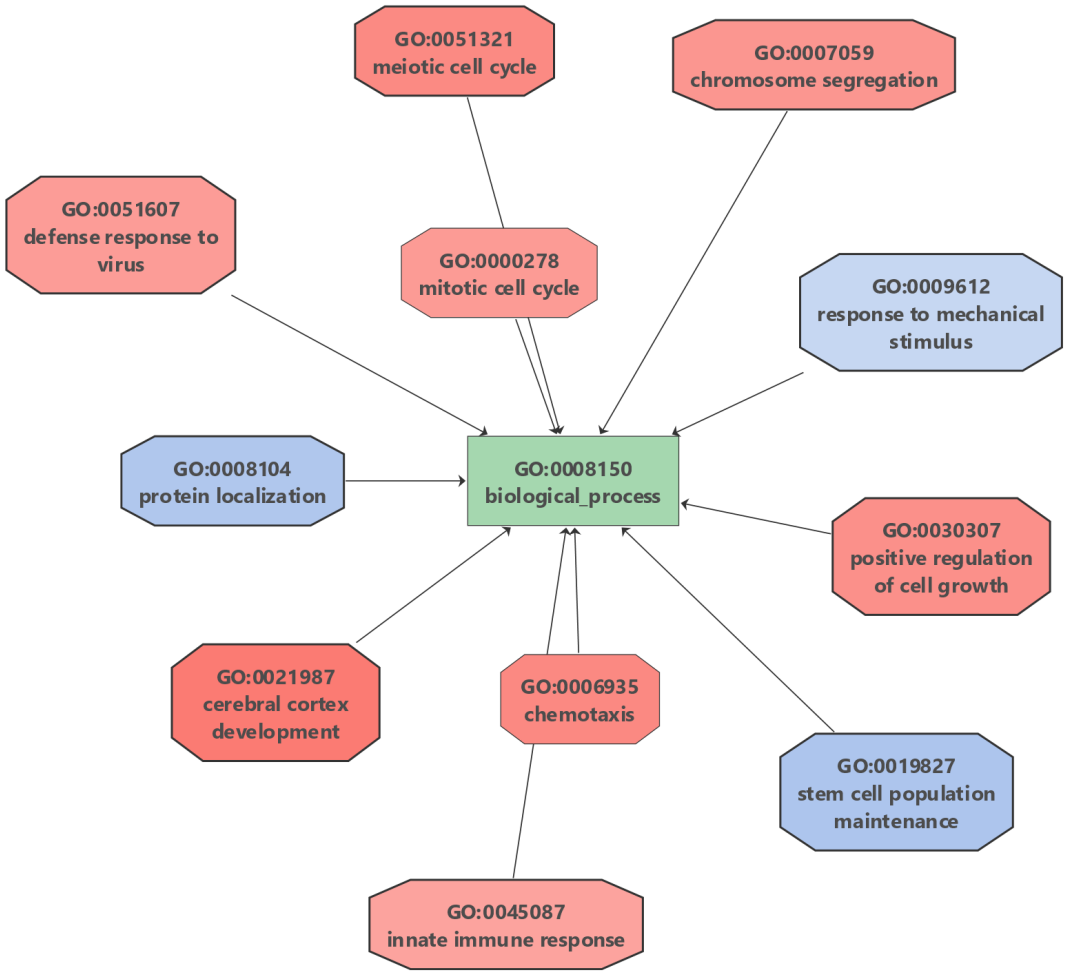

Figura 4.19 Términos GO enriquecidos (en azul los infrarrepresentados y en rojo los sobrerrepresentados, coloreados en intensidad según su FDR) en el set de datos de enfermedades dermatológicas.

Para ambos sets de datos, el resumen de los genes diferencialmente expresados y los términos GO enriquecidos detectados por cada uno de los métodos aparece representado en la Tabla 4.19, en la siguiente página.

Metaanálisis a nivel de Gen			
Método de metaanálisis	***Up***	***Down***	**Total**
Set de datos de TCGA			
Combinación ponderada de p-valores (normal inversa)	706	1596	2302
Combinación no ponderada de p-valores (Fisher)	706	1596	2302
Modelo de Efectos Fijos (FEM)	1423	3509	4932
Modelos de Efectos Aleatorios (REM)	1329	2762	4091
Combinación de rangos	1223	2941	4164
Set de datos de enfermedades dermatológicas			
Combinación ponderada de p-valores (normal inversa)	31	9	40
Combinación no ponderada de p-valores (Fisher)	31	9	40
Modelo de Efectos Fijos (FEM)	8	2	10
Modelos de Efectos Aleatorios (REM)	61	18	79
Combinación de rangos	89	37	126
Metaanálisis a nivel de Función (GO)			
Método de metaanálisis	***Top***	***Bottom***	**Total**
Set de datos de TCGA			
Combinación ponderada de p-valores (normal inversa)	44	38	82
Combinación no ponderada de p-valores (Fisher)	46	37	83
Modelo de Efectos Fijos (FEM)	208	218	426
Modelos de Efectos Aleatorios (REM)	148	144	292
Combinación de rangos	74	85	159
Set de datos de enfermedades dermatológicas			
Combinación ponderada de p-valores (normal inversa)	53	29	82
Combinación no ponderada de p-valores (Fisher)	55	28	83
Modelo de Efectos Fijos (FEM)	194	127	321
Modelos de Efectos Aleatorios (REM)	195	126	321
Combinación de rangos	113	118	231

Tabla 4.19 Número de genes diferencialmente expresados y términos GO enriquecidos detectados con cada técnica de metaanálisis en el set de datos de tumores y de enfermedades dermatológicas. En metaanálisis a nivel de gen, *up*: genes sobreexpresados, y *down*: genes infraexpresados. En metaanálisis a nivel de función, *top*: GOs sobrerrepresentados, y *bottom*: GOs infrarrepresentados.

Capítulo 5

Discusión

Los estudios de transcriptómica requieren un análisis bastante complejo hasta extraer conclusiones biológicas: análisis exploratorio, análisis diferencial de expresión, anotación funcional y enriquecimiento funcional. Además, para combinar los resultados obtenidos en varios estudios es necesario emplear técnicas de metaanálisis. En este último paso no existe un estándar: hay una gran cantidad de métodos que en este trabajo hemos revisado tanto a nivel de gen como a nivel de función molecular.

El análisis de expresión diferencial para datos de RNA-Seq, el que se ha utilizado en este trabajo, se basa en un Modelo Lineal Generalizado de la familia Binomial Negativa al estar formados estos sets de datos por conteos. Estos modelos no están exentos de problemas: los datos de RNA-seq suelen ser muy irregulares en cuanto a nivel de secuenciación por muestra y a otras consideraciones, por lo que lo más importante es implementar un buen método de normalización. Numerosos paquetes de R permiten normalizar estos datos previamente a la aplicación del modelo, siendo algo común en el mundo de la Bioinformática. Es también necesario, por tanto, realizar un análisis exploratorio exhaustivo que permita observar la distribución de conteos por muestra para comprobar que la normalización haya funcionado correctamente. En el set de datos del TCGA, el único al que se le ha aplicado un análisis de expresión diferencial, la normalización hace que la distribución de conteos sea homogénea en todas las muestras de todos los estudios, y demuestra que estos métodos son muy eficaces aplicados a cualquier set de datos de RNA-Seq.

El paso de enriquecimiento funcional que se ha seguido en este trabajo está basado en un

método de grupos de genes (GSA). El método que hemos empleado se basa en un modelo estadístico muy sencillo: una regresión logística. Este modelo, además de sencillo, consigue rescatar las anotaciones funcionales enriquecidas en sets de datos de manera muy eficiente. Además, una de las principales ventajas de esta técnica frente a otros métodos tradicionales de enriquecimiento, es que al estar basado en una regresión logística se pueden añadir covariables que no estén directamente relacionadas con la expresión diferencial (variables clínicas, ambientales, etc.) aunque en este trabajo no se hayan utilizado.

El metaanálisis es el paso que ha cubierto el grueso del trabajo. Estos métodos son muy interesantes, ya que son muy potentes a la hora de detección genes y funciones moleculares significativas y además permiten la generación de información relevante mediante abordajes *in silico*, sin la enorme inversión económica que requieren los experimentos ómicos. Estos abordajes también permiten combinar la información disponible en repositorios públicos con la información obtenida con un estudio propio.

A modo de reflexión, es importante destacar que uno de los pasos más críticos a la hora de realizar un metaanálisis es la selección de estudios. Para poder disponer de datos de estudios ya realizados es necesario que los autores hayan depositado sus datos y resultados en repositorios públicos. Actualmente, en la era del *big data*, es muy común que investigadores y revistas científicas quieran hacer accesibles sus resultados en estos repositorios para que sean aprovechables. Es por esto que se dice que actualmente la comunidad científica está volcada con la *Open Science*.

Sin embargo, algo que debe mejorar de este aspecto es el *cómo*: el cómo manejar y administrar datos científicos. En 2016 se publicaba en Nature un artículo hablando de los principios para manejar, administrar y almacenar datos científicos: los **principios FAIR** (Wilkinson y col., 2016)[1], por las siglas en inglés *Findable, Accessible, Interoperable* y *Reusable*. Estos principios actúan como guía sobre cómo operar con los datos obtenidos en investigaciones científicas para poder sacar el máximo partido a la información ya publicada. La implementación de estos principios sería ideal para la búsqueda y selección de estudios para un metaanálisis, y podría ayudar a que este tipo de técnicas fueran más frecuentes.

El metaanálisis de datos ómicos no tiene ningún método estándar establecido. En general,

[1] datos.gob.es/es/noticia/principios-fair-buena-practicas-para-la-gestion-y-administracion-de-datos-cientificos

todos los métodos son potentes frente a una combinación de estudios simple, como la que podría ser observar los genes y las funciones que han resultado significativos en todos los estudios a la vez (Figura 5.2).

El metaanálisis a nivel de función, en general, aporta mucha más información en el marco de la biología de sistemas que el de gen al desvelar en qué procesos o rutas metabólicas están involucrados los genes diferencialmente expresados. El metaanálisis a nivel de gen, sin embargo, también puede resultar interesante en determinados experimentos, como por ejemplo, en búsqueda de marcadores genéticos específicos, donde la función molecular no es un aspecto tan relevante.

Los **métodos de combinación de *p*-valores** representan el escenario más simple del metaanálisis. Solo se combinan los valores de significación y en ningún momento se tiene en cuenta el tamaño del efecto, aunque realmente resulta muy importante conocerlo, ya que representa la representación diferencial de genes y funciones en una condición respecto a la otra. En otras palabras, aunque la significación es crucial, pueden existir genes con una diferencia de expresión muy pequeña entre condiciones que se detecten como significativos aunque no sean de interés biológico.

Además, los métodos de combinación de p-valores, debido principalmente a esta limitación, requieren un preprocesado de los datos comprobando comportamientos conflictivos de genes y funciones entre estudios (p.ej. genes sobreexpresados en unos estudios e infraexpresados en otros). En ambos sets de datos, gran parte de los genes y funciones que han resultado significativos mostraban comportamientos conflictivos y tuvieron que ser excluidos. Esto demuestra que estos métodos pueden no ser los apropiados a la hora de realizar un metaanálisis.

En el siguiente escalón de complejidad, se situarían los **métodos de combinación de rangos**. Estos métodos sí tienen en cuenta el tamaño del efecto al construir el ranking de genes y funciones, pero basan su resultado en permutaciones y no permiten el enunciado de un modelo ni estiman ningún parámetro. El paso de preprocesado de los datos ya no se requiere, por lo que pueden aportar resultados más fiables que los métodos anteriores.

Por último, los métodos más complejos son los **modelos de combinación del tamaño del efecto**. Estos modelos permiten estimar la variabilidad de cada estudio, los efectos de

la heterogeneidad entre estudios para modelos de efectos aleatorios, el tamaño del efecto global, etc. y resultan más comprensibles en términos de todos los parámetros que se pueden estimar.

Además, estos métodos permiten validar el modelo utilizando los parámetros de heterogeneidad que proporciona el modelo. También permiten incorporar el análisis de sensibilidad y de estudios influyentes para obtener un mejor diagnóstico de los resultados del metaanálisis. Todo esto hace que sean los más apropiados para metaanalizar datos de esta naturaleza.

Estos métodos han sido los que han dado como resultado más genes y funciones significativas en ambos sets de datos, por lo que también son unos métodos muy potentes a la hora de la detección de genes diferencialmente expresados o funciones moleculares enriquecidas. El hecho de que el modelo con efectos aleatorios permita modelizar también un parámetro de heterogeneidad hace que este tipo de test pueda aplicarse también entre estudios que han empleado tecnologías de secuenciación, análisis de expresión diferencial o métodos de enriquecimiento funcional diferentes y que sean, por ello, heterogéneos entre sí.

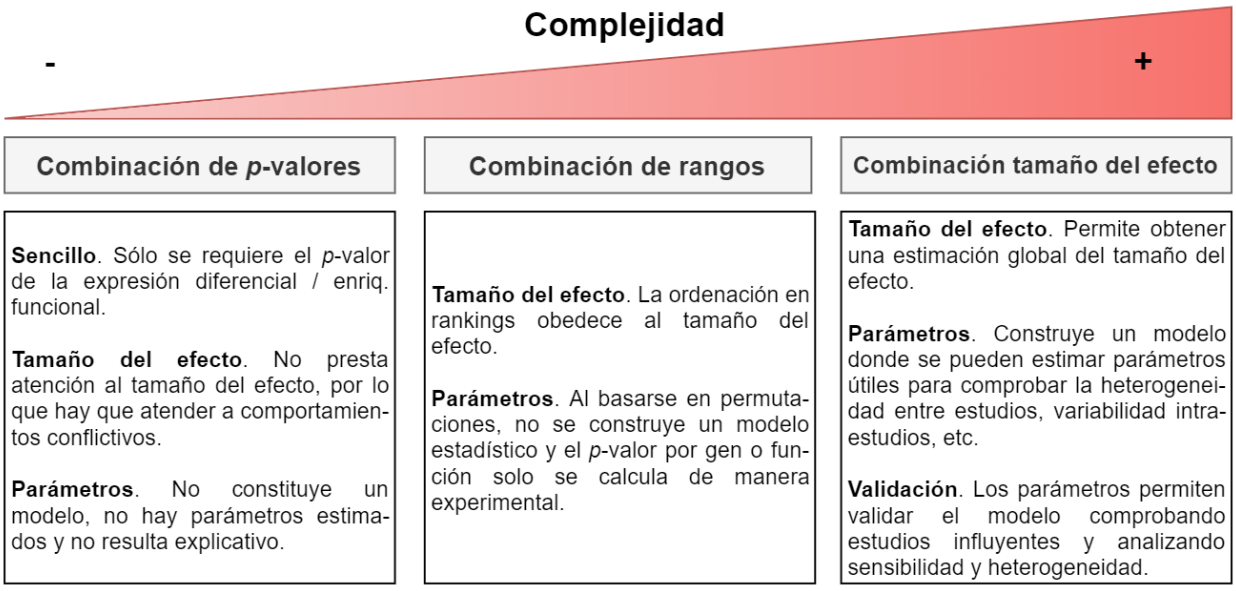

Figura 5.1 Resumen y evaluación de los métodos de metaanálisis empleados.

Capítulo 6

Conclusiones

1. La combinación de estudios en un metaanálisis es una herramienta muy potente para la generación de nuevos datos y resultados en un entorno computacional.

2. Los metaanálisis de datos ómicos no tienen una metodología estándar: existen múltiples métodos, cada uno con sus ventajas e inconvenientes. En cuanto a aplicar las técnicas de metaanálisis sobre los datos de expresión o sobre las funciones moleculares, hacerlo a nivel de función demuestra ser una mejor herramienta para obtener una panorámica de qué reacciones bioquímicas ocurren en cada condición experimental.

3. La evaluación y comparación de estos métodos se ha realizado sobre estudios de transcriptómica, pero su uso sería extendible a proteómica, metabolómica, epigenómica, etc.

4. De entre todos los métodos de metaanálisis empleados, los modelos de efectos aleatorios dentro de la combinación del tamaño del efecto demuestran ser los más adecuados, ya que permiten enunciar un modelo y estimar todos sus parámetros. Además, dentro de estos parámetros, se puede modelizar la heterogeneidad entre estudios, muy común en estudios ómicos realizados con diferentes tecnologías o en diferentes laboratorios.

5. Cabe destacar también que aunque el metaanálisis ha demostrado ser una técnica muy solvente y potente, su principal limitación es la búsqueda y selección de estudios: no existe un criterio único sobre las guías para almacenar datos en los distintos repositorios públicos ni, por tanto, para recuperarlos. Sin embargo, estándares como los principios FAIR, aunque muy limitados, comienzan a ser de uso frecuente en estos repositorios.

6. Como perspectivas futuras, señalar que este trabajo continuará con la configuración de estrategias de metaanálisis en otros tipos de resultados funcionales, como son los obtenidos en los IPA (*Inference Pathways Analysis*), que tienen como *input* las señales de expresión de los genes, los cuales son transformadas en señales de activación y recorren las rutas de señalización que describen una función biológica. Con un enfoque más aplicado, los resultados de este trabajo se utilizarán de forma inmediata en un proyecto que se ha iniciado en la Unidad de Bioinformática y Bioestadística del CIPF sobre la caracterización de las diferencias de sexo en enfermedades neurodegenerativas, donde nos centraremos en la identificación de los mecanismos diferenciales en Alzheimer, Parkinson y esclerosis múltiple entre hombres y mujeres, para mejorar su diagnóstico y tratamiento.

Referencias

Anders, S. y Huber W. (2010). "Differential expression analysis for sequence count data". En: *Genome Biology* 11.R106.

Benjamini, Y. y Hochberg Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". En: *Royal Statistical Society* 57.1.

Bentley, D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., Hall K.P., Evers D.J., Barnes C.L. y Bignell H.R. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". En: *Nature* 456, págs. 53-59.

Breitlinga, R., Armengauda P., Amtmanna A. y Herzyk P. (2004). "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". En: *Federation of European Biochemical Societies* 573, págs. 83-92.

Chen, H. y Boutros P.C. (2011). "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R". En: *BMC Bioinformatics* 12.35.

Choi, J.K., Yu U., Kim S. y Joon-Yoo O. (2003). "Combining multiple microarray studies and modeling interstudy variation". En: *Bioinformatics* 19.1, págs. i84-i90.

Cochran, W.G. (1954). "The Combination of Estimates from Different Experiments". En: *Biometrics* 10.1, págs. 101-129.

Costa, F.F. (2014). "Big data in biomedicine". En: *Drug Disc Today* 19.4, págs. 433-440.

Crabu, S. (2016). "Translational biomedicine in action: Constructing biomarkers across laboratory and benchside". En: *Soc Theory Health* 14, págs. 312-331.

Curry, S.H. (2008). "Translational science: past, present, and future". En: *Biotechniques* 44, págs. 1-8.

Edgar, R., Domrachev M. y Lash A.E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". En: *Nucleic Acid Research* 30.1, págs. 207-210.

Fang, J.Y. y Richardson B.C. (2005). "The MAPK signalling pathways and colorectal cancer". En: *The Lancet Oncology* 6.5, págs. 322-327.

Fleischmann, R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A. y Merrick J.M. (1995). "Whole-genome Random Sequencing and Assembly of Haemophilus Influenzae Rd". En: *Science* 269.5223, págs. 496-512.

GOConsortium (2004). "The Gene Ontology (GO) database and informatics resource". En: *Nucleic Acid Research* 32.1, págs. D258-D261.

Gutterson, N. y Zhang J. (2004). "Genomics applications to biotech traits: a revolution in progress?" En: *Current Opinion in Plant Biology* 7.2, págs. 226-230.

Higgins, J.P.T. y Thompson S.G. (2002). "Quantifying heterogeneity in a meta-analysis". En: *Statistics in Medicine* 21, págs. 1539-1558.

Hong, F., Breitling R., McEntee C.W., Wittner B., Nemhauser J.L. y Chory J. (2006). "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis". En: *Bioinformatics* 22.22, págs. 2825-2827.

James, G., Witten D., Hastie T. y Tibshirani R. (2013). *An Introduction to Statistical Learning. with applications in R*. Springer.

Kanehisa, M. y Goto S. (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". En: *Nucleic Acid Research* 28.1, págs. 27-30.

Kolde, R., Laur S., Adler P. y Vilo J. (2012). "Robust rank aggregation for gene list integration and meta-analysis". En: *Bioinformatics* 28.4, págs. 573-580.

Li, B. y Dewey C.N. (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". En: *BMC Bioinformatics* 12.323.

Love, M., Huber W. y Anders S. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". En: *Genome Biology* 15.12, pág. 550.

Marot, G. y Mayer C.D. (2009). "Sequential Analysis for Microarray Data Based on Sensitivity and Meta-Analysis". En: *Stat Appl Genet Mol Biol* 8.3.

Marot, G., Foulley J.L., Mayer C.D. y Jaffrezic F. (2009). "Moderated effect size and P-value combinations for microarray meta-analyses". En: *Bioinformatics* 25.20, págs. 2692-2699.

Montaner, D. y Dopazo J. (2010). "Multidimensional Gene Set Analysis of Genomics Data". En: *PLoS one* 5.4.

Nielsen, J. y Oliver S. (2005). "The next wave in metabolome analysis". En: *Trends in Biotechnology* 23.11, págs. 544-546.

Rau, A., Marot G. y Jaffrézic F. (2015). "Differential meta-analysis of RNA-seq data from multiple studies". En: *BMC Bioinformatics* 15.91.

Schmutz, J., Wheeler J., Grimwood J., Dickson M. y Yang J. (2004). "Quality assessment of the human genome sequence". En: *Nature* 429, págs. 365-368.

Seringhaus, M. y Gerstein M. (2009). "¿Qué es la ontología génica?" En: *Investigación y Ciencia* 390, págs. 73-81.

Soneson, C. y Delorenzi M. (2013). "A comparison of methods for differential expression analysis of RNA-seq data". En: *BMC Bioinformatics* 14.91.

Subramanian, A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L. y Gillette M.A. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". En: *Proc Natl Acad Sci USA* 102.43, págs. 15545-15550.

Tarca, A.L., Bhatti G. y Romero R. (2013). "A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity". En: *PLoS One* 8.11, e79217.

Viechtbauer, W. (2010). "Conducting meta-analyses in R with the metafor package". En: *Journal of Statistical Software* 36.3, págs. 1-48.

Viechtbauer, W. y Cheung M.W.L. (2010). "Outlier and influence diagnostics for meta-analysis". En: *Research Synthesis Methods* 1.2, págs. 112-125.

Wang, Z., Gerstein M. y Snyder M. (2009). "RNA-Seq: a revolutionary tool for transcriptomics". En: *Nature Reviews Genetics* 10.1, págs. 57-63.

Wasinger, V.C., Cordwell S.J., Cerpa-Poljak A., Yan J.X., Gooley A.A. y Wilkins M.R. (1995). "Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium". En: *Electrophoresis* 16.1.

Wilkinson, M.D. y col. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". En: *Scientific Data* 3.160018.