ETSE-UV
Escola Tècnica Superior d'Enginyeria

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

INTEGRACIÓN DE DATOS TRANSCRIPTÓMICOS Y METABOLÓMICOS EN ESTUDIOS ONCOLÓGICOS

AUTORA:
AYELÉN I. ROJAS BENEDICTO

DIRECTORES:
FRANCISCO GARCÍA-GARCÍA
LEONOR PUCHADES-CARRASCO

TUTORES:
VICENTE ARNAU LLOMBART

SEPTIEMBRE 2017

Resumen

El cáncer es la principal causa de muerte en los países industrializados, sólo detrás de las enfermedades cardiovasculares. En la última década, se han realizado importantes avances en el conocimiento de las bases moleculares del cáncer. Sin embargo, los métodos tradicionales de diagnóstico del cáncer presentan importantes limitaciones para conocer el pronóstico del paciente y predecir la respuesta individual a los tratamiento disponibles.

El objetivo principal de este proyecto es integrar los datos de origen metabolómico y transcriptómico para obtener una visión global del proceso de oncogénesis. Para ello, se seleccionaron datos de transcriptómica de libre acceso compatibles con estudios de metabolómica previos de la Unidad Mixta CIPF-IISLAFE de Metabolómica. A continuación, se realizó un metaanálisis en el que se identificaron las rutas que se encuentran alteradas de manera general en los distintos tipos de cáncer. Con esta información se procedió a la integración de los datos, que permitieron caracterizar los procesos moleculares y metabolómicos de la enfermedad.

Los resultados obtenidos han permitido obtener una perspectiva global de la enfermedad, al detectarse procesos típicos encontrados en las células cancerosas. Además, se han identificado posibles mejoras al método empleado, en las que se deberá profundizar en futuros trabajos.

Índice general

Abreviaturas

1D, 2D	Una dimensión, dos dimensiones
ADN	Ácido desoxirribonucleico
ADNc	Ácido desoxirribonucleico complementario
ARN	Ácido ribonucleico
ARNm	ARN mensajero
ARNt	ARN de transferencia
CLL	Leucemia linfática crónica.
CP	Cáncer de Pulmón
CPMG	Carr-Purcell-Meiboom-Gill
CRAB	Hipercalcemia, insuficiencia renal, anemia y lesiones óseas. *Hypercalcemia, renal insufficiency, anemia, bone lesions.*
Da	Dalton, unidad de masa atómica
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
GEO	*Gene Expression Omnibus.*Compilación de expresiones génicas.
GO	*Gene Ontology,Ontología de genes*
GSEA	*Gene Set Enrichment Analysis.* Análisis de Enriquecimiento de Grupos de Genes
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MM	Mieloma múltiple
NMP	Neoplasma mieloproliferativo
NOESY	Espectroscopía de efecto nuclear Overhauser. Nuclear Overhauser effect spectroscopy
OPLS-DA	*Orthogonal Projections to Latent Structures Discriminant Analysis. Análisis discriminante de Proyecciones Ortogonales a las Estructuras Latentes*
PCA	*Principal Component Analysis.Análisis de Componentes Principales*
Poli(A)	Poliadenosín fosfato
REDECAN	Red Española de Registros de Cáncer
RMA	*Robust Multi-array Average*
RMN	Resonancia Magnética Nuclear
SEOM	Sociedad Española de Oncología Médica
TSP	3-(Trimetilsilil)propionato

I. Introducción

1 Cáncer

El término cáncer engloba a un conjunto de enfermedades que se caracterizan por un desarrollo anormal de sus células, que se dividen, crecen y diseminan a otros tejidos del organismo. Se han establecido diez características esenciales (Hanahan & Weinberg, 2011) que distinguen a una célula tumoral de una normal:

- Se dividen sin necesidad de señales externas que estimulen el crecimiento.
- Su genoma es inestable y mutable.
- Crecen aún en presencia de señales inhibidoras de crecimiento.
- Evaden la apoptosis.
- Muestran angiogénesis sostenida.
- Provocan inflamación local.
- Tienen un potencial infinito de replicación.
- Son capaces de producir metástasis.
- Presentan cambios en el metabolismo energético.
- Evaden al sistema inmunitario.

Todos estos cambios son generados por una o más mutaciones a nivel genético que convierten células normales en células tumorales (Fig. 1).

Figura 1 . Características esenciales de las células tumorales (Adaptada de Hanahan & Weinberg, 2011).

Tras las enfermedades cardiovasculares, el cáncer es la principal causa de muerte en los países industrializados. Los últimos datos disponibles, a nivel mundial, sobre la incidencia del cáncer datan de 2012 y fueron realizados por el proyecto GLOBOCAN de la Organización Mundial de la Salud (OMS). Se estimó que en ese año, un total de 32.6 millones de personas padecían cáncer, 14.4 millones de personas fueron diagnosticadas y al menos 8.2 millones de muertes fueron a consecuencia de esta enfermedad. A nivel mundial los tipos de cáncer con mayor incidencia son los de pulmón, mama, colorrectal,

próstata estómago e hígado (GLOBOCAN, 2012). En España, según un registro de 2017 de REDECAN (Red Española de Registro de Cánceres), el número de nuevos casos fue de 247,771 personas. Siendo los tumores más frecuentemente diagnosticados los colorrectales, de próstata, pulmón, mama y vejiga (Puchades-Carrasco, 2013).

En la última década, se han realizado importantes avances en el conocimiento de las bases moleculares del cáncer. A pesar de ello, las técnicas actuales de diagnóstico son incapaces de predecir el pronóstico del paciente. Así como predecir la eficacia de los fármacos disponibles, que en algunos casos no es siempre efectivo.
El objetivo principal de este proyecto es integrar la metabolómica y transcriptómica para que, con esta nueva visión global del proceso de oncogénesis, sea posible identificar nuevas dianas terapéuticas y posibles biomarcadores.
En este estudio se trabajará con datos procedentes de 4 tipos de tumores concretos. Tres de los tumores son de origen hematológico: leucemia linfática crónica, mieloma múltiple y neoplasias mieloproliferativas. El cuarto es cáncer de pulmón y se agrupa de los carcinomas, cánceres de origen epitelial y el grupo más abundante.

1.1. Neoplasias Mieloproliferativas

Las neoplasias mieloproliferativas (NMP) son alteraciones neoplásicas del sistema hematopoyético mieloide. Estas enfermedades se caracterizan por presentar un número de células madre hematopoyéticas anormalmente elevado, generando una proliferación anormal de las líneas celulares en la médula ósea. Esto se traduce, en función de la línea celular principalmente afectada, en un número elevado de glóbulos rojos, glóbulos blancos o plaquetas en sangre . Dentro de las NMP crónicas se distinguen:

- Leucemia Mieloide Crónica (LMC)

Se produce un exceso de granulocitos que invaden la médula ósea y el resto del organismo hasta que finalmente alteran el funcionamiento de los órganos. Representa el 15-20% de las leucemias diagnosticadas y su incidencia en España es de un 1.6-2 casos por 100,000 habitantes (AECC, 2017). El pronóstico de la enfermedad es favorable, especialmente cuando es detectada en sus fases tempranas. Su progresión es muy lenta y puede permanecer asintomática durante años. Los síntomas, si aparecen, son bastante inespecíficos y pueden ser: debilidad, fatiga, pérdida de peso, anemia, hemorragias, hematomas no justificados, dolor ósea, entre otros.
Hoy en día el tratamiento de un paciente con LMC se basa en la administración de un inhibidor de la tirosina quinasa (ITK) que ha hecho que en los últimos años aumente la supervivencia de los paciente (REDECAN, 2014).

- Policitemia Vera (PV)

En esta enfermedad se produce un número anormal de glóbulos rojos. La enfermedad es generalmente asintomática y sus síntomas, cuando aparecen, son similares a los de LMC. La incidencia en España de 0.5-1 por cada 100,000 habitantes y la edad media es superior a los 60 años. Su pronóstico es bueno, siempre y cuando los pacientes reciban el tratamiento adecuado (Carreras, 2017).

- Trombocitemia Esencial (TE)

Se caracteriza por una cifra muy elevada de plaquetas en sangre que puede derivar en la oclusión de vasos sanguíneos. Su incidencia en la población española es de 2 casos por cada 100,000 de habitantes, es más habitual en mujeres y en individuos de edades superiores a los 60 años. Como las otras NMP no suele presentar síntomas. Si estos existen, suelen ser problemas de circulación como enrojecimiento de las extremidades, gangrenas, síncopes o trombosis. En algunos casos la sintomatología se asemeja a la PV, lo que puede dificultar su correcto diagnóstico. El tratamiento varía en función del estado de la enfermedad, asintomática o sintomática, así como la edad del paciente. Esta enfermedad se suele considerar crónica y la supervivencia de los pacientes es la misma que la de un individuo sano, sin embargo, es posible que la enfermedad evolucione a mielofibrosis o leucemia aguda (Carreras, 2017).

- Mielofibrosis primaria (MFP)

Caracterizada por la presencia de tejido fibroso en la médula ósea. En España, su incidencia es de 0.5 casos por 100,000 habitantes, tendiendo una mayor prevalencia en individuos de avanzada edad. La sintomatología no suele ser habitual y si se presenta es similar a la de LMC. El pronóstico de esta enfermedad no es muy favorable, la supervivencia media es de 6 años. El único tratamiento curativo es el trasplante de progenitores hematopoyéticos alogénico, si no fuera posible, se aplica quimioterapia como tratamiento paliativo.

Durante los últimos años se han descrito algunas características genéticas compartidas entre las distintas NMP y otras específicas de un determinado subtipo. En la mayoría de los casos de LMC hay una traslocación del cromosoma 9 y 22, alteración conocida como cromosoma de Filadelfia (Ph+) donde se encuentra el oncogén *BCR/ABL*. En los tres tipos de NMP restantes esta translocación no está presente y es habitual detectar una mutación en el gen *JAK2 V617F,* en el cromosoma 9. Con esta mutación, las enzimas tirosina quinasas fomentan el crecimiento celular independientemente de los mecanismos de control habituales. Otras alteraciones genéticas más infrecuentes son las del gen receptor de la TPO (*c-MPL*) y la del gen del receptor del factor de crecimiento derivados de las plaquetas (*PDGFR*) (Carreras, 2017). A pesar de estos avances, la etiología de estas enfermedades no está caracterizada en su totalidad y dada su similitud clínica y, en ocasiones genética, es difícil diagnosticar correctamente el tipo de NMP. Estos problemas en el diagnóstico derivan muchas veces en un tratamiento inadecuado y, por tanto, un peor pronóstico en algunos pacientes.

1.2. Leucemia linfática crónica

La leucemia linfática crónica (LLC) se caracteriza por un aumento incontrolado de linfocitos maduros que son incapaces de desarrollar su función inmunitaria y tienen un tiempo de vida medio superior a un linfocito sano. Existen dos variantes de esta enfermedad en función del tipo de linfocito que prolifera. Las variantes de LLC son la tipo T, con mayor incidencia en países asiáticos, y la tipo B, con mayor incidencia en países occidentales y de mejor pronóstico (AECC, 2017).

En España, la incidencia de la enfermedad es de 1-2 casos por cada 100,000 habitantes y en la mayoría de los casos se diagnostica en personas mayores de 60 años. Los pacientes suelen permanecer asintomáticos durante largos periodos de tiempo. En las fases más avanzadas de la enfermedad, se detectan con mayor facilidad sus síntomas: cansancio, infecciones recurrentes así como aumento del tamaño de los ganglios linfáticos. La elevada concentración de linfocitos en sangre también puede derivar en un aumento en el tamaño del hígado o bazo derivando en la aparición de dolores abdominales (AECC, 2017).

El pronóstico del paciente dependerá no sólo del tipo de leucemia, estadio de la enfermedad o tiempo de duplicación de los linfocitos, sino también de otros marcadores serológicos. Por ejemplo, si no presentan mutado el gen *IGHV* de la región variable de la cadena pesada de la inmunoglobulina, mostrarán peor pronóstico y una rápida progresión de la enfermedad. Lo mismo ocurre con las personas que tienen una elevada concentración de las proteínas ZAP-70 o CD38. Otras de las anomalías caracterizadas son deleciones en los cromosomas 11 y 17, que también indican un mal pronóstico.

El tratamiento más habitual para combatir esta enfermedad es la quimioterapia combinada con inmunoterapia. Se utilizan agentes bioterapéuticos que atacan de manera dirigida a las células leucémicas sin dañar a las sanas. Aunque la enfermedad es incurable, los tratamientos permiten conseguir remisiones de la enfermedad durante largos periodos de tiempo antes de volver a necesitar tratamiento (Carreras, 2017).

1.3. Mieloma múltiple

El mieloma múltiple (MM) es una enfermedad hematológica caracterizada por una proliferación anormal de células plasmáticas malignas en la médula ósea. Las células plasmáticas pasan a producir inmunoglobulinas con una única combinación de cadenas pesadas y ligeras, en vez del amplio rango habitual (Fig. 2). En función de la cadena pesada que presenta la inmunoglobulina monoclonal, el mieloma múltiple se puede clasificar en: IgG (50-60%), IgA (30%), IgM (2%), IgD (0.5%), y excepcionalmente IgE. El porcentaje restante corresponde a casos de mieloma de Bence-Jones, donde se producen únicamente cadenas ligeras que se filtran a la orina (Landgren *et al.,* 2011).

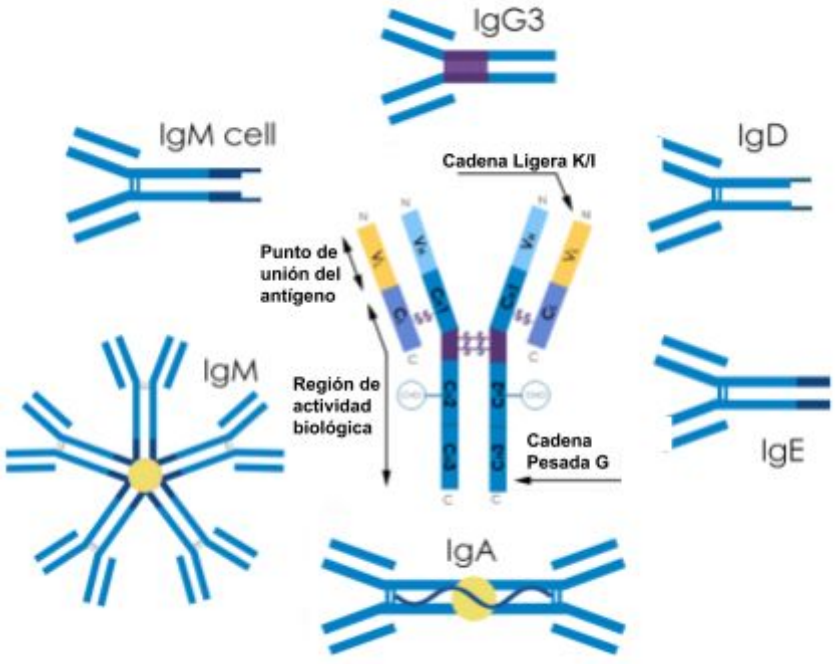

Figura 2. Estructura de un anticuerpo y sus diferentes variantes (Fuente: Abcan)

En el MM, la progresión de la enfermedad hace que se pueda dividir en dos grupos, el asintomático y el sintomático. En el estadio asintomático se establecen dos enfermedades MGUS y MM latente, que suelen preceder al MM sintomático. MGUS se caracteriza por concentraciones inferiores a 3 g./dl. de proteínas monoclonales en el suero y el porcentaje de células plasmáticas monoclonales en la médula ósea supera el 10% del total. Los valores de proteína monoclonal en suero en MM latente son ligeramente más elevados, siendo la concentración superior a 3g./dl. y el porcentaje de células plasmáticas monoclonales se mantiene al 10%. La enfermedad pasa a ser MM, y por tanto sintomática, cuando estos valores se elevan coincidiendo además con la aparición de síntomas característicos, determinados según el criterio "CRAB" (hipercalcemia, fallo renal, anemia y daños óseos). MGUS tiene una elevada incidencia en la población mayor de cincuenta años. La progresión de la enfermedad a su estadio sintomático es de tan sólo un 1% por año mientras que la progresión de MM latente a MM es de un 10% (Kyle *et al.,* 2011).

Las alteraciones genéticas que generan la enfermedad condicionan su tratamiento, siendo necesario el desarrollo de estudios genéticos para detectarlas. Algunas de las anomalías más habituales son translocaciones cromosómicas en los genes responsables de la síntesis de las cadenas pesadas, así como también deleciones en el cromosoma 13 y 14 (Corre *et al*., 2015). Entre los tratamientos disponibles está la quimioterapia en dosis altas, el trasplante de células madre o fármacos inmunomoduladores e inhibidores del proteosoma. Aún así, a día de hoy, el MM sigue considerándose una enfermedad incurable donde la evolución del paciente se caracteriza por etapas de remisión seguidas de recaídas y, finalmente una etapa refractaria donde dejan de ser sensibles a tratamientos (Fonseca *et al.,* 2009).

1.4. Cáncer de pulmón

El cáncer de pulmón es el cáncer responsable del mayor número de muertes a nivel mundial (SEOM, 2017). La mayoría de enfermos son diagnosticados en estados muy avanzados de la enfermedad, siendo ya difícil su recuperación con los tratamientos actuales y augurando tasas de supervivencia a los cinco años muy reducidas, de hasta 2-5% (Jantus-Lewintre *et al.*, 2012).
La dificultad de diagnóstico en los primeros estadios de la enfermedad, se debe a la ausencia de síntomas específicos. Los síntomas en etapas más avanzadas de la enfermedad son tos, expectoración sanguinolenta, disnea, dolor torácico, disfonía, disfagia, entre otros (AECC, 2017).

De acuerdo a sus características histopatológicas es posible distinguir dos grupos principales de cáncer de pulmón: carcinoma de células pequeñas o microcítico y carcinoma de células grandes o no microcítico.

El carcinoma de células pequeñas, o microcíticos, representa un 20% de los cánceres diagnosticados y está habitualmente asociado con el consumo de tabaco. Su nombre se debe a que las células tumorales que lo conforman son muy pequeñas y de forma ovalada, al observarse bajo el microscopio. Este tipo de cáncer se expande muy rápidamente, tanto en el pulmón como en otros tejidos, por lo que su tratamiento habitual suele basarse en tratamientos citotóxicos y no en la extirpación del tumor por cirugía.

El carcinoma de células no pequeñas o, no microcítico, según la forma de sus células al microscopio se divide en tres subgrupos:

- Carcinoma epidermoide (30%). Las células tumorales son delgadas y planas. Se localizan normalmente en la zona central de los pulmones y con frecuencia se necrosa. Tiene un crecimiento lento y es la forma más habitual de cáncer pulmonar en España.

- Adenocarcinoma (35-45%). Se suele localizar en las zonas periféricas de los pulmones y se extiende a la pleura y pared torácica. Las células tumorales producen sustancias mucosas que cubren los bronquiolos.

- Carcinoma de células grandes (10%). Es el menos frecuente de los carcinomas. Sus células son de gran tamaño y presentan un ratio citoplasma-núcleo elevado.

Como se ha indicado previamente, para tratar un carcinoma microcítico se tiende a recurrir a la quimioterapia, a pesar de que ninguno de los tratamientos actuales curan en su totalidad a los pacientes. En los carcinomas no microcíticos sí es posible el tratamiento a través de cirugía, siempre y cuando no haya diseminación, que mejora la recuperación total del paciente (Travis, 2002).

2 Tecnologías de alto rendimiento

El término *ómicas* deriva del sufijo griego -oma que indica conjunto o masa. Las tecnologías ómicas se refieren al estudio del conjunto de genes (genómica), metabolitos (metabolómica), transcritos (transcriptómico), o proteínas (proteómica), entre otros (Fig. 3).

Las ómicas permiten observar a los organismos como un todo, no solo centrarse en un área concreta. Este nuevo enfoque hace posible analizar los organismos de manera integral y desarrollar una hipótesis con los resultados obtenidos, y no desde un punto de vista reduccionista e impulsado por hipótesis a priori, como la ciencia ha venido haciendo hasta ahora (Horgan & Kenny, 2011).

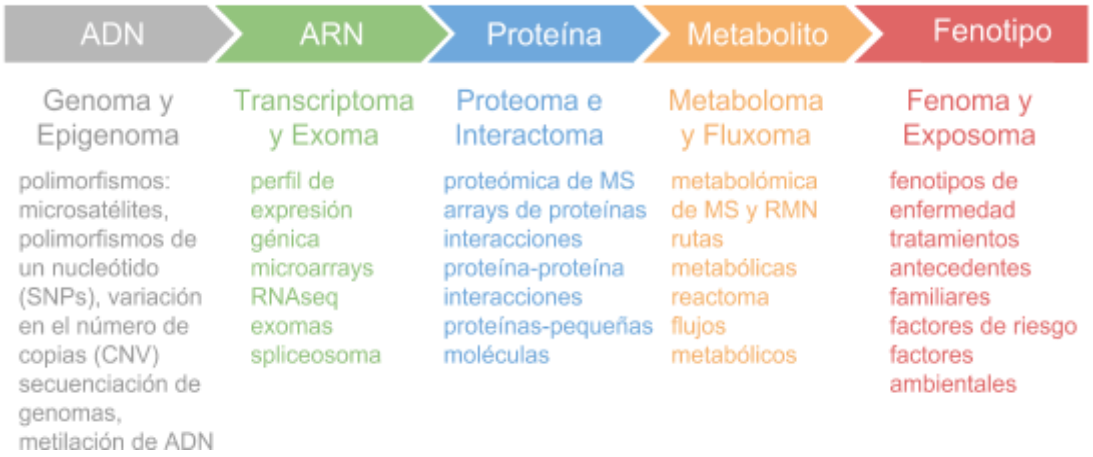

Figura 3. Tecnologías ómicas y su área de estudio (Modificada de Dumas, 2012).

El conocimiento adquirido con las tecnologías ómicas se puede aplicar en todos los ámbitos científicos. Como por ejemplo en el estudio de la simbiogénesis, en nutrición, bromatología, también en áreas clínicas como en el estudio de la patogénesis de enfermedades, la identificación de biomarcadores, el descubrimiento de fármacos o su efectividad.

Este trabajo se centra en dos técnicas ómicas concretas: la transcriptómica y la metabolómica. Ambas ómicas son reflejo del genotipo, estando además influenciadas por cambios en el fenotipo. Según el flujo de información del dogma central de la biología molecular, la transcriptómica se encuentra más alejada del fenotipo mientras que la metabolómica está directamente relacionada con este. En transcriptómica se detectan los transcritos responsables de los cambios a nivel biológico, en metabolómica lo que se estudia son las alteraciones en los niveles de los metabolitos generados por esos cambios a nivel de transcripción (Puchades *et al.*, 2015).

2.1. Transcriptómica y transcriptoma

El transcriptoma es el total de ARNm en una célula u organismo, refleja todo aquello que está siendo expresado en un momento dado. La transcriptómica es, por tanto, el estudio del conjunto de transcritos.

En el pasado, y en algunos estudios se continúa haciendo, se tendía a estudiar un único gen. Con la llegada de las ómicas, es posible medir la expresión de miles de genes a la vez. Se puede analizar la expresión diferencial entre individuos sanos y enfermos, para así identificar todos los procesos que, por ejemplo, llevan a una célula sana a convertirse en tumoral. También es aplicable en la identificación de nuevos biomarcadores para la predisposición a enfermedades o para la respuesta a ciertos fármacos. La transcriptómica es además muy utilizada en otras áreas de la ciencia, puede emplearse, por ejemplo, para estudiar los efectos causados por cambios en el hábitat o para inferir relaciones filogenéticas entre los organismos.

A lo largo del ciclo celular, el transcriptoma de una célula varía, desde que recibe su primer transcriptoma tras la división celular hasta el momento final de muerte celular. El transcriptoma cambia en respuesta a las condiciones y necesidades celulares de cada momento. Si el genoma humano tiene aproximadamente 29,000 genes, la cantidad de ARNms posibles es de seis veces ese número. Aún así, únicamente el 4% del contenido celular es ARN, del que tan sólo el 0.04% corresponde a ARN codificante (Brown, 2002).
Gracias a los estudios de transcriptómica se ha aprendido mucho sobre los genes y su traducción a proteínas. Ahora se sabe que, en la mayoría de los casos, los cambios generados a nivel celular no son debidos a un único gen. Realizando *knock-outs* se ha observado que al eliminar ciertos genes, otros suplen su función asegurando la supervivencia del organismo. Por consiguiente, el *fitness* de los organismos depende de cambios en la expresión de un conjunto de genes. Estas variaciones no tienen por qué ser elevadas pero, al darse al unísono, se generan grandes cambios a nivel celular. En cuanto a la concentración de ARNm y su traducción a proteínas se ha observado que no es una relación lineal. Una alta concentración de ARNm no implica una alta concentración de proteínas. De hecho, hay que tener en cuenta que no todos los ARNms detectados se convertirán en proteínas, así como tampoco todos tienen una misma tasa de traducción. Es decir, aunque aporte información, no se puede inferir directamente la concentración de proteínas a partir de los ARNm.

2.1.1. Plataformas analíticas

Existen diferentes plataformas que permiten evaluar simultáneamente la actividad de miles de genes. Las tecnologías más utilizadas son los microarrays y la secuenciación masiva. A continuación se explicarán ambos métodos, dando especial atención a los microarrays al ser la tecnología seleccionada para este trabajo.

Secuenciación masiva

La secuenciación masiva permite generar millones de fragmentos de ADN en un único proceso e identificar las bases nucleotídicas que los conforman. En los últimos años, el desarrollo de nuevas tecnologías de secuenciación y el descenso de sus precios ha hecho que esta metodología de trabajo deje de ser tan prohibitiva y su uso se extienda.

La secuenciación masiva abarca diversas tecnologías:

- Secuenciación exómica y genómica. Permite vincular cambios genéticos con determinadas enfermedades así como identificar el mecanismo de ciertas patologías.
- Secuenciación de ARN. Es posible estudiar diferentes tipos de ARN (microARNs, ARNts o ribosómicos). Con la detección de los niveles de expresión de estos ARN se pueden realizar comparaciones entre grupos e incluso identificar nuevos genes.
- Secuenciación epigenética. Se analizan los patrones de metilación para investigar los procesos de regulación.
- Secuenciación de inmunoprecipitación de cromatina (o ChIP-Seq). Se combina la inmunoprecipitación de la cromatina con la secuenciación masiva para identificar zonas de interacción entre la proteína y el ADN.

La principal ventaja de la secuenciación frente a los microarrays es que no se limita el área de estudio. Los microarrays deben ser diseñados con oligos previamente conocidos y elegidos de bases de datos existentes. Esto hace que haya un claro sesgo en su diseño, solo se observa aquello que se conoce. La secuenciación masiva permite estudiar el transcriptoma en su totalidad y no parte de él. A pesar de esta ventaja, los microarrays continúan siendo más populares por dos razones principales. La primera es que los investigadores están acostumbrados a su uso, conocen perfectamente el protocolo del procesado de la muestras y sus debilidades. La segunda razón es que, incluso tras la bajada de precios, los microarrays siguen siendo menos costosos y son más asequibles para la mayoría de los laboratorios.

Microarrays

Los chips o microarrays de ADN están formados por sondas de genes, secuencias o genomas específicos, sobre una superficie sólida (Fig. 4). El material es generalmente de cristal ya que es el que menos ruido de fondo produce frente a otras opciones (plástico o silicio, por ejemplo). El nivel de expresión de los genes se mide por análisis de imagen. Las moléculas de la muestra se marcan por fluorescencia, una vez hibridan con su sonda complementaria permanecen unidas al chip haciendo posible detectar su presencia en la muestra.

Existen diferentes tipos de microarray en función del estudio o biomoléculas con las que se trabaja: ADN, ARN, proteínas, compuestos químicos, tejidos, etc. Los chips de ADN más utilizados son:

- Chips de CGH (Hibridación Genómica Comparada). Se emplean para detectar la presencia de duplicaciones, amplificaciones o deleciones en el genoma.
- Chips de SNPs (*Single Nucleotide Polymorphism*). Con ellos se detectan mutaciones y polimorfismos usados como marcadores genéticos. Estos arrays se emplean para diagnóstico predictivo e inferir la respuesta a fármacos.
- Chips de embaldosado (*Tilling array)*. Se utilizan para localizar secuencias en el genoma.
- Chips de expresión. Usados para medir la expresión de diversos elementos biológicos como exones, microARNs o genes. Se emplean para identificar genes involucrados en procesos biológicos, caracterizar tumores, detectar patrones de

expresión relacionados con estados de patogenicidad, o identificar dianas terapéuticas.

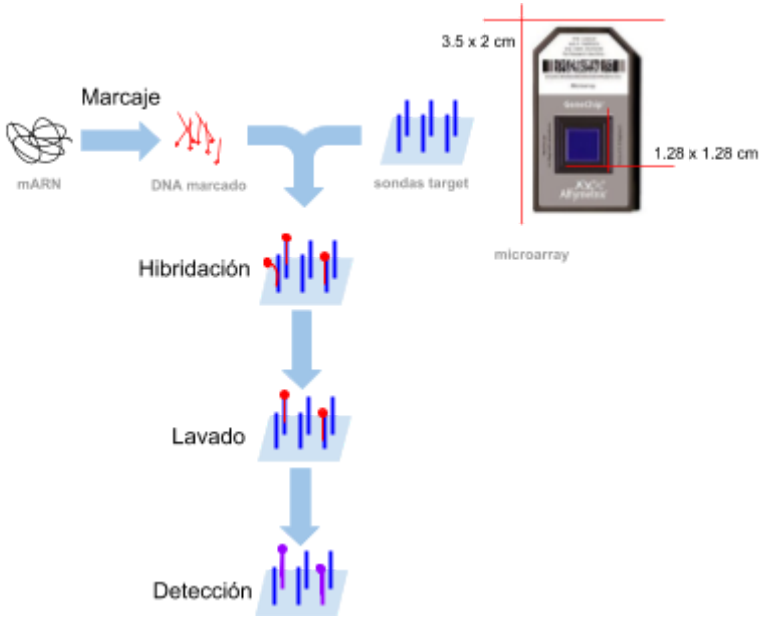

Figura 4. Funcionamiento de un microarray (Adaptada de Cañizares, 2010).

Como ya se ha indicado, las aplicaciones que se pueden dar a estos chips son múltiples. Para la detección del transcriptoma de una célula se utilizan microarrays de expresión génica. Su funcionamiento, y el de los otros chips, es sencillo y se basa en los mismos principios que el *Southern Blot* o el *Northern Blot*. En estos se hibridan las sondas a un ADN *target* unido a una membrana, mientras que en los microarrays es el ADN el que se hibrida a sondas unidas a la superficie del array. Cada gen está normalmente representado por más de una sonda, el conjunto de ellas se conoce como *probe set.*

2.1.2. Repositorios online

Existen numerosos repositorios online que dan acceso libre a estudios de transcriptómica realizados por todo el mundo. Algunas de las bases de datos que almacenan datos únicamente de cáncer son:

- *The Cancer Genome Atlas* (Cancer Genome Atlas Research Network, 2008). Almacenan información de la caracterización de 33 tipos de cáncer con datos de diferentes ómicas de 11,000 pacientes.
- *International Cancer Genome Consortium* (International Cancer Genome Consortium, 2010). Catalogan las anormalidades genómicas de 89 clases de tumores. Su acceso es restringido a usuarios registrados y es un punto de encuentro para aquellos científicos que trabajan en áreas similares para que puedan coordinarse y compartir información.
- *cBioPortal* (Cerami *et al.,* 2012). Esta web da acceso a datos procedentes de 162 estudios de cáncer . Se puede acceder a datos analizados sobre el número de copias de ADN, SNPs, la expresión de ARNms o microARNs, entre otros.

Además de iniciativas centradas en datos sobre cáncer, existen otras con recursos de transcriptomas de diferentes organismos. Por ejemplo, *Mammalian Gene Collection*

(Strausberg *et al.,* 1999), *Genotype-Tissue Expression Project* (The GTEx Consortium, 2013) y *Encyclopedia of ADN Elements* (*ENCODE*) que caracteriza las partes funcionales del genoma (NHGRI, 2015).

Todas los recursos web indicados anteriormente contienen datos procesados que provienen de bases de datos como *SRA* (*Sequence Read Archive* o SRA), *ArrayExpress* o *GEO (Gene Expression Omnibus)*. El Archivo de secuencias de genoma (Leinonen *et al.,* 2011) almacena datos no procesados y alineamientos obtenidos con plataformas de secuenciación de alto rendimiento. Mientras que la base de datos *ArrayExpress* (Parkinson *et al.,* 2005) almacena datos procedentes de estudios de microarrays bien anotados. *GEO* (Edgar *et al.*, 2002) almacena los mismos datos que las precedentes y muchos más, convirtiéndola en la mayor base de datos de expresión génica. En este trabajo se han utilizado datos de la base de datos, que se describe a continuación, con más detalle:

***Gene Expression Omnibus* (*GEO*)**

Gene Expression Omnibus (GEO) es un repositorio internacional de datos para su almacenamiento y libre distribución. Fue fundado en 2000 y desde entonces incluye datos de microarrays, *next-generation sequencing* y otros *datasets* obtenidos por tecnologías de alto rendimiento (Edgar *et al.,* 2002). Muchas ayudas para la financiación de proyectos y revistas requieren que los datos obtenidos y publicados sean almacenados en bases de datos de libre acceso. De esta manera, los datos pasan a estar disponibles a toda la comunidad investigadora para poner a prueba los resultados e incluso realizar nuevos descubrimientos a partir del mismo *dataset*. En *GEO* se almacenan no sólo datos sin procesar, sino también los datos procesados o los datos descriptivos de las muestras. Además, la plataforma web incluye herramientas online que permiten explorar, analizar y visualizar los estudios o perfiles de expresión de interés.

El 90% de los datos en *GEO* proceden de estudios de expresión génica. Estos datos se usan para investigar enfermedades, evolución, inmunidad, toxicología, metabolismo, etc. El 10% restante son datos de estudios de genómica funcional o epigenómica, que analizan la metilación del genoma, la estructura de la cromatina o la interacción entre el genoma y las proteínas (Barnett, 2013).

En la siguiente tabla (Tabla 1) se enumera el contenido actual de la base de datos. En total, hay más de 80,000 estudios correspondientes a más de 3,900 organismos.

Tabla 1. Contenido actualizado de la base de datos GEO NCBI.

	Públicos	Privados	Total
Series	86.856	10.718	97.574
Plataformas	17.492	297	17.789
Muestras	2,137,331	327,832	2,465,163
DataSets	4.348	-	4.348

Los datos se organizan en tres niveles principales de datos aportados por los usuarios:

- Plataforma (*GPLXXX*). Se refiere a la plataforma física donde se ha hecho el análisis. Se incluye una lista de los elementos del array (ADNc, oligonucleótidos, sondas…) o los elementos que se han detectado o cuantificado en el experimento (péptido, marcadores…). Una misma plataforma puede tener diferentes muestras de varios usuarios.
- Muestra (*GSMXXX*). Describe las condiciones bajo las cuales cada muestra se ha adquirido y las mediciones obtenidas. Cada muestra tiene una única plataforma y puede estar incluida en varias series.
- Series (*GSEXXX*). Una serie comprende todas las muestras que son de un mismo grupo. En la serie se incluyen descriptores de los datos así como conclusiones y los análisis realizados.

Además existe un cuarto nivel que son los *Datasets* (*GDSXXX*). Estos datos han sido seleccionados por el personal de *GEO* a partir de las muestras de los usuarios. Un *dataset* integra todas aquellas muestras que son biológicamente y estadísticamente comparables, que pueden pertenecer a diferentes series pero siempre a la misma plataforma de análisis.

Desde su fundación, *GEO* ha ido incluyendo mejoras graduales. Una de ellas es establecer requisitos más restrictivos al admitir información, así como actualizar los datos admitidos en función de las nuevas tecnologías. También desarrollaron el estándar MIAME (*Minimum Information About a Microarray Experiment*) para la descripción de los microarrays (Barnett, 2013).

2.2. Metabolómica y metaboloma

El metaboloma es el conjunto de perfiles de metabolitos en un sistema biológico bajo unas condiciones concretas. Los metabolitos participan en las redes metabólicas, formadas por enzimas y reguladas por factores genéticos y redes de señalización. en las que participan los metabolitos.. El metaboloma es el punto final de todas las rutas biológicas y el mejor indicador del fenotipo del individuo. Los estudios de metabolómica permiten estudiar el efecto de cambios producidos a nivel genético, transcriptómico y proteómico estudiando directamente el metaboloma (Dumas, 2012). Además de su cercanía con el fenotipo, otra de las ventajas de la metabolómica es que el número de objetos a estudio es inferior. Si se compara con otras ómicas, como la transcriptómica o proteómica, el número de metabolitos de un organismo es considerablemente más bajo que el de transcritos o proteínas (Gupta *et al.*, 2014).

Los estudios de metabolómica se centran en el estudio de los niveles de los metabolitos presentes en una muestra biológica. Los metabolitos son compuestos con un peso molecular inferior a 1,500 Da, como aminoácidos, ácidos orgánicos o lípidos. En la bibliografía, normalmente se hace referencia a unos 3,000 metabolitos. Sin embargo, con las metodologías actuales de RMN y EM se han detectado unos 300,000 metabolitos, de los cuales, sólo un 15% está caracterizado (Dumas, 2012).

A pesar de la escasa madurez del área, la metabolómica ya genera un gran impacto en muchas áreas de la ciencia. Se utiliza en toxicología, seguridad alimentaria o monitorización de tratamientos. De especial interés es su uso en la identificación de biomarcadores para la detección de enfermedades o respuesta a tratamientos. La importancia de esta técnica recae en que las muestras biológicas analizadas se pueden obtener de forma no invasiva. Los metabolitos se pueden obtener a partir de tejidos o biofluidos, como suero, orina o sangre, aportando información no sólo sobre las rutas metabólicas implicadas en el desarrollo de la enfermedad sino también en posibles aproximaciones para su diagnóstico y tratamiento. Así pues, la información extraída de los estudios metabolómicos puede ser utilizada para la identificación de nuevas dianas terapéuticas (Puchades-Carrasco, 2013).

2.2.2. Plataformas analíticas

Las técnicas analíticas más utilizadas para la caracterización de perfiles metabólicos son la resonancia magnética nuclear (RMN) y la espectroscopía de masas (EM). La selección de la plataforma analítica depende del tipo de muestra con la que se trabaja y los objetivos establecidos en el estudio. En la Tabla 2 se realiza una comparación de las ventajas e inconvenientes de ambas técnicas.

Tabla 2. Ventajas e inconvenientes de RMN y EM (Adaptada de Puchades-Carrasco, 2013; y Bonneau *et al.,* 2016).

	Espectrometría de masas	Resonancia magnética nuclear
Ventajas	Elevada sensibilidad (pg-ng) Permite mayor selectividad Posibilidad de hacer análisis dirigido Mantenimiento del equipamiento coste medio Detecta muchos metabolitos (aprox. 500)	No destructiva No selectiva Detecta todos los metabolitos que contienen en su estructura el núcleo analizado (^{1}H) Mínima preparación de la muestra Altamente reproducible Permite análisis directo de tejidos Asignación de espectros basada en bases de datos públicas Duración experimento (5 - 60 min) Resultados cuantitativos
Inconvenientes	Requiere preparación y separación previa de la muestra Baja reproducibilidad Necesidad de patrones internos para la identificación y cuantificación de compuestos Técnica destructiva Equipamiento costoso Duración experimento hasta 60 min	Baja resolución en muestras complejas con superposición de señales Baja sensibilidad (µg) Detecta pocos metabolitos (aprox. 200) Equipamiento y mantenimiento costoso

	con cromatografía Sesgo en la detección (pos/neg mode)	

Algo que se debe tener en cuenta es que ambas técnicas aportan información diferente sobre las moléculas presentes en una misma muestra. EM informa sobre la masa de los fragmentos analizados, y su tiempo de retención si se usa cromatografía. Mientras que con los datos de RMN es posible inferir la estructura de las moléculas o analizar moléculas que se ionizan con dificultad. Compuestos con la misma masa, como la isoleucina y la leucina, se asemejan mucho en EM pero en RMN son fácilmente distinguibles.

Puesto que en este trabajo se utilizan datos obtenidos por RMN, se hará una descripción más detallada de esta técnica.

Resonancia magnética nuclear

La RMN se basa en las propiedades magnéticas de los núcleos analizados. En presencia de un campo magnético externo constante los momentos magnéticos de estos núcleos se encuentran alineados hasta el momento en el que se aplica un pulso de radiofrecuencia, perturbando este equilibrio. El efecto de esta perturbación y la vuelta al equilibrio tras el pulso de radiofrecuencia es el resultado interpretado por la RMN.

En los seres vivos los elementos químicos más comunes son: carbono, hidrógeno, oxígeno y nitrógeno, fósforo y sulfuro. Con estos seis elementos se forman la mayoría de las moléculas de la tierra y organismos que habitan en ella. En la RMN aplicada a metabolómica se suelen estudiar núcleos de protón (^{1}H), por ser los núcleos activos en RMN con mayor abundancia natural. Los núcleos de fosfato (^{31}P) también son muy abundantes pero la dificultad de distinguir compuestos fosforilados de compuestos sin fosforilar, al solaparse, hace que el análisis de protones sea el método más extendido (Markley *et al.,* 2017).

La RMN de protón (^{1}H-RMN) se basa en la capacidad de este núcleo para ser magnéticamente activo (Fig. 5). En ausencia de un campo magnético, los estados de espín posibles se distribuyen de manera aleatoria y con la misma energía. Si se genera un campo magnético externo se produce una polarización parcial, los espines se redistribuyen a favor o en contra del campo magnético, ocupando los estados de menor energía. La diferencia de energía entre los distintos niveles determina la frecuencia de radiación necesaria para excitar el núcleo a un estado de espín de mayor energía. Esta frecuencia, que es característica para cada núcleo, se conoce como frecuencia de *Larmor.* Si se aplica a la muestra un pulso de esta radiofrecuencia perpendicular al campo magnético, todos los espines nucleares pasarán a su estado más alto de energía. Al cortar el pulso, los espines se relajan y vuelven al estado de equilibrio produciendo una pequeña cantidad de radiación detectable, llamado decaimiento de la inducción libre (*Free Induction Decay*, FID). Esta señal se convierte en un espectro de RMN usando la transformada de Fourier. Se pasa de tener un pulso de radiofrecuencia, a un espectro con

picos de diferentes formas y tamaños, distribuidos a lo largo del eje de frecuencias en función de su desplazamiento químico, que viene determinado por su entorno químico. Las unidades del desplazamiento químico se expresan por partes por millón (ppm) y equivalen a 10^6 veces la relación entre la separación de las señales y la frecuencia del campo magnético externo. El área de los picos es además proporcional al número de protones que contribuyen a esa señal, de esta forma las señales pueden emplearse con fines cuantitativos (Puchades-Carrasco, 2013).

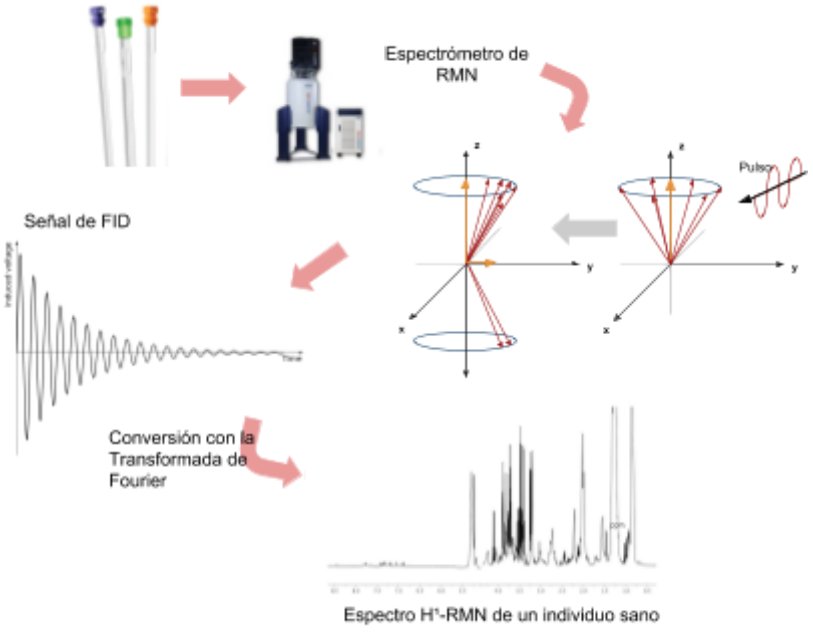

Figura 5. Diagrama explicativo de la obtención de un espectro H^1 por RMN.

En metabolómica se realizan diferentes experimentos de RMN para analizar las muestras, estos son (Beckonert *et al.*, 2007; Puchades-Carrasco, 2013):

- 1D NOESY (presaturación). Con esta secuencia de pulsos se satura la señal del agua suprimiendo su señal. Es útil para cuantificar compuestos de bajo y alto peso molecular en una solución acuosa.
- CPMG (editado basado en las propiedades de relajación). Las moléculas de gran tamaño, como las proteínas y lipoproteínas, son las primeras en relajarse tras un pulso. Una vez todas las moléculas son excitadas, se espera a que estas se relajen y, pasado su tiempo de relajación, se obtienen las señales de resonancia que pertenecerán únicamente a las moléculas de pequeño tamaño de la muestra. Es el método de detección utilizado en suero y plasma, dado su alto contenido proteico.
- Difusión (editado basado en las propiedades de difusión). Las moléculas de pequeño tamaño presentan coeficientes de difusión rápidos. Aprovechando esta propiedad es posible obtener señales de únicamente moléculas de elevado peso molecular como lípidos. Este método es muy utilizado en suero y plasma.
- J-resuelto (experimento homonuclear 2D). En este experimento se separa la señal en 2D, la información relacionada con el acoplamiento de la señal del

desplazamiento químico. Gracias a este método es posible interpretar mejor las señales con solapamiento.

2.2.2. Repositorios online

La metabolómica es una ómica que, si se compara con la transcriptómica, aún está en sus fases iniciales. No existen muchos repositorios online de *datasets* de metabolómica, principalmente porque la mayoría de las revistas no requieren depositar los datos para la publicación de los resultados. Aún así, hay algunas bases de datos que almacenan datos metabolómicos de especial interés como:

- *MetaboLights* (Haug *et al.,* 2013). Es una base de datos creada por la organización europea *EMBL-EBI* y el repositorio de metabolómica donde muchas revistas piden a los autores subir sus datos. Almacena un total de 266 *datasets* de análisis metabolómicos hechos con EM o RMN y procedentes de diferentes especies, tan sólo 86 de ellos son de humanos. La página incluye información referente a los metabolitos, como su función biológica, localización o resultados en diferentes experimentos, aunque la mayoría de los metabolitos están incompletos.
- *Metabolomics Workbench* (Sud *et al.,* 2016). Este repositorio estadounidense alberga 744 estudios de metabolómica (355 de ellos procedentes de humanos). La técnica de obtención de datos más utilizada es EM aunque hay además datos de RMN. Sirve como punto de coordinación de diferentes iniciativas en el área de metabolómica tanto a nivel internacional como de su país de origen.
- *MetabolomeXchange* (http://www.metabolomexchange.org/). El repositorio alberga 703 *datasets* de los cuales 255 son de estudios realizados en humanos. El objetivo de este repositorio, también creado por *EMBL*, es facilitar la búsqueda de datos metabolómicos para los investigadores aunando la información disponible en los principales repositorios en la actualidad (los dos previamente mencionados y *Metabolomic Repository Bordeaux* repositorio de plantas).

Además de estos repositorios que almacenan estudios completos de metabolómica existen bases de datos que albergan información sobre los metabolitos, sus espectros, participación en rutas metabólicas, etc. La más importante en el área es *The Human Metabolome Database (HMDB)* (Wishart *et al.*, 2007). La base de datos contiene tres tipos de información de los metabolitos del cuerpo humano: información química, clínica e información relativa a la biología molecular y bioquímica del metabolito. En la base de datos hay información de 74,462 metabolitos y en cada uno de ellos se pueden encontrar links que redirigen a bases de datos sobre su bioquímica y participación en rutas metabólicas, como *KEGG (*Kanehisa & Goto, 2000)*, UniProt* (The UniProt Consortium *et al.,* 2017)*, GenBank* (Benson *et al.,* 2013) o *Drugbank* (Law *et al.,* 2014).
Otra base de datos de interés en metabolómica, pero para datos de EM, es *Metlin* (Smith *et al.,* 2005). Gracias a esta base de datos es posible identificar los resultados obtenidos por EM y EM/EM, así como obtener links de acceso a otras bases de datos como *KEGG, PubMed* o *HMDB,* para cada uno de los 960,000 compuestos que contiene.

2.3. Estrategias de análisis de datos ómicos

La estrategia de análisis de datos ómicos varía en función del tipo de ómica que se esté utilizando, pero presentan todos ellas unas líneas comunes generales.

Diseño experimental

Este paso es fundamental en todos los estudios ómicos.
Es importante establecer un adecuado diseño experimental adaptado a la enfermedad de estudio, objetivos establecidos y tecnología ómica empleada.
Algunos de los aspectos que se deben tener en cuenta son:

- El número de sujetos a incluir en el estudio y la distribución de edades, sexo, procedencia étnica y otras posibles variables de interés.
- El tipo de muestra a analizar.
- Los procedimientos para la recogida de las muestras.
- El procesamiento de las mismas.
- La plataforma analítica.
- Los datos clínicos asociados a las muestras necesarios para la interpretación biológica de esos resultados.

Procesado de los datos

El procesamiento de los datos es específico del dato ómico con el que se esté trabajando. A continuación se detalla el procesado típico de los datos ómicos utilizados en este trabajo:

- Datos de metabolómica obtenidos con RMN

Los espectros de RMN obtenidos deben ser pretratados, eliminando regiones del espectro que no son informativas y además pueden introducir variabilidad (como el agua o la urea), y calibrados utilizando una señal de referencia como puede ser la señal del grupo metilo de la alanina o la señal del TSP (compuesto de referencia habitualmente utilizado en estudios de RMN).
A continuación se procede al *bucketing* del espectro, que consiste en dividirlo en pequeños cubos (entre 0.001 y 0.04 ppm, dependiendo del tipo de muestra) para integrar el área del espectro en su interior.
Se obtendrá una matriz, con tantas filas como muestras analizadas y tantas columnas como en *buckets* se divide el espectro.
En RMN la normalización se hace en función de cada *bucket* y de cada muestra, y es necesaria siempre que se esté trabajando con muestras que tienen diferentes factores de dilución (por ejemplo, orina) o espectros que presenten señales de muy distinta intensidad. Para la normalización en función de muestras se suele recurrir a la normalización del espectro por su área total, aunque otros métodos más específicos según el tipo de muestras son normalización por *PQN* o cuantiles. El escalado por columnas permite que todos los metabolitos, independientemente de cuál sea la intensidad de su señal en el espectro, tengan el mismo peso dentro de cada muestra. Se pueden utilizar diferentes escalados: Centrado, Centrado con Pareto o Centrado con Varianza Unitaria. Adicionalmente, puede ser necesaria otra normalización de los datos dirigida a eliminar otras fuentes de variabilidad que puedan interferir en el análisis,

debido, por ejemplo a problemas en la reproducibilidad al obtener los espectros en días diferentes o por diferentes personas. En estas ocasiones se suele recurrir a algoritmos específicos, como puede ser del paquete *sva* (Leek *et al.*, 2017), que permite eliminar la variabilidad entre los sets de muestras (Puchades-Carrasco *et al.,* 2016).

El análisis estadístico puede preceder o no a la anotación del espectro. En cualquier caso, la asignación de los picos del espectro o de *buckets* de relevancia estadística, se realiza con información disponible en bases de datos públicas como *Human Metabolome Database* (*HMDB*) (Wishart *et al.,* 2009), *Biological Magnetic Resonance Data Bank* (*BMRB*) (Ulrich *et al.,* 2008) o privadas como *BBIOREFCODE* 2.0. provista por la empresa Bruker, empresa líder en la fabricación de espectrómetros de RMN. Una vez anotado el espectro, se calcula el área de intensidad que le corresponde a cada metabolito a partir de los *buckets.*

- Datos de transcriptómica obtenidos con microarrays

Tras la hibridación, se obtiene una imagen digital que representa la intensidad de hibridaciones por celda. Con el software adecuado, esta imagen es convertida a un fichero CEL que contiene la intensidad de hibridación a nivel de sondas.
Una vez leído el fichero CEL y transformado a una matriz, se debe corregir el ruido de fondo y normalizar las muestras para que sean comparables entre sí.
La normalización pretende eliminar la variabilidad sistemática de los datos a causa de los procesos técnicos (concentración de ARN, tiempos de hibridación, soluciones utilizadas, aparato de detección) más que a cambios biológicos. Aunque cada *dataset* utiliza chips de la misma casa comercial cada uno de ellos arroja resultados diferentes, incluso cuando se usa exactamente la misma muestra. Para eliminar esta variabilidad aleatoria hay que realizar primero una normalización intra-array, que elimine la variabilidad de los mismos genes dentro de un chip, seguida de una normalización inter-array, que permitirá comparar los resultados de los diferentes arrays (Ayala, 2017).
Existen diferentes métodos para llevar a cabo todos los pasos de la normalización de modo secuencial o de manera combinada como pueden ser: MAS 5.0 de Affymetrix (Hubbell *et al.*, 2002), GC-RMA (Wu *et al.,* 2002), RMA (Irizarry *et al.,* 2003), entre otros.
En ocasiones puede ser necesario volver a procesar los datos tras la normalización. Por ejemplo, aplicar una transformación logarítmica para disminuir el rango de valores de niveles de expresión, eliminar aquellos genes que presenten valores nulos en la mayoría de las muestras o imputar valores ausentes asignándoles el valor de la media o mediana de los restantes. Estos métodos no son habituales y pueden inducir a error, ya que se pueden descartar o modificar genes que actúan en conjunto con otros y marginalmente no tienen una actividad apreciable (Ayala, 2017).
Por último, es habitual recurrir a una media o mediana representativa de las sondas que representan a un mismo gen reduciéndose la dimensión de la matriz que contiene los niveles de expresión (García-García, 2016).

Análisis exploratorio

Se realiza un análisis exploratorio de los datos, que variará en función del dato ómico. Se pueden realizar diagramas de cajas, análisis de *clusters* o PCAs que permiten obtener

una visión global de la distribución de los datos y detectar comportamientos anómalos en las muestras.

Análisis estadístico

La estrategia a seguir en el análisis estadístico de los datos depende del objetivo planteado al inicio de la investigación. Los datos pueden ser sometidos a diferentes abordajes estadísticos dependiendo de su origen ómico.

El análisis estadístico puede tener dos enfoques diferentes. El primero es un análisis dirigido en el que se pretende identificar y cuantificar el comportamiento de elementos concretos, asociados a una ruta o patología concreta. El segundo enfoque consiste en un estudio del perfil de los datos ómicos con el que se esté trabajando. Entre los métodos estadísticos aplicados se incluyen análisis no supervisados o de clusterizado, como los PCAs, y análisis supervisados o de predicción de clases, como OPLS-DA. Estos análisis deben ir acompañados de tests de validación y análisis univariante para dar legitimidad a los análisis multivariantes (Puchades-Carrasco, 2013; García-García, 2016).

Interpretación biológica

Finalmente, se procede a interpretar biológicamente los resultados obtenidos.
La interpretación biológica de los resultados de metabolómica suele depender de la información disponible en la bibliográfica o en bases de datos como *HMDB* (Wishart *et al.,* 2007), *ConsensusPathDB-human* (Kamburov *et al.*, 2009), y *KEGG* (Kanehisa & Goto, 2000).
En transcriptómica, es importante tener en cuenta que los genes y sus transcritos suelen trabajar de forma cooperativa. A la hora de analizar los resultados se deben considerar los genes agrupados aplicando métodos de enriquecimiento funcional (Subramanian et al., 2005; Al-Shahrour *et al.,* 2007) para poder detectar así qué grupos de genes están más diferencialmente expresados y no gen a gen.
Este método, es también aplicable en metabolómica, primero se seleccionan las variables de interés, por ejemplo aquellas con un comportamiento diferencial entre los dos grupos de análisis, para luego consultar en bases de datos, como *KEGG* (Kanehisa & Goto, 2000) o términos *GO* (Ashburner *et al.,* 2000), las posibles relaciones entre los elementos y determinar aquellas funciones sobrerrepresentadas, tanto sobreexpresados como reprimidas. Otros procedimientos para el análisis de grupos son aquellos que aplican regresión logística (Sartor et al., 2009; Montaner & Dopazo, 2010) que estudian la dependencia entre una variable binaria y otra continua permitiendo además incluir otras variables convirtiendo el análisis en multidimensional.

3 Metaanálisis de datos ómicos

3.1. Fundamentos del metaanálisis

El metaanálisis es una metodología estadística que permite combinar los resultados de diferentes estudios con una misma hipótesis, nula o alternativa, y diseño experimental. La unión de sus resultados permite obtener una medida combinada del efecto de interés con una mayor precisión que la ofrecida por los estudios de manera individual. Tanto los

estudios como los métodos estadísticos utilizados deben ser seleccionados cuidadosamente ya que su aplicación incorrecta puede llevar a resultados erróneos (García-García, 2016; Catalá-López & Tobías, 2014).

Los metaanálisis fueron aplicados inicialmente en estudios médicos sobre incidencia de enfermedades, mortalidad, etc. A escala ómica se comenzaron a utilizar recientemente para identificar los genes altamente expresados en múltiples microarrays.

La aplicación de metaanálisis en el análisis de los datos supone la obtención de resultados más robustos y persistentes, ya que se pueden filtrar y eliminar aquellas relaciones espurias que pueden surgir en los *datasets*. Esto es especialmente importante, ya que con el descenso de costes en las tecnologías ómicas, el número de estudios disponibles ha aumentado pero la mayoría de las veces los estudios realizados incluyen un número reducido de muestras derivando en un escaso poder de detección.

En la actualidad, la mayoría de los métodos de metaanálisis trabajan a nivel de gen, en estudios de expresión génica, o a nivel de variante, buscando la relación de los polimorfismos con las enfermedades. Sin embargo, el uso de procedimientos de metaanálisis a nivel de función proporciona una nueva vía para la obtención e interpretación de resultados al facilitar su comprensión biológica y clínica.
La combinación permite mejorar la detección de aquellas clases de genes verdaderamente enriquecidas, al disminuir el ruido causado por la tecnología, la estructura del experimento o el bajo tamaño muestral (García-García, 2016). Pero además trabajar con rutas en lugar de hacerlo con genes mejora la consistencia de los resultados. Se sabe que los genes de estudios independientes pueden presentar pocos solapamientos entre sí, mientras que el estudio de rutas biológicas es más consistente. Al trabajar a nivel de rutas no es necesario que los genes de interés sean comunes en todos los estudios, lo que se pretende es detectar aquellas rutas comunes que pueden estar activadas por diferentes genes en cada estudio (Shen *et al.,* 2010).

3.2. Métodos existentes

Existen pocos métodos orientados al metaanálisis de datos a nivel de función y cada uno de ellos con un enfoque diferente:

- Shen & Tseng (2010). Propusieron los métodos MAPE (*Meta-Analysis for Pathway Enrichment*) (Fig. 6). Los métodos están orientados a la detección de las rutas metabólicas basándose en el enriquecimiento de grupos de genes (Subramanian *et al.,* 2005). Los métodos son MAPE_G, MAPE_P y MAPE_I. El primer método es un metaanálisis de enriquecimiento de rutas a nivel de gen. Se asume que todos los genes a estudiar están presentes en todos los estudios, se realiza metaanálisis asignando a cada gen un valor basado en su expresión en todos ellos y se hace un análisis de enriquecimiento funcional. El segundo método analiza las rutas enriquecidas y no asume que los genes presentes son comunes en todos los estudios. El análisis de enriquecimiento funcional se hace de forma individualizada en cada estudio. Se hace metaanálisis de los resultados individuales obteniéndose valores de representación para cada una de las rutas. El hecho de no unificar los datos antes del análisis hace que

MAPE_P sea estadísticamente más efectivo que MAPE_G. Sin embargo MAPE_G pasa a ser más efectivo cuando todos los genes son comunes en los estudios. Por este motivo, la opción óptima es la integración de ambos métodos con MAPE_I. El último método integra los resultados de los dos metaanálisis de enriquecimiento obteniéndose resultados más robustos y consistentes.

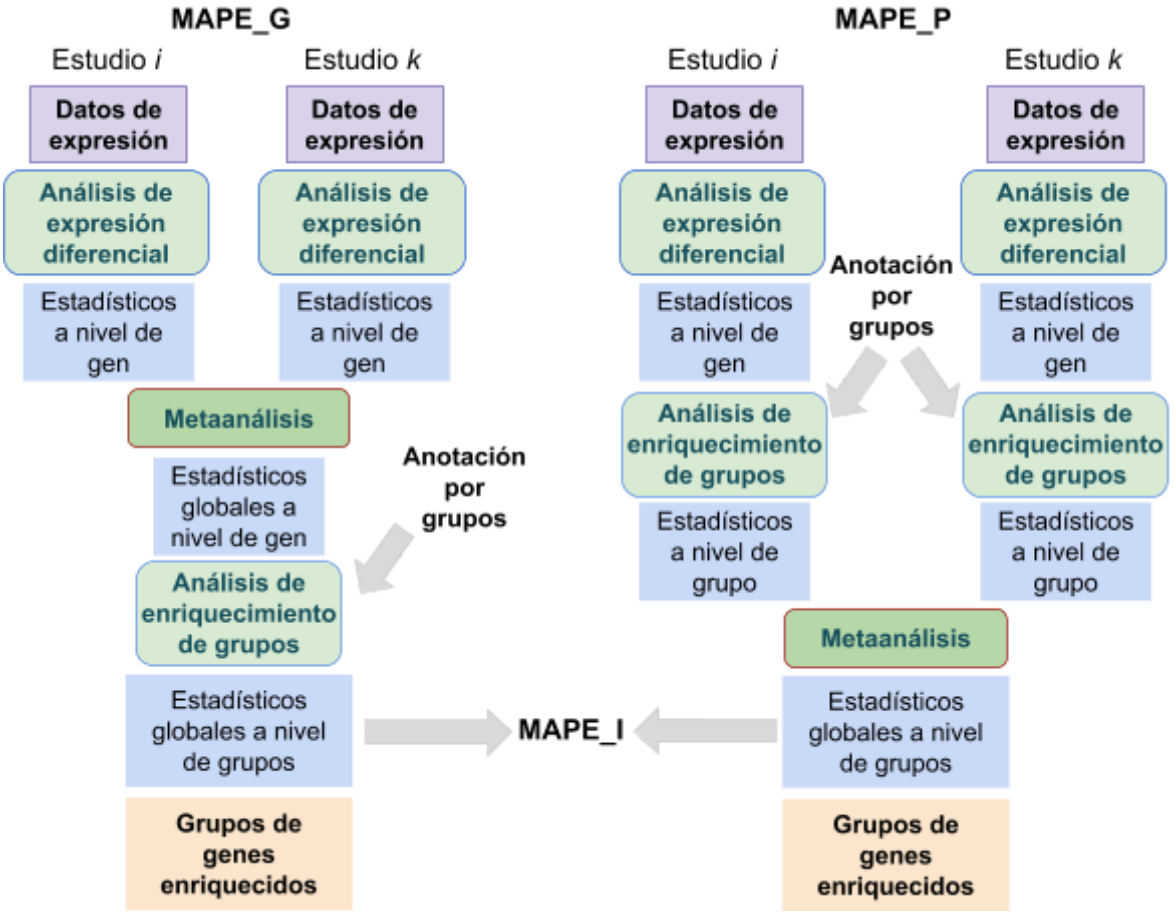

Figura 6. Cuadro representativo del flujo de trabajo del método MAPE propuesto por Shen *et al.* (*2010*) (Diagrama adaptado de Chen *et al.,* 2013).

- Chen *et al.* (2013). Proponen un método con enfoque bayesiano de los datos de expresión (Fig. 7). Con el método se pretende, de forma flexible, inferir la probabilidad de identificar los genes diferencialmente expresados teniendo en cuenta la variabilidad proceden de los estudios individuales. Esta puede ser el hecho de que no todos los estudios tienen el mismo número de genes, número muestras o que la calidad de los estudios es heterogénea. Aplicando este método se consiguen modelar de forma conjunta todos los datos, obteniendo los grupos de genes diferencialmente expresados.

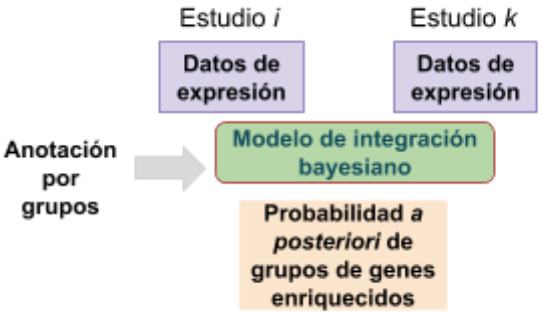

Figura 7. Cuadro representativo del flujo de trabajo del método propuesto por Chen *et al., 2013* (Diagrama adaptado de Chen *et al.* 2013).

Estos dos métodos están orientados a datos procedentes de estudios transcriptómicos. El método aplicado en este trabajo, de García-García (2016), es utilizable con datos de diferentes ómicas desde transcriptómica, con datos de expresión génica o microARNs, a datos de proteómica o metabolómica. En esta metodología se necesita únicamente información funcional sobre los datos, de bases de datos o propia de los investigadores, para hacer un metaanálisis funcional de los estudios mejorando la interpretación de los resultados con respecto a los métodos propuestos hasta la fecha (Fig. 8).

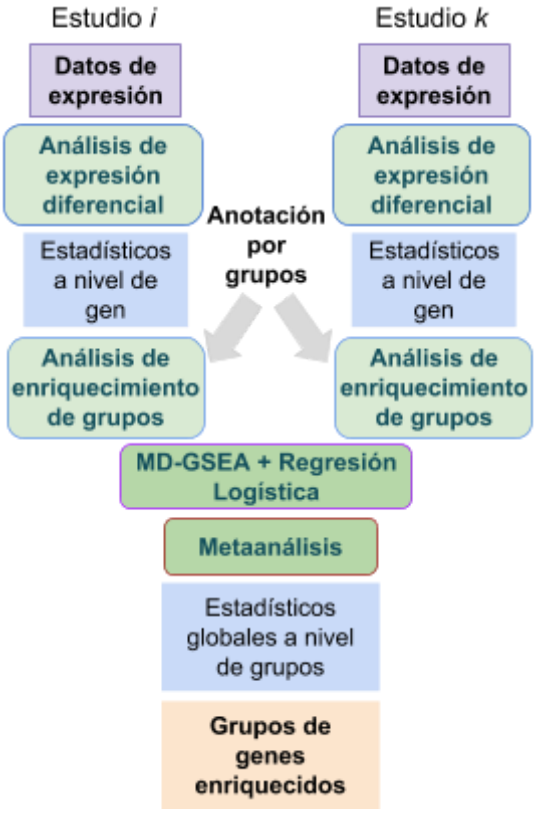

Figura 8. Cuadro representativo del flujo de trabajo del método utilizado en este trabajo.

En este método se realiza un análisis de expresión diferencial de los datos y a continuación un *Gene Set Analysis* basado en modelos de regresión logística (Sartor *et al.*, 2009; Montaner & Dopazo, 2010) que permite obtener bloques funcionales en cualquiera de las condiciones. Sobre los resultados del enriquecimiento funcional, se lleva

a cabo el metaanálisis funcional que asigna estimadores numéricos para caracterizar el poder estadístico de los resultados y evaluar la consistencia de los resultados.

4 Integración de datos ómicos

4.1. Fundamentos de la integración de datos ómicos

En los últimos diez años, los artículos en cuyo título se incluía el término *"data integration"* pasaron de 1,504 en 2006 a 3,292 en 2016 (https://www.ncbi.nlm.nih.gov/pubmed/). El aumento generalizado en el uso de técnicas ómicas ha hecho que cada vez haya más información accesible para los investigadores, aunque no necesariamente creada por ellos. Esta creciente disponibilidad para toda la comunidad científica permite reutilizar los datos para probar la veracidad del análisis original (reproducibilidad) pero también para buscar nuevas formas de explotarlos y extraer nueva información a partir de ellos.

La integración de datos ómicos no es más que el siguiente paso lógico en un abordaje global del análisis de datos. Es algo que a día de hoy se hace habitualmente y es imprescindible en la investigación, como es unir información de dos fuentes. Se podría, por tanto, definir la integración como el uso de información de diferente origen para obtener un mejor entendimiento de un sistema/situación/asociación, uniendo la información (Gómez-Cabrero *et al.,* 2014). El objetivo final es obtener una nueva perspectiva de los resultados, que permitirá no sólo predecir mejor las respuestas sino también abordar los datos desde un punto de vista de biología de sistemas.

La evaluación del conjunto total de los procesos biológicos proporciona un ensamblaje del sistema biológico real que conforma. Integrando información de diferentes ómicas se pueden compensar las carencias que algunas presentan así como desechar datos poco fiables. Además, el hecho que los resultados de diferentes ómicas converjan en un resultado común, un gen o ruta metabólica de interés, hace que los resultados sean más robustos y disminuya la probabilidad de falsos positivos (Ritchie *et al.,* 2015).

4.2. Métodos existentes

Aún no existe un clara metodología para llevar a cabo la integración de datos ómicos. El análisis variará en función de las tecnologías ómicas que se pretendan integrar. Por ejemplo, los datos transcriptómicos y metabolómicos al ser integrados no tienen una asociación directa como sí pueden tener los datos transcriptómicos y proteómicos. La mayoría de los transcritos serán traducidos a una proteína determinada pero un transcrito concreto puede inducir a la alteración de los niveles de más de un metabolito diferente.

De la misma forma que ocurre en los estudios de una sola ómica, al realizar la integración de varias de ellas es importante diseñar bien el experimento (Fig. 9). Se pueden considerar hasta cuatro diseños experimentales posibles para la obtención de las muestras ómicas (Cavill *et al.,* 2016).

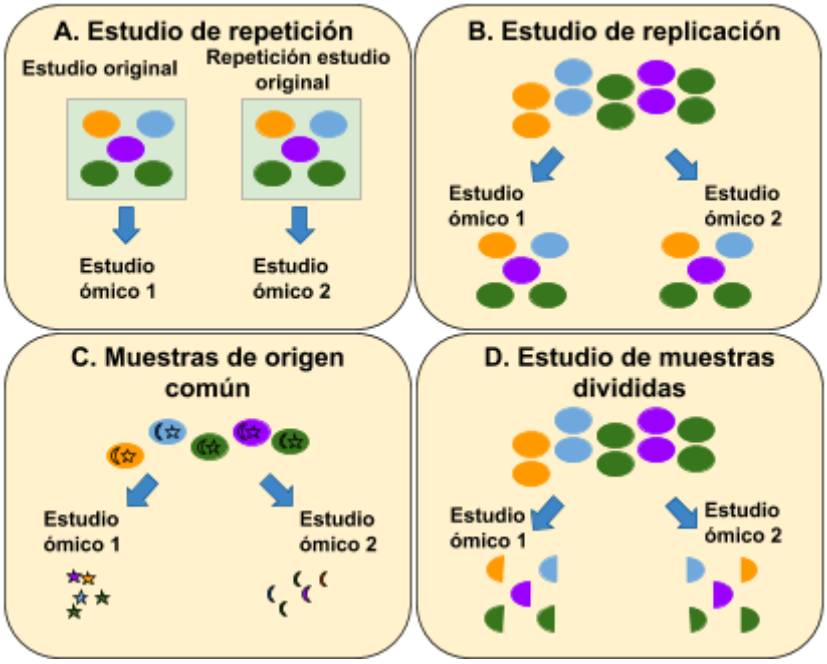

Figura 9. Diseños experimentales comunes en los estudios multiómicos (Diagrama adaptado de Cavill *et al.*, 2016).

Estos son:

- Estudio de repetición (Fig. 9A). En este caso se aplican los mismos protocolos para la obtención de las muestras pero son tomadas y analizadas en momentos distintos. Al trabajar con muestras diferentes, los datos deberán ser ajustados considerando un posible *'batch effect'.*
- Estudio de replicación (Fig. 9B). Se utilizan réplicas biológicas obtenidas en el mismo experimento pero son preparadas de forma aislada para ser analizadas por cada técnica ómica.
- Muestras de origen común (Fig. 9C). No son muestras biológicas idénticas, sino que proceden de diferentes tejidos de un mismo individuo o con características similares. Por ejemplo, pueden ser muestras de sangre y muestras histológicas de un paciente enfermo.
- Estudio de división de muestras (Fig. 9D). Se utiliza una misma muestras biológica que es dividida en dos para ser analizadas por cada técnica ómica.

Desde un punto de vista global la integración se puede llevar a cabo por tres vías diferentes: conceptual, estadística o basada en modelos (Ebbels & Cavill, 2009) (Fig. 10). La primera es un enfoque conceptual en el que cada ómica es analizada de forma individual y las conclusiones extraídas se unifican en una resolución final. La integración estadística, la más habitual cuando se refiere a integración como tal, busca relaciones estadísticas entre los diferentes elementos de ambas ómicas. La integración basada en modelos es aquella en la que, una vez se conoce a la perfección el sistema biológico, es posible usar modelos computacionales y matemáticos que infieran los cambios que han ocurrido partiendo de los datos ómicos obtenidos. Esta última opción es algo que, con el conocimiento actual, es imposible de lograr (Cavill *et al.,* 2016).

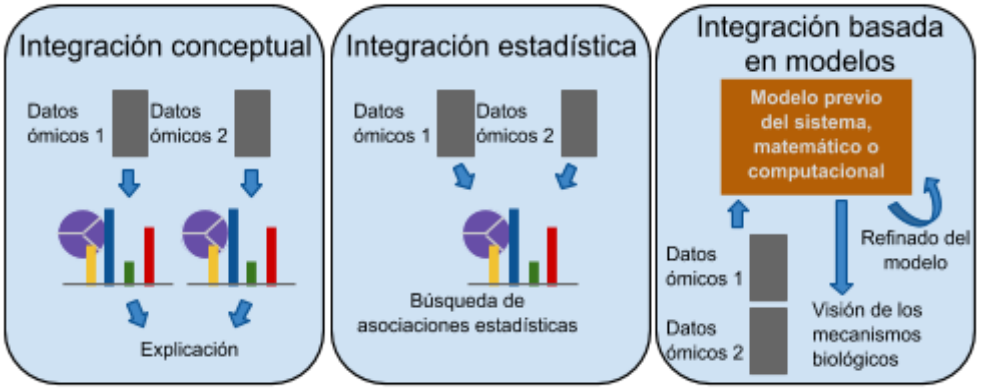

Fig. 10. Diseños experimentales comunes en los estudios multiómicos (Diagrama traducido de Cavill *et al.*, 2016).

Existe un plétora de métodos estadísticos para integrar los datos que se pueden dividir en integración supervisada, no supervisada y semi-supervisada. En la integración supervisada las muestras se separan en función de su fenotipo, por ejemplo sanos y enfermos, y se utilizan algoritmos basados en *machine learning* para evaluar los modelos e integrarlos. La integración no supervisada, en cambio, no tiene en cuenta el origen de los datos a la hora de integrarlos. La integración semi-supervisada es un punto intermedio entre las dos anteriores, utiliza algoritmos de aprendizaje para integrarlos considerando los datos tanto asociados a un fenotipo como sin asociar.

Dentro de cada vertiente hay a su vez una gran variedad de algoritmos con diferentes criterios. Por ejemplo, dentro de los métodos no supervisados se incluyen métodos basados en matrices de factorización, métodos que correlacionan las variables a los sets de datos, métodos bayesianos que asumen las posibles distribuciones, métodos de análisis de *kernel* u otros basados en redes y grafos. En un artículo reciente Huang *et al.* (2017) hace un repaso de todos estos métodos y sus aplicaciones.

En la Figura 11 se muestra un esquema de algunos de los métodos que han sido descritos en la bibliografía para la integración de datos de origen metabolómico y transcriptómico (Cavill *et al.,*2016).

Integración por correlación

Este tipo de concatenación busca correlacionar la información entre los dos *datasets* (Fig. 11A). Existen muchos métodos estadísticos que correlacionan *datasets*, como la correlación de Pearson, Spearman, modelos lineales o correlaciones parciales. El principal problema de este procedimiento es, según los resultados obtenidos hasta el momento en la bibliografía, que las rutas más relacionadas entre sí normalmente no se muestran correlacionadas mientras que las correlaciones más alejadas entre sí son las que más se identifican.

Integración por concatenación

La integración por concatenación consiste en unificar en una misma matriz los datos procedentes de las dos ómicas para ser analizados por métodos estadísticos (Fig. 11B). Este método, aunque conceptualmente sencillo, es problemático cuando el volumen y la

dimensión de las variables son muy distintos. En estos casos es necesario escalar correctamente las variables para conseguir que el peso final sea homogéneo y proporcionado, sin importar el origen ómico de la variable. Sin embargo, las variables tenderán a agruparse con aquellas de su mismo origen ómico. Existen métodos como es *iCluster* (Shen *et al.,* 2009) que permite solucionar en parte estos problemas al clusterizar los datos concatenados y sin concatenar.

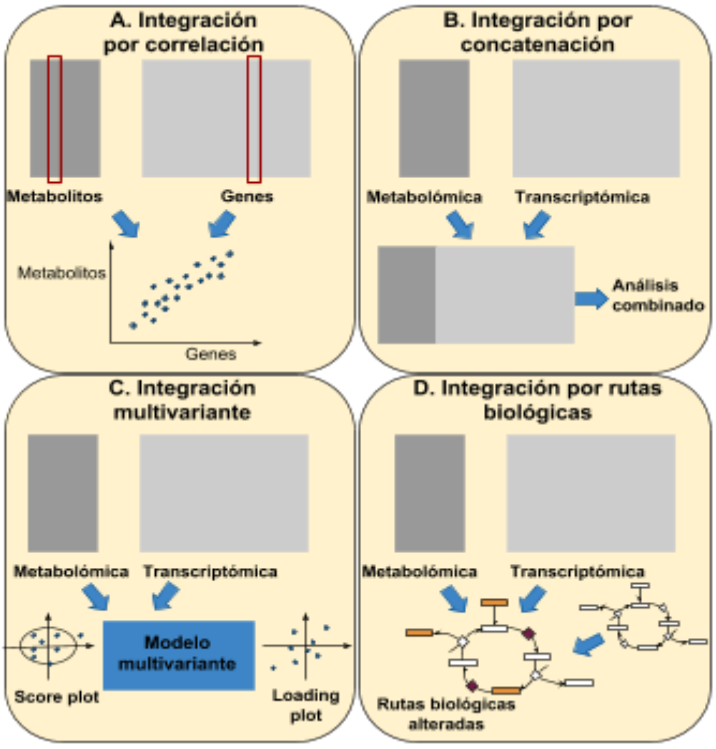

Figura 11. Métodos de integración estadísticos habituales de datos metabolómicos y transcriptómicos (Diagrama adaptado de Cavill *et al.*, 2016).

Integración multivariante

Esta técnica de integración es útil cuando se trabaja con datos metabolómicos y transcriptómicos con un alto nivel de colinealidad (Fig. 11C). La integración por multivariantes mantiene separados los datos de diferente origen y aplicando algoritmos, como O2PLS, se pueden modelar las relaciones entre los dos grupos. Con el uso de este método, *a priori* mucho más efectivo que los dos anteriores, se ha observado que tan sólo un tercio de los datos que se tienen son modelables. Es decir, gran parte de la información con la que se trabaja normalmente es única de cada *dataset*.

Otros métodos de integración multivariante son aquellos procesos multi-vía que permiten añadir una nueva capa de información a los datos ómicos o los modelos, modelos jerárquicos utilizando PCA o PLS-DA y por último modelos de correlación canónica que busca combinaciones lineales entre las variables de los dos sets de datos.

Integración por rutas biológicas

El uso de bases de datos de rutas biológicas se aplica en esta metodología. Se consigue examinar directamente la relación entre los genes expresados y la respuesta en

metabolitos generados (Fig. 11D). Para emplear este método existen recursos web y software específicos que permiten integrar la información de forma directa. Una de ellas es *IMPaLa* (Kamburov *et al.*, 2011) que realiza un análisis de enriquecimiento de los datos transcriptómicos y metabolómicos de forma aislada que después combina identificando si las rutas a las que pertenecen están diferencialmente expresadas. Otras herramientas como *PathVisio* (Kutmon *et al.,* 2015*)*, *Paintomics* (García-Alcalde *et al.,* 2011) o *INMEX* (Xia *et al.,* 2013) obtienen de forma conjunta los elementos diferencialmente expresados para detectar las rutas de interés.
Al igual que ocurría cuando se concatenaban los datos, al integrar por rutas metabólicas es importante ser conscientes de que es probable que los datos resulten desbalanceados si hay un exceso de genes o metabolitos. También se debe considerar la presencia de un sesgo en la obtención de resultados. Con RMN se más probable identificar aminoácidos que otras biomoléculas, por lo tanto la rutas sobrerrepresentadas detectadas suelen estar vinculadas a la síntesis de estos aminoácidos. Al identificar los metabolitos es posible que algunos de ellos no se identifiquen correctamente al estar superpuestos con otros picos de mayor intensidad, o bien que no sea posible asignarle correctamente su quiralidad química, lo que supone asociarlo a rutas en las que no tienen porqué participar además de una pérdida de información. Por último, otro posible sesgo que hay en la asignación de los metabolitos de interés, es que en RMN es habitual asignar únicamente aquellos metabolitos que se encuentran alterados entre las dos condiciones de estudio, perdiéndose información de muchos metabolitos.

En este trabajo, no se parte de los mismos estudios para transcriptómica y metabolómica, sino que se dispone de un grupo de estudios diferentes en cada ómica, aunque correspondientes a la misma enfermedad. Por este motivo, no es posible realizar muchos de los métodos antes aplicados. La integración de los datos se deberá hacer desde un punto de vista conceptual, basada en los resultados del análisis estadístico de las rutas biológicas implicadas.

II. Objetivos

La finalidad del presente trabajo es por un lado, mediante el metaanálisis de los datos incluidos en cada estudio, identificar cuales son las rutas que se encuentran alteradas de manera general en los distintos tipos de cáncer y, en segundo lugar, mediante la integración de datos transcriptómicos y metabolómicos, caracterizar los procesos moleculares y metabólicos alterados en esta enfermedad.

Los tres objetivos que se plantearon para el trabajo fueron:

1. La realización de una revisión sistemática de estudios de transcriptómica de las enfermedades de interés. Para después, seleccionar aquellos estudios que fueran compatibles con los datos de metabolómica de los que se disponía.
2. La aplicación de una metodología de metaanálisis funcional (García-García, 2016) que ofrezca una doble caracterización funcional: a nivel transcriptómico y a nivel metabolómico.
3. La integración de los resultados obtenidos en los metaanálisis de ambas ómicas y la realización de una interpretación funcional en el contexto de la enfermedad.

III. Material y métodos

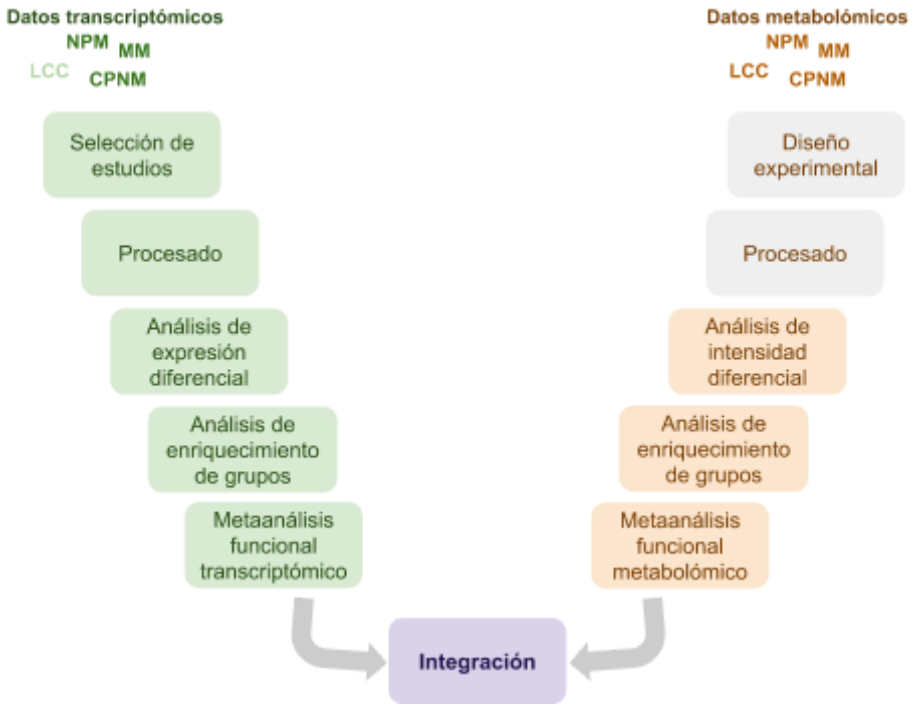

Figura 12. Protocolo de análisis en un estudio de integración ómica.

La Figura 12 resume la estrategia de análisis realizada en este trabajo. Los dos grupos de datos ómicos siguen un proceso paralelo y muy similar. La única diferencia se encuentra en el procesado inicial de los datos y el enriquecimiento de grupos, donde usan listas de anotación diferentes, adaptadas al origen ómico de la información.

En la Tabla 3 se indica el tiempo estimado necesario para el desarrollo de las diferentes tareas que conforman el trabajo.

Tabla 3. Tiempo estimado para el desarrollo de cada uno de los objetivos.

Cronograma de trabajo		
	Tarea	**Tiempo**
	Revisión bibliográfica	2 semanas
Objetivo 1	Revisión sistemática y selección de estudios	2 semanas
	Procesamiento datos transcriptómica	2 semanas
	Análisis expresión diferencial transcriptómica	3 semanas
	Análisis intensidad diferencial metabolómica	1 semana
Objetivo 2	Análisis de enriquecimiento de grupos	1 semana
	Metaanálisis	1 semana

Objetivo 3	Integración e interpretación de resultados	1 semana
	total	13 semanas

1 Revisión sistemática y selección de estudios

La revisión y selección de los estudios incluidos en los metaanálisis y su integración, es fundamental para la obtención de resultados de interés que puedan ser interpretados. Seleccionar estudios con poca calidad puede llevar a resultados con sesgo así como a una pérdida de eficiencia. La definición de criterios de inclusión y exclusión de los estudios consensuados, incrementa la robustez de los resultados y facilita su interpretación.

1.1. Metabolómica

Los datos de metabolómica utilizados en este trabajo proceden de estudios realizados en la Unidad Mixta CIPF-IISLAFE de Metabolómica. La unidad se especializa en la identificación de biomarcadores oncológicos de utilidad clínica a través de RMN, así como también en la caracterización de las alteraciones metabólicas asociadas a la progresión de la enfermedad.

1.1.1. Neoplasias Mieloproliferativas

Los datos proceden de un estudio cuyo objetivo es evaluar la capacidad de la RMN para distinguir entre los diferentes casos de neoplasias mieloproliferativas ante la similitud que presentan sus cuadros clínicos (Tabla 4). Las muestras son de suero obtenido de pacientes con enfermedades enmarcadas bajo NMP y de otros pacientes que sufrían enfermedades hematológicas no NMP. Los controles eran individuos sanos sin ninguna patología. Las muestras se clasificaron también en función de otra variable de interés como es la presencia o ausencia de mutaciones en el gen *JAK2 V617F*.

Tabla 4. Muestras incluidas en el estudio metabolómico de neoplasia mieloproliferativa.

Grupos	*n*	*Descripción*
Controles (C)	21	Individuos sanos
Policitemia vera (PV)	22	Pacientes con cáncer
Trombocitemia esencial (TE)	46	Pacientes con cáncer
Trombocitemia secundaria (TS)	12	Pacientes con enfermedad hematológica
total	101	

1.1.2. Leucemia linfática crónica

Los datos proceden de un proyecto (MacIntyre *et al.,* 2010) que tenía como objetivo determinar si era posible distinguir los perfiles metabolómicos de los pacientes de LLC

que presentaban marcadores de diagnósticos conocidos como son las mutaciones en el gen *IGHV* o la concentración de la proteína ZAP-70 (Tabla 5). Las muestras de suero procedían de pacientes de LLC sin tratar y fueron analizadas por ^{1}H-RMN.

Tabla 5. Muestras incluidas en el estudio metabolómico de leucemia linfática crónica.

Grupos	*n*	*Descripción*
Controles (C)	7	Individuos sanos
Sin mutación (mut -)	27	Pacientes con cáncer
Mutación (mut +)	8	Pacientes con cáncer
total	42	

1.1.3. Mieloma múltiple

Los datos se obtuvieron por ^{1}H-RMN y provienen de dos ensayos clínicos en pacientes con MM (Puchades-Carrasco *et al.*, 2011). En el estudio, se recogieron muestras de suero sanguíneo de pacientes de MM antes del tratamiento y tras entrar en remisión, es decir, ausencia de detección de las proteínas monoclonales y una concentración inferior al 5% de células plasmáticas en la médula ósea. Los controles eran individuos sanos de la misma edad y sexo que los pacientes (Tabla 6).
Las variantes de mieloma múltiple que presentaban los pacientes también se conocían. Un total de 13 pacientes presentaba proliferación de la inmunoglobulina IgG, 9 de IgA y 5 presentaba mieloma de Bence-Jones.

Tabla 6. Muestras incluidas en el estudio metabolómico de mieloma múltiple.

Grupos	*n*	*Descripción*
Controles (C)	31	Individuos sanos
Diagnosis (D)	27	Pacientes con cáncer
Remisión (R)	23	Pacientes con cáncer tratados con fármacos
total	81	

1.1.4. Cáncer de pulmón

Los datos proceden de dos grupos de muestras de suero analizadas por ^{1}H-RMN (Puchades-Carrasco *et al.,* 2016). En el primero se obtuvieron muestras de individuos sanos y pacientes con cáncer. El segundo grupo de datos recogía se introdujeron además muestras de personas sin cáncer con enfermedades pulmonares benignas. Los datos procedentes de pacientes con CP se clasificaron en dos grupos según el estadio de la enfermedad: fase inicial (estadios I-IIIB) y fase avanzada (estadios IIIB o IV) (Tabla 7).

El estudio en el que se usaron las muestras originalmente tenía como objetivo desarrollar una técnica no invasiva para la detección de cáncer pulmonar. Para ello se compararon

los diferentes perfiles metabólicos de los pacientes con cáncer de pulmón frente a individuos con otras enfermedades pulmonares e individuos sanos.

Tabla 7. Muestras incluidas en el estudio metabolómico de cáncer de pulmón.

Grupos	*n*	*Descripción*
Controles (C)	84	Individuos sanos
Fase inicial (FI)	89	Pacientes con cáncer
Fase avanzada (AS)	86	Pacientes con cáncer
Enfermedad benigna (EB)	27	Pacientes con enfermedades pulmonares
total	286	

Además de los diferentes estadios de la enfermedad también existe información sobre la histología del tumor (adenocarcinoma, epidermoide o células grandes) y del posible tabaquismo del individuo (no fumadores, fumadores o exfumadores). Sin embargo esta información no está disponible para todas las muestras presentes. De un total de 286 muestras 219 están clasificadas según su histología y tan solo 182 indican si el individuo es o no fumador.

1.2. Transcriptómica

Los estudios de microarrays se seleccionaron del repositorio GEO. Los estudios seleccionados debían ceñirse a los distintos tipos de cáncer de los que ya se tenían datos de metabolómica. Por ello, los criterios de inclusión iniciales fueron:

- Estudios realizados en humanos.
- Individuos de ascendencia europea.
- Diseño experimental preferiblemente C*aso vs. Control* y con las mismas comparaciones analizadas en metabolómica.
- Preferiblemente un modelo de array común en todos ellos.

Con esta selección se pretende limitar al máximo los posibles problemas que puedan acarrear el uso de diferentes poblaciones en el origen de nuestros datos y minimizar la heterogeneidad de las muestras.La búsqueda inicial en el repositorio proporcionó 28 estudios que cumplían alguna de las características de interés pero no todas. Tras una revisión más exhaustiva de todos ellos, se descartaron los que, por ejemplo, utilizaban una plataforma analítica muy diferente del resto, no realizaban todas las comparaciones de interés, o bien el número de muestras era relativamente bajo.

1.2.1. Neoplasias Mieloproliferativas

Se plantearon hasta ocho estudios con los que trabajar, pero se descartaron en base a que tenían un número bajo de muestras o no realizaban la comparación de interés. Finalmente se seleccionó la serie GSE54644 que presentaba un número elevado de muestras y además indicaba la presencia o ausencia de mutaciones en el gen *JAK2 V617F,* variable también presente en las muestras de metabolómica (Tabla 8).

Los datos de esta serie se utilizaron en el estudio original para la caracterización de la ruta de JAK-STAT (Rampal *et al.,* 2014). Se pretendía evaluar el efecto causado por las mutaciones en el gen *JAK2 V617F,* en la activación y desarrollo del NMP.
El array utilizado fue *GeneChip HT-HG_U133A Early Access Array* de Affymetrix. Como su nombre indica es un array previo al desarrollo de la versión definitiva del array *HG U133A.*

Tabla 8. Muestras incluidas en la serie GSE54644 del estudio de neoplasia mieloproliferativa.

Grupos	*n*	*Descripción*
Controles (C)	11	Individuos sanos
Policitemia Vera (PV)	28	Pacientes con cáncer
Trombocitemia esencial (TE)	47	Pacientes con cáncer
Mielofibrosis primaria (MF)	18	Pacientes con pre-cáncer
total	104	

1.2.2. Leucemia linfática crónica

Se seleccionaron un total de 13 estudios de LLC como posibles *set* de estudio. Sin embargo se optó por descartar el uso de muestras de transcriptómica para este tipo de cáncer. Si bien algunos *datasets* presentaban las comparaciones de interés, el número de muestras era demasiado bajo, o bien no incluían controles. Por ejemplo, la serie GSE22529 hacía la comparación de interés pero tan sólo incluía 52 muestras. Otra serie que se consideró para el análisis fue GSE28654 ya que contaba con variables de interés, como eran las mutaciones en *IGHV* o los valores de ZAP-70, pero carecía de controles sanos.

1.2.3. Mieloma múltiple

La serie GSE6477 se seleccionó como *dataset* de transcriptómica de MM. Se barajaron hasta 8 posibles *datasets* pero el GSE6477 era el que mejor cumplía las condiciones a pesar de no contar con todas las posibles comparaciones de interés (Tabla 9). El objetivo original de estos datos era estudiar el MM causado por hiperploidías cromosómicas (Chng *et al.,* 2007). Por este motivo, las muestras además de contar con distinciones entre MGUS, MM latente o recaídas, distingue entre diploidías, hiperdiplodías o no diploidía, así como también la de deleciones en el cromosoma 13, que suele estar relacionado con un peor diagnóstico de MM. Las muestras de ARN proceden de células plasmáticas de la médula ósea de los individuos. En el estudio se analizaron un total de 162 muestras con el array *Human Genome U133A* de Affymetrix.

Tabla 9. Muestras incluidas en la serie GSE6477 del estudio de mieloma múltiple.

Grupos	*n*	*Descripción*
Controles (C)	15	Individuos sanos
Diagnosis (D)	75	Pacientes con cáncer
Recaída (R)	28	Pacientes con cáncer

Latente (L)	23	Pacientes con pre-cáncer
MGUS	21	Pacientes con pre-cáncer
total	162	

1.2.4 Cáncer de pulmón

Se plantearon varios posibles estudios a utilizar, ya que ninguno de los disponibles en GEO cumplía en su totalidad con los requisitos esperados. Había estudios con pocas muestras, no incluían diferenciación histológica de los tipos de CP, no se indicaba si los individuos eran fumadores, o la plataforma de análisis no era común al resto de estudios. Finalmente se optó por utilizar el estudio GSE43458. Los datos fueron empleados originalmente en un trabajo que estudiaba la diferencia de expresión del oncogen *ETS2* en cáncer pulmonar. El microarray utilizado para el análisis fue *Human Gene 1.0 ST Array* de Affymetrix, distinto al de los anteriores pero dentro de la misma casa comercial. El estudio fue seleccionado porque era el que más muestras analizaba, y aunque solo estudiaba adenocarcinomas, también hacía distinción de muestras dependiendo de si los individuos eran fumadores o no fumadores (Tabla 10). Todas las muestras de ARN se obtuvieron de tejido pulmonar, los controles proceden de tejido sano de pacientes no fumadores de los que se obtuvo tejido canceroso para el estudio.

Tabla 10. Muestras incluidas en la serie GSE43458 del estudio de cáncer de pulmón.

Grupos	*n*	*Descripción*
Controles (C)	30	Individuos sanos
Fumadores (F)	40	Pacientes con cáncer
No fumadores (NF)	40	Pacientes con cáncer
total	110	

2 Análisis primario de los datos

2.1. Transcriptómica

En este trabajo se han utilizado datos procedentes de chips de oligonucleótidos. Estos se diferencian de los microarrays de ADNc en la forma que miden la expresión génica. Los microarrays de oligonucleótidos obtienen el valor absoluto de la expresión, mientras que los de ADNc estiman la diferencia de expresión entre los diferentes genes (Yauk *et al.,* 2004). Los arrays de la compañía Affymetrix, al contrario que otras empresas como Agilent o Illumina, incluyen grupos de sondas (*probe sets*) que cubren toda la extensión de los transcritos de interés y permiten calcular con más precisión la expresión de los genes. Existen más de 10 grupos de sondas diferentes por secuencia de ARNm y además, están distribuidas sobre toda la superficie del microchip para eliminar la posible variabilidad intra-array. Otro punto a tener en cuenta en el diseño del chip es el ruido de fondo generado por las hibridaciones inespecíficas. Para calcularlo en el chip hay dos tipos de sondas: sondas con hibridación perfecta con el transcrito de interés (PM, *perfect*

match) y sondas con hibridación imperfecta (MM, *mismatch*). Estas sondas imperfectas tienen una única base nucleotídica cambiada y se usan para calcular las posibles uniones inespecíficas por otros transcritos que serán eliminadas durante la normalización de los resultados (Ayala, 2017). Con todos estos mecanismos implementados en los arrays es posible obtener resultados precisos, certeros y reproducibles (Affymetrix, 2017).

Los datos utilizados en el presente trabajo proceden de dos modelos de arrays de la línea *GeneChip*™ de Affymetrix, *Human Genome U133 Plus 2.0 Array* y *Human Exon 1.0 ST Array.*

A pesar de pertenecer a la misma línea comercial, el array *Exon 1.0 ST* es muy diferente al resto de arrays convencionales de la línea, como podría ser *U133 Plus 2.0.* La primera diferencia es que, tal y como su nombre indica, sus sondas hibridan con los exones de los transcritos. Esto permite analizar a dos niveles, se consigue la expresión de los genes y además se conoce la isoforma del transcrito. Otro cambio aparece en el diseño de las sondas *targets*. Los *targets* del array *U133 Plus 2.0* se obtienen con *primers* que reconocen la cola poli(A) de los ARNms de las muestras de interés. Mientras que las sondas *target* para el array *Exon 1.0 ST* se obtienen por *random priming,* resultando en sondas homogéneamente distribuidas y sin sesgo hacia sondas con colas poli(A) (cabe recordar que algunos ARNms carecen de ellas).

El procesado de las muestras también es diferente. En los arrays *U133 Plus 2.0* se utiliza directamente el ARNm purificado para obtener las sondas antisentido de ADNc *target* que se unirán a las sondas del array. En la preparación de las sondas *target* para *Exon 1.0 ST*, el ARNm debe ser previamente transcrito a ADNc para luego ser convertido al ARN antisentido con el que se obtendrán las sondas *target* de ADNc (Zimmermann & Leser, 2010).

Tabla 11. Tabla comparativa de ambos arrays.

	HG U133 Plus 2.0	Human Exon 1.0 ST
Sondas por gen	~ 11-20	~ 40
Sondas por array	~ 1,300,000	~ 5,500,000
Probesets por array	54	1,400,000
***Background* probes por array**	650	40

En la Tabla 11 se comparan los números de sondas y genes detectados por ambos arrays. El número de sondas unidas al array *Exon 1.0 ST* es hasta cuatro veces superior que las sondas de *U133 Plus 2.0*. Esto se debe a que, al hibridar con exones y no con el transcrito, se necesitan muchas más sondas para cubrir el número de exones de estudio. El elevado número de sondas a utilizar, obliga a limitar el número de sondas que detectan el ruido de fondo. Por este motivo, no hay el mismo número de sondas con hibridación imperfecta que sondas con hibridación perfecta como ocurre en el array *U133 Plus 2.0*. En el *Exon 1.0 ST* array se usan tan sólo 40,000 sondas distribuidas por el array para calcular el posible ruido de fondo (Zimmermann & Leser, 2010).

2.1.1. Procesamiento de los datos

Tras la lectura de los datos, se utilizó el mismo procedimiento de normalización para todos los datasets: *Robust Multi-array Average* (RMA), inclusive para la versión inicial de los datos del array *HG_HT-U133A.*
Este procesamiento común para todas las muestras reduce la variabilidad entre estudios (biológica y experimental propia de cada grupo de datos), mejorando su evaluación conjunta.

La única diferencia en el proceso fue la lectura inicial de los ficheros CEL. Los datos de CP, al proceder de un array de oligonucleótidos, era recomendable que fueran leídos con la función *read.celfiles* del paquete *oligos* (Carvalho & Irizarry, 2010) mientras que el resto de datos fueron leídos con la función *ReadAffy* del paquete *affy* (Gautier *et al.,* 2004).
La función seleccionada para la normalización fue la función *rma* del paquete *affy*. Esta función es una de las más utilizadas para la normalización de microarrays de la línea GeneChip. Ya que no requiere los datos de las sondas de MM para normalizar, permitiendo su uso en tanto arrays de oligonucleótidos como de ADNc.

La función *Robust Multi-array Average* (*rma*) lleva a cabo cuatro pasos:

- Se realiza una corrección de fondo de forma que ninguna sonda tenga valores negativos
- Normalización por cuantiles consiguiendo que la distribución de todos los arrays sea la misma al redistribuir los valores en función de la media en cada posición.
- Cálculo del logaritmo en base 2 ($\log_2$) de los niveles de expresión.
- Se hace un resumen de las distintas sondas de cada *probe set* en todos los microarrays con el método de *median polish* de Tukey (Mosteller & Tukey, 1977). Consiste en la normalización por medianas de los chips y genes hasta que converjan.

Los datos fenotípicos de los *datasets* se adicionaron al *ExpressionSet* de los datos normalizados descargando directamente el fichero Series Matrix de GEO. Los arrays fueron anotados utilizando la función *select* de la librería *AnnotationDbi* (Pagès *et al.,* 2017) con su correspondiente paquete, salvo para *HG_HT-U133A Early Access*, para el que se usó la versión definitiva del mismo. Con esta función es posible asignar a cada sonda su *Gene Symbol* correspondiente, así como *Ensembl Gene ID id* o *Entrez Gene ID.* Puesto que un gen puede estar representado por varias sondas diferentes esta información ha de unificarse bajo un único identificador. Esto se realizó con la función *avereps* de *limma* (Ritchie *et al.,* 2015) estandarizando todos los valores con la media total facilitando los futuros análisis.

Concluido el procesado y normalización, se hizo un análisis exploratorio de los estudios para la identificación de posibles *outliers.* Se analizó la distribución de las muestras con diagramas de cajas, *clusters* jerárquicos y PCAs.

2.1.2. Análisis de expresión diferencial

Los contrastes estadísticos que se realizaron fueron siempre comparando *Casos* frente a *Controles*, es decir, enfermos frente a individuos sanos. Se realizaron varias comparaciones con los datos disponibles, seleccionando finalmente 8 contrastes (Tabla 12). A la hora de realizar un metaanálisis es muy importante mantener equilibrio entre los contrastes y muestras que se incluyen. Si se añaden más contrastes en un tipo de enfermedad que en otro, puede suponer una sobrerrepresentación de ese grupo y sus funciones, infiriendo de manera negativa en la calidad de los resultados.

Tabla 12. Contrastes estadísticos de transcriptómica seleccionados para el metaanálisis.

Análisis de expresión diferencial transcriptómica	
Estudio	**Comparación**
CP	*Adeno No fumador vs. No fumador*
CP	*Adeno Fumador* vs. *No fumador*
MM	*Diagnosis* vs. *Control*
MM	*Recaída vs. Control*
MM	*Diag + Re* vs. *MGUS + Latente*
NMP	*PV* vs. *Control*
NMP	*TE* vs. *Control*
NMP	*JAK-* vs. *JAK+*

En las muestras de CP se compararon de pacientes con CP frente a individuos sanos en función si los pacientes enfermos eran fumadores o no.
Los pacientes de MM agrupados bajo Recaída y Diagnosis son los únicos que realmente padecen MM, por este motivo se enfrentaron a los Controles, MGUS y Latentes. Siendo estas dos últimos enfermedades fases previas al desarrollo completo de la enfermedad MM.
Para las comparaciones de los datos de NMP, los contrastes fueron de dos de las variantes de la enfermedad frente a los controles, así como también en función de la presencia o ausencia de mutaciones en el gen *JAK2 V617F*.

El nivel de expresión entre los grupos se analizó utilizando la función *lmfit* del paquete *limma* (Ritchie, 2015). Con esta función ajusta cada uno de los genes a un modelo lineal que permite ver la respuesta de cada gen frente a las condiciones de interés, no sin antes ajustarlo con la función *eBayes* que corrige los errores estándar producidos. A continuación, se aplicó a los valores de P la corrección de Benjamini & Hochberg (1995) con un FDR de 0.05, evitando falsos positivos y falsos negativos consecuencia del elevado número de comparaciones realizadas.

2.1.3. Análisis de enriquecimiento de grupos de genes

Tras el análisis de expresión diferencial se obtiene una lista de todos los genes ordenados según su patrón de expresión. En la parte superior de la lista se encuentran

aquellos genes diferencialmente expresados en los individuos enfermos, en la inferior aquellos en los que la expresión es mayor en los individuos sanos.
El procedimiento de enriquecimiento de grupos de genes requiere información sobre las posibles agrupaciones que los genes pueden conformar. Los genes pueden ser agrupados en función de sus términos *GO* (Ashburner *et al.,* 2000) o su participación en las rutas metabólicas, información disponible en *KEGG* (Kanehisa & Goto, 2000) . La anotación de los términos *GO* se obtuvo directamente desde *BioMart* (Aken *et al.,* 2016) dado que la información de *EnrichmentBrowser* (Geistlinger *et al.*, 2016) no estaba tan actualizada. En cambio, la información disponible sobre la agrupación de los genes en función de su participación en rutas metabólicas se descargó con *EnrichmentBrowser* al estar más actualizadas que otras alternativas como Reactome.db (Ligtenberg, 2017) o *KEGGrest* (Tenenbaum, 2017).
El modelo de enriquecimiento funcional aplicado está basado en modelos de regresión logística que permiten agrupar los genes basándose no sólo en su valor de P sino también por su agrupación funcional (García-García, 2016; Montaner *et al.,* 2009; Sartor *et al.,* 2009; Montaner & Dopazo 2010). Las funciones utilizadas se encuentran en el paquete *mdgsa* (Montaner, 2009). Primero, con *pval2index* se transforman los valores de P y estadísticos que asocian la direccionalidad del contraste a un ranking de los genes diferencialmente expresados Luego, con la función *indexTransform* se transforma el ranking en una variable apta para ser sujeta a un modelo de regresión logística. Con *annotFilter* se contrasta el listado de genes con los genes agrupados según los términos *GO* o rutas *KEGG.* Con esta función se eliminaron aquellos grupos que tenían más de 500 o menos de 5 genes asociados. Por último, la función *uvGSA* (*Uni-Variate Gene Set Analysis*) aplica el modelo de regresión logística que comprueba si alguno de esos grupos de genes están sobrerrepresentando a alguna de las condiciones experimentales de interés.

Tras hacer el análisis de enriquecimiento de grupos se obtienen los datos de partida para la realización del metaanálisis funcional. La función *uvGsa* del análisis de enriquecimiento proporciona como *output* una matriz (Tabla 13) que contiene los grupos de rutas KEGG o términos GO con mayor presencia en alguna de los grupos experimentales comparados.

Tabla 13. Resultados obtenidos del análisis de enriquecimiento de grupos del contraste de Recaída vs. Control de los datos transcriptómicos de MM.

Ruta *KEGG*	Descripción	N	*LOR*	*pval*	*padj*	*SD*	*t*	*conv*
hsa00010	Glicólisis / Gluconeogénesis	57	0.2691	4.361E-02	7.000E-01	0.1334	2.0180	1
hsa00020	Ciclo del ácido cítrico	29	0.7468	3.571E-05	1.820E-03	0.1806	4.1351	1
hsa00030	Vía pentosa fosfato	24	0.3063	1.318E-01	1	0.2032	1.5072	1
hsa00040	Interconversiones de pentosa y glucuronato	21	0.4030	6.449E-02	9.665E-01	0.2180	1.8489	1

hsa00051	Metabolismo de fructosa y manosa	31	-0.0939	6.017E-01	1	0.1799	-0.5220	1
hsa00052	Metabolismo de la galactosa	25	0.0123	9.510E-01	1	0.2002	0.0615	1
hsa00053	Metabolismo de ascorbato y aldarato	16	0.3009	2.283E-01	1	0.2497	1.2049	1
hsa00061	Biosíntesis de ácidos grasos	13	0.1054	7.041E-01	1	0.2775	0.3798	1

Las columnas de la matriz contienen estadísticos que permiten valorar la importancia de cada uno de los bloques formados. Estos valores son:

- *N*, número de genes o metabolitos anotados en cada grupo.
- *LOR* (*log Odds Ratio*), probabilidad estadística estimada para cada grupo.
- *pval*, valor de P asociado a cada probabilidad estadística.
- *padj*, valor ajustado de P.
- *SD*, desviación estándar de cada *log Odds Ratio.*
- *t*, estadístico t asociado a cada contraste.

Los estadísticos de esta tabla serán valorados de forma conjunta en el metaanálisis, combinando el grupo de funciones y genes sobrerrepresentados de manera común en todos los estudios.

2.2. Metabolómica

2.2.1. Procesamiento de los datos

Tras la adquisición de los espectros ^{1}H-RMN CPMG correspondientes a todas las muestras incluidas en los estudios, la estrategia de análisis de los datos fue similar al protocolo habitualmente empleado en este tipo de estudios, explicado en la introducción de este trabajo. Una vez anotados los espectros, se procedió a la cuantificación de todos los metabolitos a partir de la intensidad de sus picos. En este trabajo se utilizan esos datos, convertidos a su correspondiente *KEGG Compound ID*, para el metaanálisis e integración.

Antes de proceder al análisis de intensidad diferencial se hizo un análisis exploratorio de los datos, incluyendo diagramas de cajas, *clusters* jerárquicos y PCAs obtenidos con funciones del paquete *ggplot2* (Wickham, 2009). El análisis permite comprobar la distribución de la intensidad de los metabolitos e identificar posibles anomalías en los datos.

2.2.2. Análisis de intensidad diferencial

Antes de la realización del análisis de intensidad diferencial, comprobaremos la presencia de valores negativos. Sí existen, para corregirlos se debe sumar a todas las variables el valor mínimo existente, transformándose todos los datos valores positivos.

Una vez escalados los datos, se realizó el análisis de intensidad diferencial de metabolómica aplicando el mismo procedimiento que en transcriptómica. Los contrastes fueron siempre enfermo frente a control y las comparaciones fueron adaptadas a los datos disponibles. Se hicieron diversas comparaciones combinando las variables clínicas, descartando aquellas que no eran informativas para la integración. Finalmente se optó por utilizar los datos procedentes de 8 contrastes (Tabla 14).

Tabla 14. Contrastes estadísticos de metabolómica seleccionados para el metaanálisis.

Análisis de intensidad diferencial metabolómica	
Estudio	**Comparación**
CP	*FI + FA* vs. *Control*
CP	*FI + FA* vs. *EB*
LLC	*IGHV-* + *IGHV+* vs. *Control*
MM	*Diagnosis* vs. *Remisión*
MM	*Diagnosis* vs. *Control*
NMP	*PV* vs. *Control*
NMP	*TE* vs. *Control*
NMP	*TE* vs. *TS*

Siguiendo el mismo procedimiento que en transcriptómica, la expresión de los genes se calculó con la función *lmfit* del paquete *limma* (Ritchie, 2015) y se ajustó con *eBayes.* El ratio de falsos descubrimientos se corrigió con la tasa de falsos positivos de Benjamini & Hochberg (1995).

2.2.3. Análisis de enriquecimiento de grupos de metabolitos

Una vez se ha concluido el análisis de expresión diferencial se tiene un listado de metabolitos ordenados según la magnitud de las diferencias encontradas en su intensidad en cada comparación. Para realizar el análisis de enriquecimiento de grupos de metabolitos se necesitan listados que agrupen los metabolitos por sus funciones o por las rutas comunes en las que participan.
Los listados correspondientes a reacciones metabólicas de *KEGG* y los metabolitos participantes habían sido descargados utilizando el paquete *RbioRXN* (Min *et al.,* 2015). Los términos *GO* a los que estaban asociados los metabolitos se obtuvieron asociando las rutas *KEGG* con sus correspondientes términos *GO*.

Se siguió el mismo procedimiento que en transcriptómica, la expresión de los grupos se obtuvo con un método de enriquecimiento funcional de grupos de genes basado en modelos de regresión logística (García-García, 2016; Montaner *et al.* 2009; Sartor *et al.,* 2009; Montaner & Dopazo 2010). La única diferencia es que usaron ficheros específicos para la anotación de metabolitos, por términos *GO* y rutas KEGG, y que no se eliminaron aquellos grupos con pocos metabolitos con la función *annotFilter*. Tras aplicar la función *uvGSA* (*Uni-Variate Gene Set Analysis*) se obtuvieron los grupos de metabolitos con un

patrón común de intensidad que además estaban asociados a una determinada función, para cada cada uno de los contrastes.

Tabla 15. Resultados obtenidos del análisis de enriquecimiento de grupos del contraste de PV vs. Control de los datos metabolómicos de NMP.

Ruta KEGG	Descripción	N	*LOR*	*pval*	*padj*	*SD*	*t*	*conv*
hsa00010	Glicólisis / Gluconeogénesis	4	1.292	0.047	1.000	0.631	2.049	1.000
hsa00020	Ciclo del ácido cítrico	2	-0.092	0.903	1.000	0.751	-0.123	1.000
hsa00030	Vía pentosa fosfato	2	0.443	0.556	1.000	0.746	0.594	1.000
hsa00040	Interconversiones de pentosa y glucuronato	3	-0.193	0.757	1.000	0.620	-0.312	1.000
hsa00051	Metabolismo de fructosa y manosa	2	2.186	0.003	1.000	0.694	3.150	1.000
hsa00052	Metabolismo de la galactosa	4	0.105	0.848	1.000	0.545	0.193	1.000
hsa00053	Metabolismo de ascorbato y aldarato	3	-0.223	0.720	1.000	0.618	-0.361	1.000
hsa00061	Biosíntesis de ácidos grasos	0	0.000	1.000	1.000	0.157	0.000	1.000

Con los datos de metabólica el *output* de la función *uvGsa* fue una matriz idéntica a la de transcriptómica que se usará para el metaanálisis (Tabla 15). En este caso, la *N* indica el número de metabolitos participantes de la ruta que se encuentran sobrerrepresentados. El resto de estadísticos ya fueron explicados en el apartado de transcriptómica.

3 Metaanálisis

El metaanálisis permite valorar de manera conjunta los estadísticos de cada una de las comparaciones *Casos* frente a *Control* realizadas en cada ómica. Para ello el primer paso consiste en extraer de cada una de las matrices los valores de *LOR, SD* y valor de P ajustado asociados a los grupos de estudio. Estos datos se guardan en tres matrices individuales que son los datos de entrada de los metaanálisis que realizaremos.

Una vez se dispone de las matrices que contienen los estadísticos de los contrastes se realiza el metaanálisis que combina todos los resultados de los estudios individuales. Primero, se realiza una representación gráfica de todos ellos para determinar si hay algún grupo de muestras o genes que presentan un comportamiento anómalo en los contrastes y que se debiera revisar (Fig. 13).

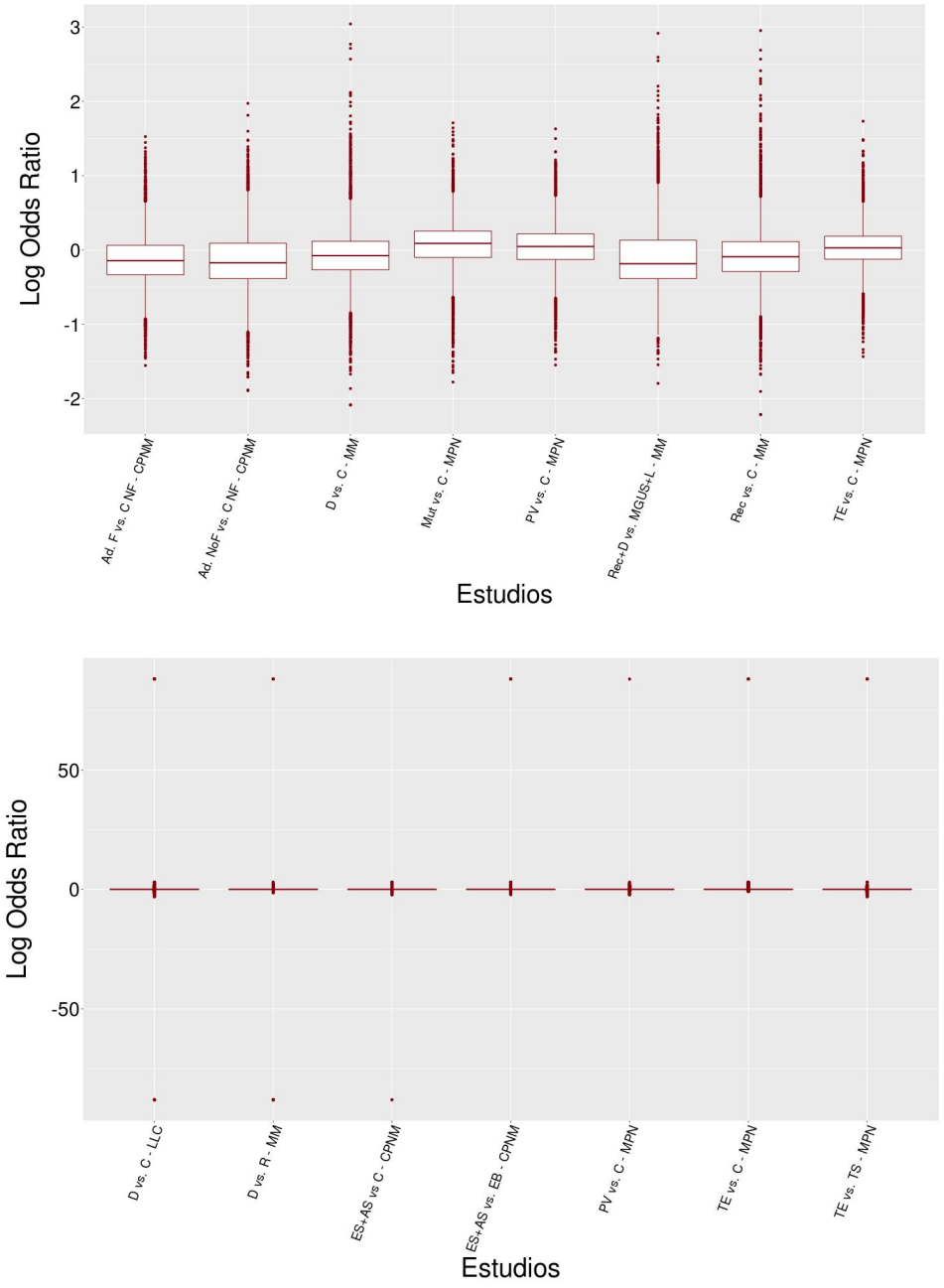

Figura 13. Distribución de la medida del efecto (LOR) por estudio de los grupos de funciones moleculares sobrerrepresentadas en (A) metabolómica y (B) transcriptómica.

La combinación de los resultados anteriores se podrá realizar utilizando un modelo de efectos fijos o aleatorios, en función de la variabilidad existente entre los estudios individuales.

El modelo de métodos de efectos fijos asume que no hay heterogeneidad entre los resultados y que las diferencias observadas se deben únicamente al azar. Con este método se le aplica a todos los modelos el mismo peso a la hora de realizar del metaanálisis. En un estudio de metaanálisis aplicar esta técnica no sería la metodología adecuada ya que, por una parte los resultados de estudios con muchas muestras ocultan a aquellos de menor tamaño y también se asume, de manera errónea, que todos los estudios proceden de una misma población, las muestras no tienen variabilidad técnica o

que se usan siempre las mismas variables en los contrastes (García-García, 2016; Steele & Tucker, 2008).
Algunos de los diferentes modelos de efectos aleatorios que se pueden emplear son: DL (DerSimonian & Laird, 1986), HE (Hedges *et al.,* 1999), HS (Schmidt & Hunter, 2014), entre otros.
La aplicación de métodos de efectos aleatorios será más adecuada en los escenarios con presencia de heterogeneidad entre los estudios y también dentro de sus muestras. El paquete *metafor* (Viechtbauer, 2010) contiene una colección de funciones que permiten realizar metaanálisis en R utilizando los métodos de efectos antes mencionados así como también funciones para realizar gráficas que permiten analizar los resultados.

La función que utilizada para el metaanálisis fue *rma.uni*. Esta función necesita como datos de entrada dos de la matrices ya creadas, la que contenía los valores del contraste (*LOR*) y una segunda con los errores estándar. También se debe indicar a la función el modelo aleatorio que modelo de efectos se desea aplicar. En total, no se efectúa un sólo metaanálisis, si no uno por cada función a analizar y modelo de efectos seleccionado. Con la información de estudios similares y los resultados definitivos ya obtenidos, se decidió que el modelo de efectos adecuado era DL, tanto para los datos de metabolómica como transcriptómica. La selección del modelo depende de los datos con los que se trabaja, en caso de no conocer cual es modelo adecuado se debe repetir el metaanálisis con cada modelo de efectos y valorar los resultados estadísticamente (Anexo A.1).

Como *output* de la función metanálisis se obtiene una tabla por cada método de efectos aleatorios aplicados (Tabla 16). A la hora de seleccionar los valores significativos se aplicó la tasa de falsos descubrimientos de Benjamini & Hochberg (1995). Para los datos de transcriptómica el nivel de significación utilizado fue de 0.05, mientras que en el metaanálisis de metabolómica se utilizó 0.01 ya que 0.05 era demasiado restrictivo.

Tabla 16. Resultados obtenidos del metaanálisis de transcriptómica según las rutas *KEGG* aplicando el modelo de efectos aleatorios DL.

Ruta KEGG	L. inferior	LOR medio	L. superior	P-valor	P-ajustado
hsa00010	0.104	0.194	0.285	0	0.001
hsa00020	-0.021	0.319	0.659	0.066	0.21
hsa00030	0.073	0.211	0.349	0.003	0.027
hsa00040	0.087	0.239	0.391	0.002	0.023
hsa00051	0.009	0.195	0.382	0.04	0.162
hsa00052	0.056	0.212	0.368	0.008	0.053
hsa00053	-0.14	0.034	0.208	0.7	0.805
hsa00061	-0.098	0.181	0.461	0.204	0.401

Ruta KEGG	QE	QEp	SE	tau2	I2	H2
hsa00010	1.7	0.975	0.046	0	0	1

hsa00020	51.213	0	0.173	0.208	86.332	7.316
hsa00030	4.853	0.678	0.07	0	0	1
hsa00040	4.214	0.755	0.078	0	0	1
hsa00051	15.626	0.029	0.095	0.04	55.202	2.232
hsa00052	9.086	0.247	0.08	0.012	22.958	1.298
hsa00053	5.924	0.549	0.089	0	0	1
hsa00061	14.312	0.046	0.143	0.083	51.09	2.045

La tabla de resultados contiene diversos estimadores que permiten valorar la medida del efecto, el valor de *log odds ratio,* de la ruta o el término *GO* en el estudio y de la heterogeneidad del metaanálisis. A continuación, se explicará el significado de cada indicador utilizando los resultados del metanálisis de transcriptómica como ejemplo.
Los estimadores son (García-García, 2016; Viechtbauer, 2010):

- *LOR* medio. Es el logaritmo de *odds ratio* (*LOR*), es la medida del efecto combinado de todos los estudios. Dado que los contrastes de los que se obtiene este valor son *Caso* frente a *Control*, un *LOR* con valor negativo significa que este grupo se encuentra sobrerrepresentado en el grupo control. El valor del estadístico indica la proporción exacta de la variabilidad entre ambos grupos.

Las rutas *KEGG* de transcriptómica presentes en la tabla superior se encuentran todas sobreexpresadas en el grupo de individuos enfermos frente a los controles, dado que todos los *LOR* medio son positivos.

- Límite inferior y superior. Son los valores máximos y mínimos, dentro del intervalo de confianza al 95%, de los contrastes realizados para determinar la heterogeneidad o homogeneidad de las muestras. A partir de este intervalo se establece la significatividad del estadístico *LOR*.

Aplicando el intervalo de confianza a los valores superiores (Tabla 16), el valor de *LOR* confirma que hsa00010 es una ruta significativa, ya que su intervalo se mantiene dentro de valores positivos, entre 0.104 y 0.285. Sin embargo, en tres de la rutas, por ejemplo hsa00061, el intervalo de los contrastes va desde un valor negativo, -0.098, a un positivo, 0.461. Esto indica que hay variabilidad en los contrastes de las muestras, en algunos la ruta está más activa en los *Casos*, y en otros, en los *Controles*, restando por tanto significatividad a esta ruta en el metaanálisis.

- *QE* y *QEp*. Son el estadístico de contraste y su respectivo valor de P (*QEp*) obtenidos al aplicar el método de efectos aleatorios seleccionado. Con ellos es posible determinar la presencia o ausencia de heterogeneidad entre las muestras. El nivel de significación del *QEp* habitual para aceptar la hipótesis alternativa es de 0.05, el mismo que se aplica en análisis de expresión diferencial o similares.

Basándonos en los resultados del valor de *P* de *QE* se aceptará la hipótesis alternativa, que existe heterogeneidad entre los estudios, siempre que el valor sea inferior a 0.05. Según los resultados de la tabla superior existe heterogeneidad entre las muestras,

mientras que otras son más homogéneas. En este contexto de variabilidad, sería correcto aplicar un modelo de efectos aleatorios.

- Valor de *P* y valor de *P* ajustado. Indican el nivel de significación del efecto combinado de todos los resultados y, dado que se está trabajando con un número elevado de contrastes, es necesario ajustar el valor aplicando métodos de corrección estadísticos. En este caso, se ha aplicado la tasa de falsos descubrimientos (Benjamini & Hochberg, 1995).

Comparando nuevamente la ruta hsa00010 frente a la hsa00061, el valor de *P* de la primera tiene un valor significativo en los contrastes del metaanálisis. Mientras que en la ruta hsa00061 el valor de *P* no le da validez, esto es consistente con los valores de LOR donde se detecta variabilidad en la expresión de esta ruta en los diferentes contrastes del metaanálisis.

- *tau2*. Es el indicador de heterogeneidad entre los estudios utilizados para el metaanálisis. Si se aplica un modelo de efectos fijos sobre los datos el valor será siempre igual a cero.

Según los resultados obtenidos, en algunas funciones existe heterogeneidad en el comportamiento de las muestras y, en otros el estimador es cero indicando que hay homogeneidad en su comportamiento. Este valor, similar a *QEp*, nos permiten identificar aquellas rutas que tienen un comportamiento muy parecido, sin importar el estudio en cuestión, rutas centrales en el desarrollo de la cancerogénesis.

- *I2*. Otro indicador empleado para valorar la heterogeneidad de los estudios que representa el porcentaje total de variabilidad entre los estudios debida a su heterogeneidad. Los valores negativos son convertidos a cero de forma que el porcentaje siempre se encuentra entre 0 y 100. Un 0% indica que no hay heterogeneidad entre las muestras y un valor elevado que si la hay (Thompson *et al.,* 2003).

Como ocurría con el estadístico *tau2*, *I2* también indica la heterogeneidad arrojando resultados similares. La diferencia es que ahora es posible estimar de manera porcentual la diferencia de expresión de la ruta en los diferentes contrastes.
Un detalle que se debe tener en cuenta es que este indicador, y también tau2, presenta sesgos cuando el número de muestras en el estudio es inferior a siete (von Hipper, 2015). Por este motivo, en ausencia de heterogeneidad entre las muestras, puede sobrerrepresentarla y, en presencia, infrarrepresentarla. Dado que el estudio realizado presenta un número pequeño de muestras, a la hora de validar los resultados es más certero hacerlo utilizando el intervalo de confianza del 95% calculado en los límites superiores e inferiores de *LOR*.

- *H2*. Este estadístico es similar al anterior pero indica el cociente entre la variabilidad total y la variabilidad de las muestras. Por tanto, si no existe heterogeneidad este valor será igual a 1.

Si se observan los datos de la Tabla 16, cuando *tau2* es igual a 0, *H2* es igual a 1. Mientras que en aquellos contrastes en los que había mucha variabilidad dentro de las distintas muestras este valor es superior a 1.

En resumen, los estadísticos obtenidos sirven para valorar la heterogeneidad de las rutas en los diferentes contrastes incluidos en el metaanálisis. Se observa que, en la mayoría de los casos, encontramos heterogeneidad dentro de los diferentes estudios. Pero que, a pesar de ello, los valores de *P* obtenidos tienen valores que indican la significancia de estas rutas. Esto quiere decir que, estas rutas presentan comportamientos diferentes en función del tipo de cáncer o contraste realizado pero que, en general, dentro de la disparidad de sus valores, estos son siempre diferentes a las muestras control. En otros casos, vemos que una misma ruta tiene un comportamientos claramente opuestos, como se observaba en la ruta hsa00061. Este resultado debe ser valorado desde dos ópticas distintas, la primera es considerar un posible error de muestreo y la segunda es atribuirla a la variabilidad intrínseca de los procesos cancerosos. Debemos recordar que los datos de este metaanálisis no proceden de un mismo tipo de cáncer sino de cuatro diferentes. De forma que, es de esperar, no todos tengan el mismo comportamiento. Con este metaanálisis lo que se podrá detectar, con significancia estadística, es la base central del proceso de carcinogénesis.

3.1. Representación e interpretación de resultados globales

El primer resultado obtenido tras la realización del metaanálisis, es una tabla resumen que contiene el número anotaciones significativas según la ontología de términos *GO* (proceso biológico, componente celular o función molecular) o rutas de señalización *KEGG* (Tabla 17).

Tabla 17. Resultados globales del metaanálisis funcional de transcriptómica utilizando rutas de señalización *KEGG*.

Métodos	*Casos*	*Control*	*Sig.Casos*	*Sig.Control*	*Sig.LOR.Casos*	*Sig.LOR.Control*
DL	120	201	23	21	1	0
HE	119	201	23	23	1	0
HS	119	202	24	25	1	0
FE	119	202	56	120	1	1

En la tabla superior, de los resultados del metanálisis de transcriptómica, tenemos por filas las funciones identificadas como significativa por cuatro métodos de estimación de la variabilidad del efecto diferentes. Las columnas contienen información referente a las rutas y anotaciones sobrerrepresentadas e infrarrepresentadas. A continuación analizamos el significado de las columnas una a una:

- *Control*, hace referencia al número de elementos sobrerrepresentados de grupos de genes con un nivel de expresión alto en el grupo de *Casos*, individuos enfermos.
- *Casos*, indica el número de elementos sobrerrepresentados de grupo de genes con un nivel de expresión alto en el grupo de controles, los individuos sanos.

Hay 120 rutas *KEGG* se encuentran sobrerrepresentadas en el grupo de *Casos* frente a 201 en los *Controles*.

- *Sig.Control* y *Sig.Casos*, son el número de elementos sobrerrepresentados y significativos de cada grupo experimental.

Se aplicó sobre los valores de *P* anteriores la tasa de falsos descubrimientos (Benjamini & Hochberg, 1995) para corregir el número de falsos positivos consecuencia de las comparaciones múltiples. En transcriptómica, el nivel de significación elegido fue de 0.05; al seleccionar las rutas con un valor de *P* inferior, el número original de rutas *KEGG* desciende a 23 significativas en el grupo *Casos* y 21 en el grupo *Control*.
Comparando los valores en función del modelo de efectos aplicado se aprecia que cambia significativamente según el método. Por ejemplo, si se aplica el modelo de efectos fijos se obtiene un número elevado de rutas y funciones significativas, al obviarse la heterogeneidad existente entre las muestras y asumir que cualquier cambio es consecuencia de las diferencias entre los grupos *Caso* y *Control*.

- *Sig.LOR.Control* y *Sig.LOR.Casos*, son el número de elementos sobrerrepresentados de manera significativa y cuyo efecto de sobrerrepresentación, *LOR,* es mayor de 0.5.

Este estadístico selecciona de forma mucho más restrictiva las funciones y rutas de interés. Al aplicarlo, el número de rutas *KEGG* sobrerrepresentadas y significativas en el metaanálisis desciende a 1 para los *Casos* y es nula para los *Controles*.

Gráficos volcán

Los gráficos volcán son diagramas de puntos que permiten representar numerosos datos ómicos y detectar cambios de interés. Se pueden obtener con la función *ggplot2* (Wickham, 2009) utilizando los datos obtenidos con la función *rma.uni* para un modelo de estimación de efectos determinado.. En el eje X se indica la magnitud del cambio, el logaritmo en base 2 del *odds ratio,* mientras que en la Y se indica una medida de significación estadística, en este caso el log_{10} P-value ajustado.

3.2.Representación e interpretación de resultados a nivel de función del metaanálisis

Dado el alto volumen de datos con el que se trabaja en los metaanálisis, es habitual representarlos gráficamente para su valoración. A continuación, se resumen algunos gráficos utilizados para analizar la heterogeneidad de los estudios y la presencia de sesgos en las funciones o rutas biológicas estudiadas en función del modelo de efectos seleccionado.

Gráficos de bosque

Los gráficos de bosque (Fig. 14) se obtienen con la función *forest* del paquete *metafor* (Viechtbauer, 2010). El gráfico representa un “bosque”, donde los árboles son los contrastes de la función de interés en cada uno de los estudios del metaanálisis. El eje X es el valor del efecto (*odds ratio*), a la derecha aparecen los valores del efecto globales y su intervalo de confianza al 95%. Los cuadrados negros sobre los ‘’árboles’’ son el valor medio para ese contraste, mientras que el rombo rojo inferior indica el resultado global de efecto para todos los contrastes. La anchura de ambos símbolos simboliza la precisión del resultado.

En el ejemplo de la Figura 14, la mayoría de los contrastes están a la derecha de la línea perpendicular discontinua que indica donde el efecto es nulo. La posición del rombo, también a la derecha, corrobora la tendencia global de los contrastes y por tanto la función estaría sobrerrepresentada en los *Casos*. En este caso, también se aprecia que en los contrastes de una misma enfermedad, el comportamiento es similar, aunque no siempre es así. Cuanto más alejado esté el rombo del cero, más consistente será el resultado, al indicar que hay variabilidad entre los *Casos* y los *Controles*.

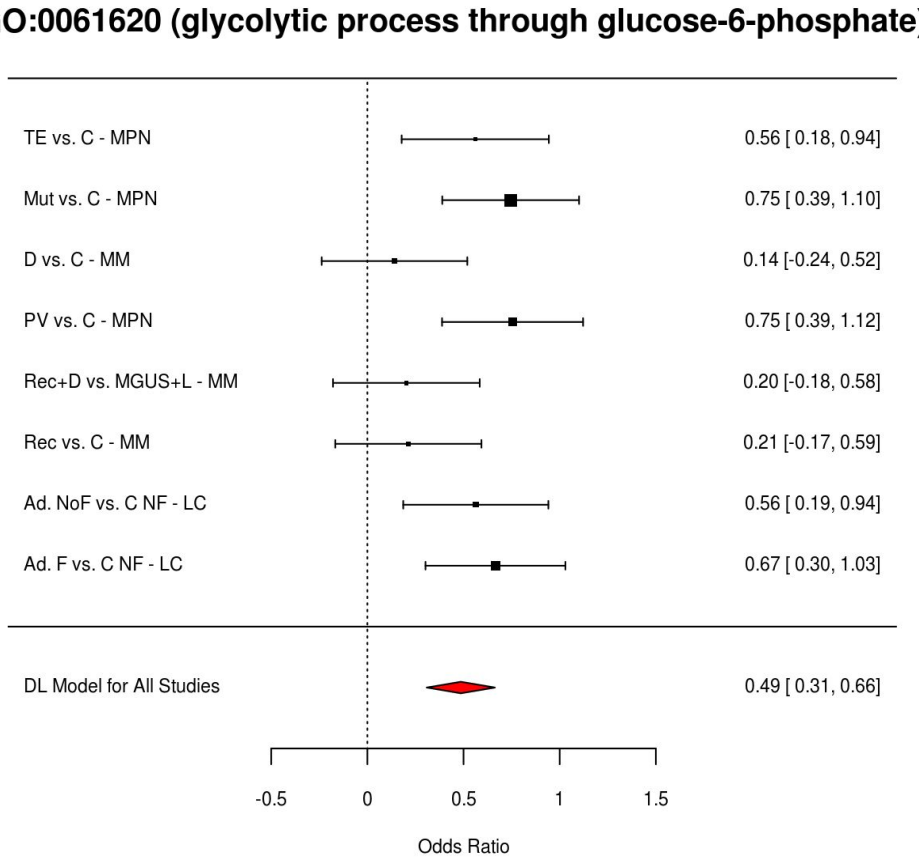

Figura 14. Gráfico de bosque. Distribución del efecto para la función GO:0061620 en transcriptómica.

Gráficos de embudo

Los gráficos de embudo evalúan la presencia de sesgos y heterogeneidad en los datos utilizados (García-García, 2016; Sterne & Egger, 2001). Se han realizado con la función *funnel* presente en el paquete *metafor* (Viechtbauer, 2010).

En ausencia de sesgo, se espera que los datos estén próximos a la media central y aquellos de poca precisión se encuentren fuera del embudo, la región de confianza. En un gráfico de embudo (Fig. 15) se encuentra en el eje X la magnitud del efecto medido, en este caso se trata del *log odds ratio*, mientras que en el eje Y hay un estadístico de dispersión, como la desviación estándar o el inverso de la varianza.

En la Figura 15 se representa, por tanto, la relación entre el *log odds ratio* y el error estándar para el proceso GO:0000002 según el modelo de efectos aleatorios DL. En ausencia de sesgo y variabilidad, la distribución de las muestras debe ser siempre

homogénea y distribuida a lo largo del eje X. Si no lo fuera, y los puntos estuvieran dispersos por toda el área, significaría que hay un sesgo muestral que puede deberse a la falta de muestras en el estudio. Otro indicador de heterogeneidad es la distribución de los puntos en el eje, en este caso vemos como la mayoría de valores se concentran en el lado positivo del eje X. La medida del efecto tiene una tendencia positiva en los contrastes analizados, lo ideal sería trabajar con un balance más equilibrado entre positivos y negativos.

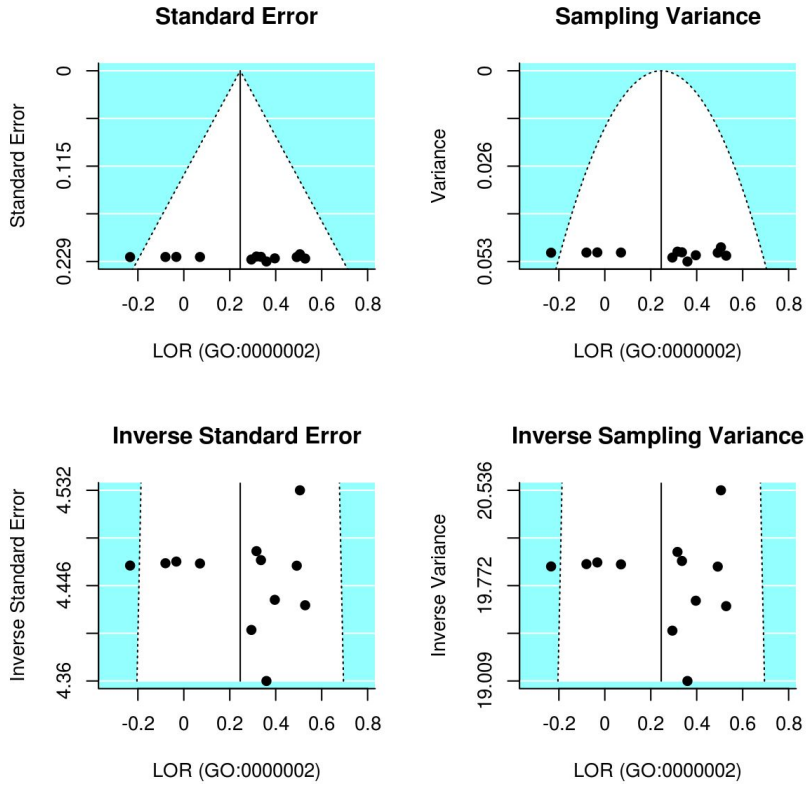

Figura 15. Gráfico de embudo. Variabilidad del efecto estudiado en la función GO:0000002 aplicando el modelo de efectos aleatorios DL en los datos de transcriptómica.

Gráficos radial

Los gráficos radial valoran la consistencia de los efectos según su nivel de precisión (García-García, 2016; Galbraith, 1988). La función *radial,* utilizada para realizar las gráficas se encuentra incluida en el paquete *metafor* (Viechtbauer, 2010).
En el eje X se representa un estadístico de dispersión, por ejemplo el error estándar, y en el eje Y el tamaño del efecto observado estandarizado por el estadístico de dispersión (Fig. 16). El arco en la derecha de la gráfica representa la localización del efecto observado pero sin estandarizar. Para valorar la magnitud del efecto, se debe trazar una proyección desde el eje de coordenadas que atraviese el estudio de interés y corte sobre el arco, obteniendo la magnitud del efecto.

En la Figura 16 se representa el gráfico radial de los resultados obtenidos para función GO:0098634 en el metaanálisis de transcriptómica. El ejemplo seleccionado es atípico en comparación a los otros gráficos obtenidos. En el gráfico se observan sólo 3 puntos. Uno de ellos, corresponde a los 3 contrastes de la función en MM que al ser idénticos se solapan. Otro de los puntos, es consecuencia del solapamiento de dos de los tres

estudios de NMP. Los dos contrastes que faltan en esta gráfica son los de CP, cuyos datos de transcriptómica proceden de un array diferente. Por lo tanto, o bien falta información de genes que participan en este proceso, o bien esta función no está sobrerrepresentada en CP. Otro motivo que convierte a esta gráfica en peculiar es que se distinguen dos patrones de variabilidad distintos. Normalmente la distribución de la línea de puntos debe ser vertical y perpendicular al eje Y. La diferente distribución en dos grupos indica diferencias en la precisión del cálculo de la medida del efecto.

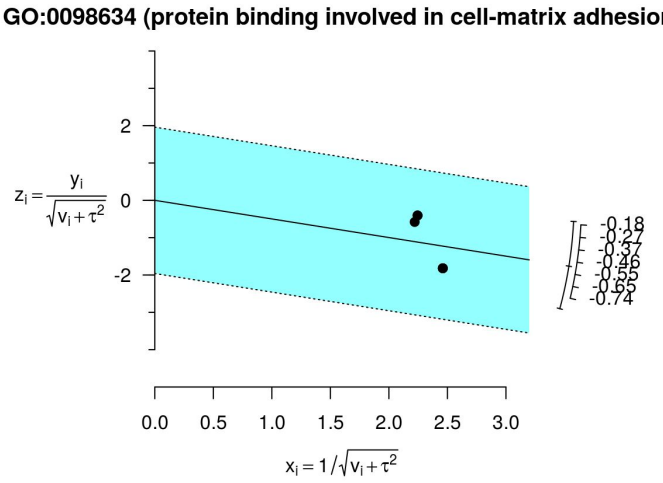

Figura 16. Gráfico radial. Variabilidad del efecto estudiado en la función GO:0098634 aplicando el modelo de efectos aleatorios DL en los datos de transcriptómica.

4 Integración

La integración de los resultados ofrece información desde tantos enfoques como fuentes utilizadas. En el inicio de este trabajo se resumieron las diferentes metodologías actuales empleadas en la integración de datos ómicos. La mayoría de ellas están basadas en la integración de dos grupos de datos, por ejemplo, muestras de transcriptómica y metabolómica de un grupo de enfermos. En este trabajo se ha dado un paso más allá y antes de la integración se realiza un metaanálisis funcional de varios sets de datos, que proporciona la integración de los diferentes estudios correspondientes a una misma ómica. De modo que tras la aplicación de un metaanálisis funcional, no dispondremos de matrices de datos con la expresión de genes o intensidad de metabolitos, sino que dispondremos de resultados nivel funcional.

La estructura de estos resultados no permite la aplicación directa de la mayoría de metodologías de integración desarrolladas hasta el momento. Por ello, presentaremos una combinación de dos de ellas, donde se realizará una integración conceptual de los resultados pero partiendo de resultados de funciones significativas obtenidas por la anotación funcional de los genes/metabolitos de los datasets.

IV. Resultados y discusión

1 Transcriptómica

1.1. Procesamiento de los datos

Una vez los datos fueron procesados, se hizo un análisis exploratorio para detectar posibles outliers entre las muestras o algún otro comportamiento anormal de los datos. Se hicieron para ello diagramas de cajas, análisis de *clusters* así como análisis de componentes principales.

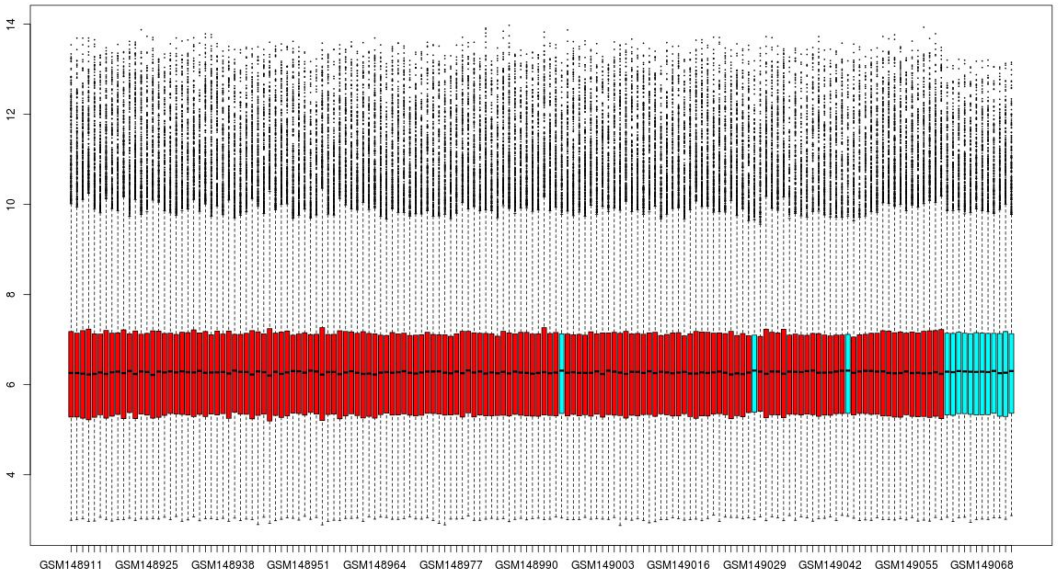

Figura 17. Diagrama de cajas correspondiente a las muestras de transcriptómica de MM.

En general, todas las muestras presentaban una distribución similar (Fig. 17) y que confirma que la normalización de los estudios fue correcta. Las representaciones gráficos de los resultados del análisis de componentes principales y análisis de clúster permitieron la exploración de los datos. Dado el alto volumen de muestras en los PCAs no son siempre visibles pero en los *clusters* jerárquicos se visualiza fácilmente (Fig. 18). En el clúster jerarquizado se observa una clara distinción entre las muestras procedentes de individuos sanos e individuos enfermos. Aún así a la hora de realizar los *clusters*, en algún caso, las muestras de individuos sanos no se agrupaban todas juntas. Por ejemplo, en el clúster jerarquizado por distancia euclídea de las muestras de CP, dos de ellas se agrupaban dentro de los *Controles* a pesar de ser muestras de *Casos*, y otra procedente de un individuo sano se agrupaba con los enfermos. Lo mismo ocurre en los otros *clusters*, donde una o dos muestras se encuentran distantes de su clase. Ni en los artículos originales donde se utilizaron estas muestras, ni en la base de datos *GEO* se hacía referencia a la presencia de muestras anómalas y consideradas como *outliers.* Por esta razón, y también para no perder controles, se optó por conservar todas las muestras en el metaanálisis.

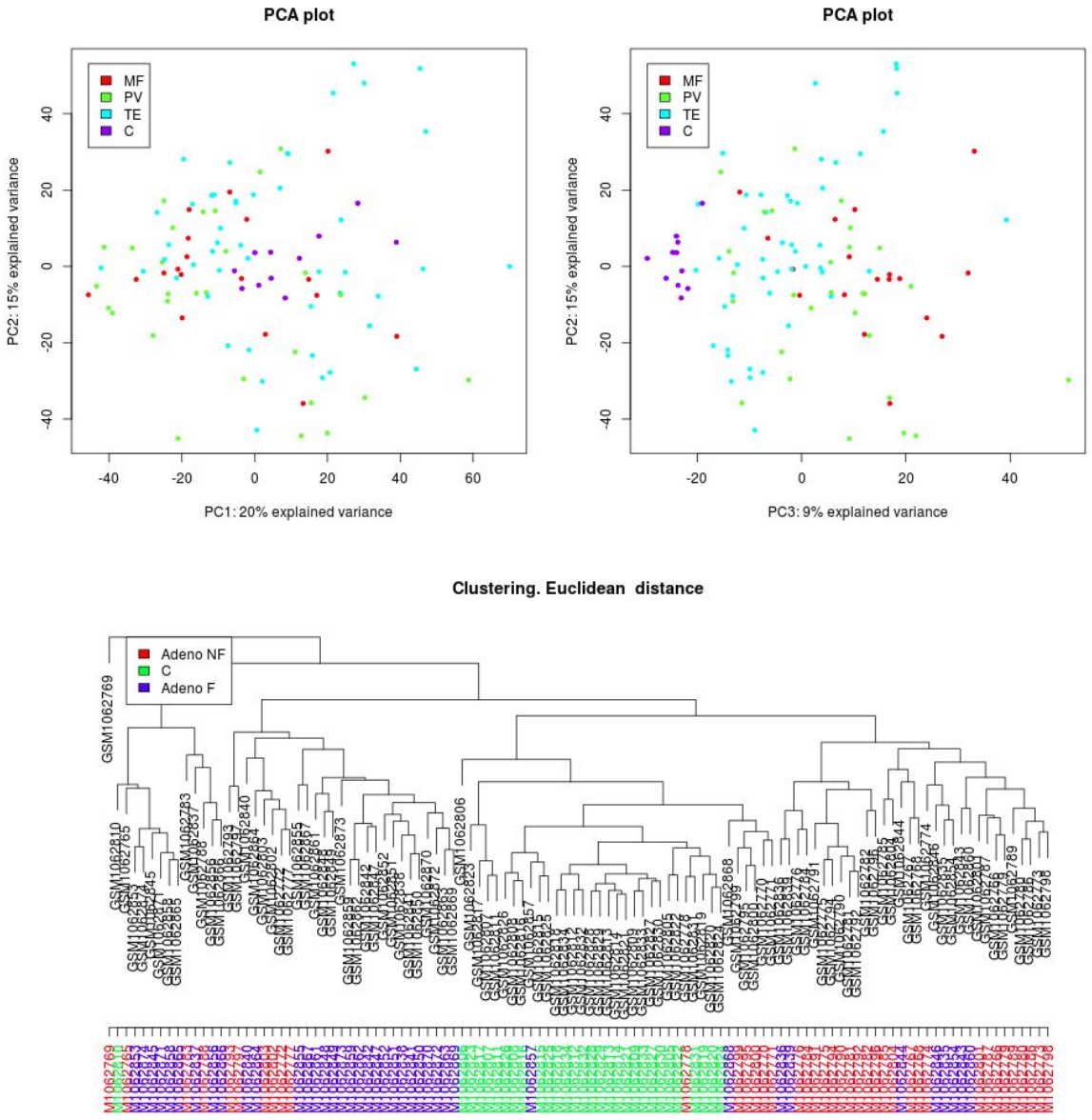

Figura 18. B) PCA de las muestras de NMP. C) Clúster jerárquico por correlación de las muestras de CP.

1.2. Análisis de expresión diferencial

En los contrastes estadísticos del análisis de expresión diferencial de transcriptómica, el orden de comparación de grupos experimentales fue: *Casos* frente a *Controles,* o pacientes frente a individuos sanos. En resultados obtenidos en los 8 contrastes definitivos (Tabla 18) se observa una tendencia alcista de los genes sobreexpresados (más expresados en los *Casos*) frente a los controles. El número más elevado es el de genes sin expresión diferencial, aquellos que a pesar del proceso de carcinogénesis no cambian su comportamiento.

Tabla 18. Nivel de expresión diferencial entre los grupos experimentales comparados de transcriptómica.

Análisis de expresión diferencial transcriptómica				
Estudio	**Comparación**	**Genes infraexpresados**	**G. sin expresión diferencial**	**Genes sobreexpresados**
CP	*Adeno NoF* vs. *Control NoF*	3,546	11,777	4,039
CP	*Adeno F* vs. *Control NoF*	5,521	6,878	6,963
MM	*Diagnosis* vs. *Control*	1,745	8,926	2,074
MM	*Recaída* vs. *Control*	1,451	9,350	1,944
MM	*Diag + Re* vs. *MGUS + Latente*	1,788	8,825	2,132
NMP	*PV* vs. *Control*	1,160	10,416	1,169
NMP	*TE* vs. *Control*	894	11,121	730
NMP	*JAK-* vs. *JAK+*	285	12,064	396

1.3. Análisis de enriquecimiento de grupos de genes

El análisis de enriquecimiento de grupo de genes se hizo para los 8 contrastes y utilizando la anotación de las rutas *KEGG* y los términos *GO* de las tres ontologías (componentes celulares, funciones moleculares y procesos biológicos).

Los resultados dependen del análisis de expresión diferencial y el contraste realizado, así como también los datos de origen. En las cuatro tablas inferiores (Tabla 19-22) se presentan los resultados obtenidos del análisis de enriquecimiento de CP, MM y NMP, utilizando la anotación basada en los términos *GO* y las rutas *KEGG*. Los grupos que se consideraron significativos fueron aquellos con un valor de *P* superior a 0.05.

Como se observa en las tablas, el número de grupos diferencialmente expresados en *Controles* es ligeramente mayor que el de *Casos*.

Tabla 19. Distribución de las funciones significativas según el término GO de Funciones biológicas en transcriptómica.

Análisis de Enriquecimiento. Funciones significativas *GO* BP					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F.sig.Casos***	***F.sig.Controles***
LC	*Adeno NoF* vs. *Control NoF*	6,654	1023	207	816
LC	*Adeno F* vs. *Control NoF*	7,139	538	62	476
MM	*Diagnosis* vs. *Control*	7,096	581	263	318
MM	*Recaída* vs. *Control*	6,991	686	265	421
MM	*Diag + Re* vs. *MGUS + Latente*	6,446	1231	575	656
NMP	*PV* vs. *Control*	7,520	157	89	68
NMP	*TE* vs. *Control*	7,653	24	7	17
NMP	*JAK-* vs. *JAK+*	7,357	320	172	148

Tabla 20. Distribución de las funciones significativas según el término GO de funciones moleculares en transcriptómica.

Análisis de Enriquecimiento. Funciones significativas *GO* MF					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F.sig.Casos***	***F. sig.Controles***
LC	*Adeno NoF* vs. *Control NoF*	1,492	74	32	42
LC	*Adeno F* vs. *Control NoF*	1,501	65	18	47
MM	*Diagnosis* vs. *Control*	1,490	76	49	27
MM	*Recaída* vs. *Control*	1,489	77	48	29
MM	*Diag + Re* vs. *MGUS + Latente*	1,355	211	124	87
NMP	*PV* vs. *Control*	1,533	33	17	16
NMP	*TE* vs. *Control*	1,547	19	12	7
NMP	*JAK-* vs. *JAK+*	1,543	23	3	20

Tabla 21. Distribución de las funciones significativas según el término GO de componentes celulares en transcriptómica.

Análisis de Enriquecimiento. Funciones significativas *GO* CC					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F. sig.Casos***	***F. sig. Controles***
LC	*Adeno NoF* vs. *Control NoF*	769	145	66	79
LC	*Adeno F* vs. *Control NoF*	808	106	17	89
MM	*Diagnosis* vs. *Control*	775	139	90	49
MM	*Recaída* vs. *Control*	760	154	89	65
MM	*Diag + Re* vs. *MGUS + Latente*	657	257	178	79
NMP	*PV* vs. *Control*	811	103	48	55
NMP	*TE* vs. *Control*	877	37	13	24
NMP	*JAK-* vs. *JAK+*	806	108	41	67

Tabla 22. Distribución de las funciones significativas según las rutas KEGG en transcriptómica.

Análisis de Enriquecimiento. Funciones significativas *KEGG*					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F. sig. Casos***	***F. sig. Controles***
LC	*Adeno NoF* vs. *Control NoF*	207	114	25	89
LC	*Adeno F* vs. *Control NoF*	253	68	8	60
MM	*Diagnosis* vs. *Control*	277	44	18	26
MM	*Recaída* vs. *Control*	255	66	25	41
MM	*Diag + Re* vs. *MGUS + Latente*	218	103	40	63
NMP	*PV* vs. *Control*	298	23	11	12
NMP	*TE* vs. *Control*	311	10	2	8
NMP	*JAK-* vs. *JAK+*	287	34	24	10

Al comparar los contrastes uno a uno se detecta variabilidad en las funciones más significativas. En la Tabla 23 se incluyen las tres funciones más significativas en MM, en NMP estas funciones también son las más significativas sin embargo su valor de *LOR* es inverso. En CP, estas funciones no ocupan las primeras posiciones.
Al ser datos procedentes de 3 tipos de tumores concretos es esperable que tengan un comportamiento diferente. Con el metaanálisis será posible identificar las homologías más allá de estas diferencias.

Tabla 23. Comparativa de los resultados estadísticos obtenidos según las rutas KEGG en transcriptómica.

***GO* BP**	**Descripción**	**N**	**LOR**	**padj**
GO:0034660	MM - *Diagnosis* vs. *Control*	451	1.0169	2.38E-67
Proceso metabólico del ncRNA	NPM - *PV* vs. *Control*	451	-0.5822	8.02E-27
	CP - *Adeno NoF* vs. *Control NoF*	448	0.4182	9.21E-15
GO:0022613	MM - *Diagnosis* vs. *Control*	387	1.0602	1.49E-55
Biogénesis del complejo ribonucleoproteico	NPM - *PV* vs. *Control*	387	-0.7689	1.33E-34
	CP - Adeno NoF vs. Control NoF	383	0.3477	1.34E-08
GO:0034470	MM - *Diagnosis* vs. *Control*	322	1.1819	5.26E-55
Procesamiento de ncRNA	NPM - *PV* vs. *Control*	322	-0.7689	1.29E-31
	CP - Adeno NoF vs. Control NoF	319	0.4095	4.27E-10

1.4. Metaanálisis

A partir de los resultados obtenidos con los datos y la bibliografía, se había optado por utilizar el modelo de efectos aleatorios DL. A pesar de existir otras alternativas recomendadas, el modelo sigue siendo el más utilizado dado lo sencillo que es aplicarlo y los buenos resultados que se obtienen (Veroniki *et al*., 2015).
En la Tabla 24 se resumen aquellas funciones que el metaanálisis ha identificado como significativas, y que además su valor de efecto es superior al 0.5.

Tabla 24. Resultado global del metaanálisis funcional de transcriptómica aplicando DL.

Anotación	*Casos*	*Control*	*Sig.Casos*	*Sig.Control*	*Sig.LOR.Casos*	*Sig.LOR.Control*
GO F. molecular	737	832	56	68	11	12
GO P. biológico	2,633	5,032	260	420	43	41
GO C. celular	479	433	42	33	13	5
KEGG	120	201	23	21	1	0

Puesto que el número de anotaciones que corresponden a la ontología de procesos biológicos es superior, detectamos un mayor número de grupos sobrerrepresentados. Como ocurría en el análisis de expresión por grupos de metabolitos, el grupo de funciones sobrerrepresentadas de los *Controles* es mayor, sin embargo las funciones significativas mayoritarias son de los *Casos*.

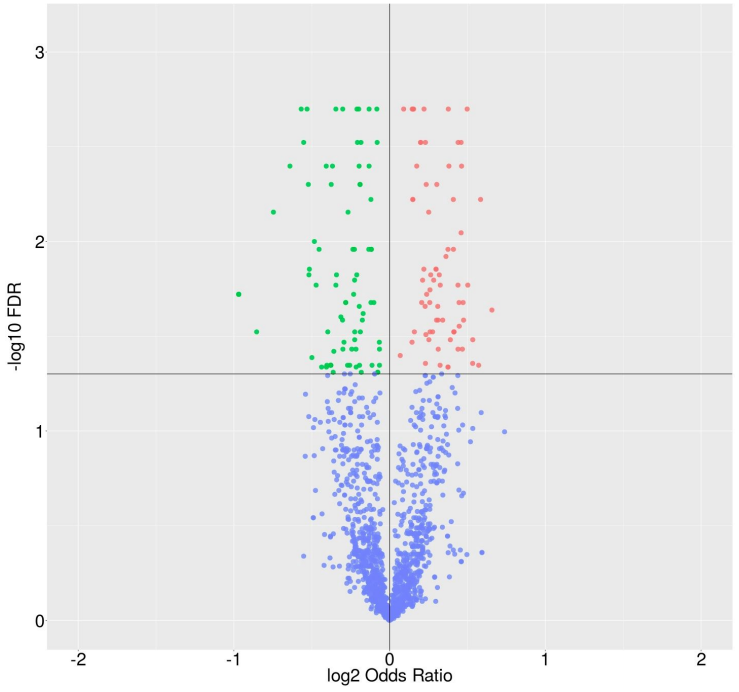

Figura 19. Resultados del metaanálisis de funciones moleculares en estudios de cáncer aplicando un modelo de efectos DL.

La mejor forma de visualizar los datos de una manera global es con un gráfico volcán (Fig. 19). Cada punto del gráfico es una función biológica según los términos *GO*. Las funciones significativas serán aquellas con un valor de *P* ajustado superior a 0.05. Los puntos rojos son las funciones en las que el efecto combinado de todos los contrastes es negativo (sobrerrepresentadas en el grupo *Control*) y los puntos verdes cuando es positivo (sobrerrepresentadas en el grupo *Casos*). La presencia de ambos grupos está equilibrada para las funciones biológicas, hay un número similar de funciones sobrerrepresentadas en los dos grupos. Los puntos azules simbolizan todas las funciones que, a pesar de estar sobrerrepresentadas, no lo hacen de manera significativa.

1.4.1. Funciones moleculares significativas del metaanálisis funcional

Se detectaron un total de 11 funciones significativas con un valor del efecto superior a 0.5 en los *Casos*, frente a un total de 18 funciones significativas con un valor del efecto superior a -0.5 en los *Controles*.

Tabla 25. Términos GO de las 8 funciones moleculares más sobrerrepresentadas y estadísticos seleccionados del metaanálisis de trascriptómica.

GO F.molecular	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
GO:0008494	Activación del activador de la traducción	0.679	1.00E-03	0.14	0
GO:0048256	Activación de la endonucleasa Flap1	0.645	0.001	0.139	0
GO:0002161	Actividad de edición por Aminoacil ARNt sintetasa	0.625	0.017	0.183	0.144
GO:0019238	Actividad ciclohidrolasa	0.623	0.006	0.163	0
GO:0070061	Unión de fructosa	0.615	0.005	0.155	0
GO:0005384	Actividad del transportador transmembrana de manganeso	0.606	0.006	0.156	0
GO:0003688	Unión al origen de replicación de DNA	0.578	0.011	0.161	0.132
GO:0043138	Actividad helicasa 3'-5'	0.575	0.006	0.149	0.082

En las funciones más sobrerrepresentadas de los *Casos* existen varias implicadas en los procesos de replicación y traducción (Tabla 25). También se sobreexpresa la endonucleasa FLAP1, enzima que participa en replicación y reparación de ADN, además de fragmentarlo durante la apoptosis. Se ha observado que su sobreexpresión aumenta la proliferación de las células cancerosas y las mutaciones en el ADN (Signh *et al.,* 2008). El aumento de calcio intracelular también está asociado a las células cancerosas y a la

proliferación celular, el manganeso puede funcionar como su sustituto en algunas reacciones, de ahí la activación de su transportador transmembrana (GO:0005384) (Braun *et al.,* 2012).

En los *Controles* también hay funciones sobrerrepresentadas con respecto a los *Casos*. Una de ellas es la función de actividad N,N-dimetilanilina monooxigenasa (GO:0004499). La rápida proliferación de las células cancerosas hace que carezcan de la vascularización necesaria para recibir nutrientes y oxígeno de manera eficiente. Una vez que la célula ha consumido todos sus recursos de oxígeno, se genera un ambiente de hipoxia, por lo que deja de usar las reacciones que utilizan oxígeno como aceptor de electrones. La inhibición de la función GO:0004499 se puede deber a la falta de oxígeno en el medio (Pavlola & Thompson, 2016; Eales *et al.,* 2016).

1.4.2. Procesos biológicos significativos del metaanálisis funcional

En la Tabla 26 se indican los procesos biológicos que tienen un *LOR* mayor, es decir, están más sobrerrepresentados en los *Casos*. Entre estos procesos biológicos se encuentra la síntesis de inositol (IMP) precursor de purinas (GO:0046040, GO:0006188, GO:0006189). Dos de los procesos más desregulados están relacionados con el procesamiento de RNA codificante (snoRNA y snRNA). Estas moléculas tienen funciones reguladoras y se ha observado que en el cáncer juegan un papel importante en la regulación de los procesos de metástasis y proliferación celular (Mannoor *et al.,* 2013).

Tabla 26. Términos GO de los 8 procesos biológicos más sobrerrepresentados y estadísticos seleccionados del metaanálisis de transcriptómica.

GO P. biológicos	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
GO:0006189	Síntesis de novo de inosina monofosfato	1.286	0.018	0.386	1.088
GO:0034472	Procesamiento del extremo 3' de snRNA	1.08	0	0.149	0.035
GO:0031126	Procesamiento del extremo 3' de snoRNA	1.067	0.025	0.337	0.777
GO:0006188	Biosíntesis de inosina monofosfato	0.855	0.001	0.192	0.173
GO:0000727	Vía de la replicación de ruptura inducida	0.777	0	0.144	0
GO:0046040	Proceso metabólico IMP	0.777	0	0.164	0.121
GO:0031507	Ensamblaje de heterocromatina	0.752	0	0.161	0.082
GO:0009396	Biosíntesis de compuestos con ácido fólico	0.732	0.001	0.16	0.051
GO:0034501	Transporte por el cinetocoro	0.732	0	0.12	0

Otra de las funciones sobrerrepresentadas, es la vía de la replicación de ruptura inducida. Esta ruta está implicada en la reparación de ADN pero en condiciones de carcinogénesis puede causar reordenamientos cromosómicos (Hasty & Montagna, 2014). También, en algunos tipos de tumores, es la responsable de convertir en inmortales a las células tumorales ya que participa, en sustitución de la telomerasa, en la elongación de sus telómeros (Dilley *et al.*, 2016).

Otra función de interés, pero que no presenta valor de LOR de 0.5, es la función GO:000000077 (Punto de control de daño de ADN). Como se puede ver en la Figura 20, la ruta se encuentra sobrerrepresentada en los *Casos*, aunque no todos los contrastes tienen el mismo comportamiento independientemente de su origen.

Esta variabilidad es de esperar, ya que estamos haciendo diferentes contrastes en los que puede existir un cierta variabilidad en el proceso de oncogénesis. Según consta en la bibliografía, en algunos tipos de cáncer, el comportamiento de esta función es distinto. Al iniciarse el proceso canceroso las células responden al punto de control, pero a medida que avanza, o bien las proteínas que participan de la función dejan de expresarse, o bien aumenta su expresión augurando una peor prognosis para el paciente (Wang *et al.,* 2015).

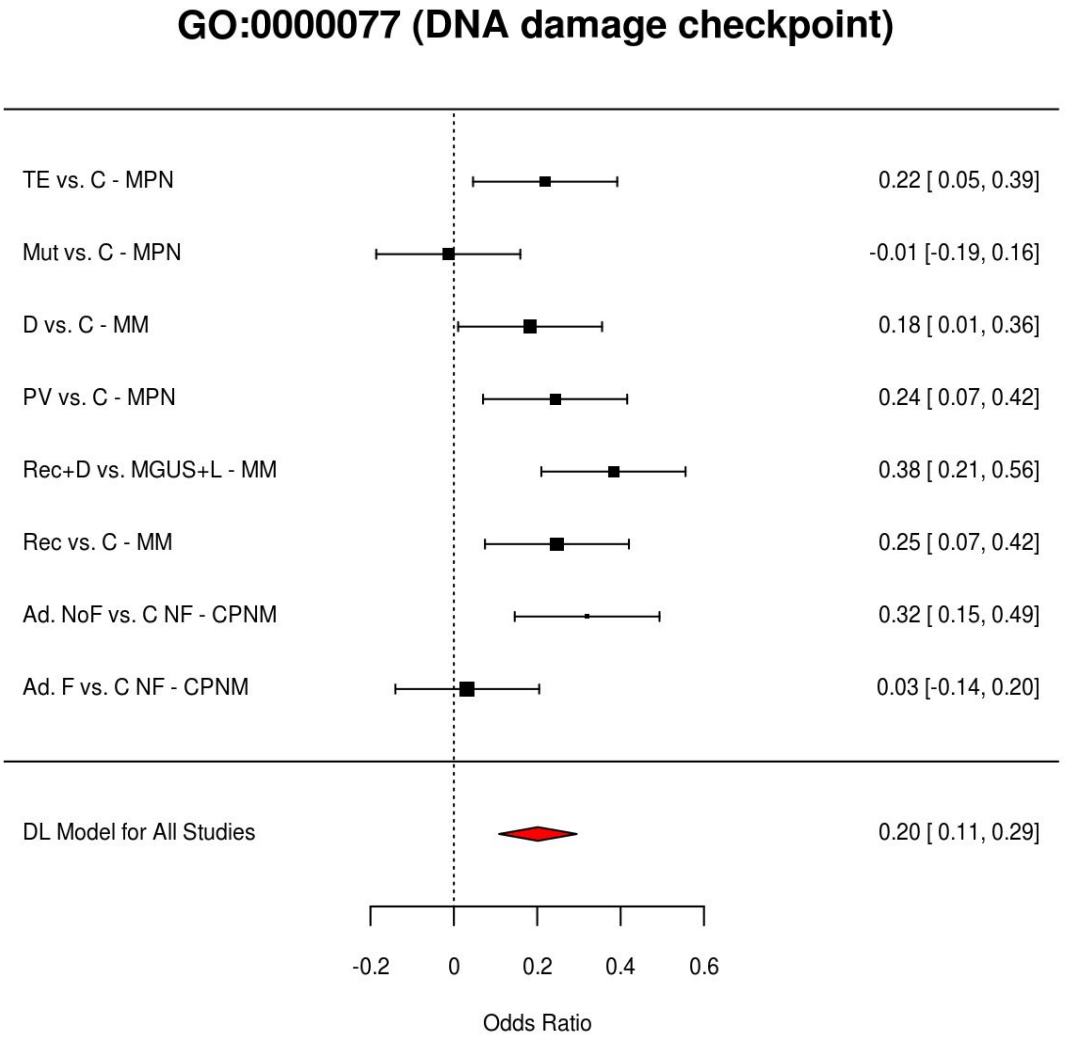

Figura 20. Distribución del efecto para la función GO:000000077.

Por otro lado, tres de las rutas más sobrerrepresentadas en los *Controles* (LOR < -0.5) están relacionadas con la regulación de la apoptosis: reconocimiento de células apoptóticas (GO:0043654), proceso apoptótico de células dendríticas (GO:0097048) y regulación del proceso apoptótico de células dendríticas (GO:2000668). Esta desregulación coincide con el proceso oncogénico, donde las células cancerosas no responden a las señales de apoptosis (Hanahan & Weinberg, 2011).

1.4.3. Componentes celulares significativos del metaanálisis funcional

Esta ontología hace referencia a localizaciones y estructuras celulares donde están ocurriendo las funciones o procesos biológicos. Los grupos más sobrerrepresentados en los *Casos* (LOR > 0.5) están relacionados con la replicación de ADN y división celular (Tabla 27). Se incluyen complejos responsables de la replicación de ADN, así como complejos encargados de la condensación y ensamblaje de los cromosomas o su reparto durante la mitosis (cinetocoros).

Tabla 27. Términos GO de los 8 componentes celulares más sobrerrepresentados y estadísticos seleccionados del metaanálisis.

GO C.celular	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
GO:0000940	Región externa del cinetocoro	0.865	8.00E-03	0.232	0.356
GO:0000796	Complejo de condensina	0.741	0.024	0.228	0.298
GO:0017101	Complejo aminoacilo-ARNt sintetasa multienzima	0.739	0.035	0.242	0.394
GO:0031261	Complejo de preiniciación de replicación de ADN	0.733	0.009	0.2	0.143
GO:0000808	Complejo de reconocimiento del origen cromosomal	0.711	0.004	0.179	0.145
GO:0005664	Complejo de reconocimiento de origen nuclear de replicación	0.711	0.004	0.179	0.145
GO:0031298	Complejo de protección de la horquilla de replicación	0.665	0.001	0.141	0
GO:0031465	Complejo de ubiquitina ligasa Cul4B-RING E3	0.625	0.002	0.147	0

Otros componentes de interés, y con valores *LOR* elevados pero inferiores a 0.5, son el complejo BRCA1 (GO:0070531) y complejo TORC1 (GO:0031931). El complejo TORC1 regula el crecimiento celular y la proliferación celular (Bhola *et al.,* 2016). Mientras que el complejo BRCA1 es responsable de la estabilidad genómica, entre sus muchas funciones se incluye la reparación de la cadena doble de ADN, la regulación de la transcripción y el *splicing* de ARNm (Savage & Harkin, 2014). La sobreexpresión de BRCA1 en los *Casos*

podría ser reflejo de la respuesta de la célula tumoral frente a un elevado daño en su DNA.

Entre los procesos sobrerrepresentados en *Controles* se encuentra el complejo fosfoinositol 3-quinasa clase III. Existen 3 clases de fosfoinositol 3-quinasas, cada una de ellas sintetiza un fosfatidilinositol diferente. El complejo clase I tiene potencial cancerígeno conocido. Sin embargo, las clases II y III carecen de este potencial y, además, su expresión se ve menguada en presencia de AKT, activador del complejo de clase I (Vogt *et al.,* 2010). En la Figura 21 se representa el comportamiento de esta función. Aunque el valor de *LOR* varía en función del contraste, la función está mayoritariamente sobrerrepresentada en *Controles*.

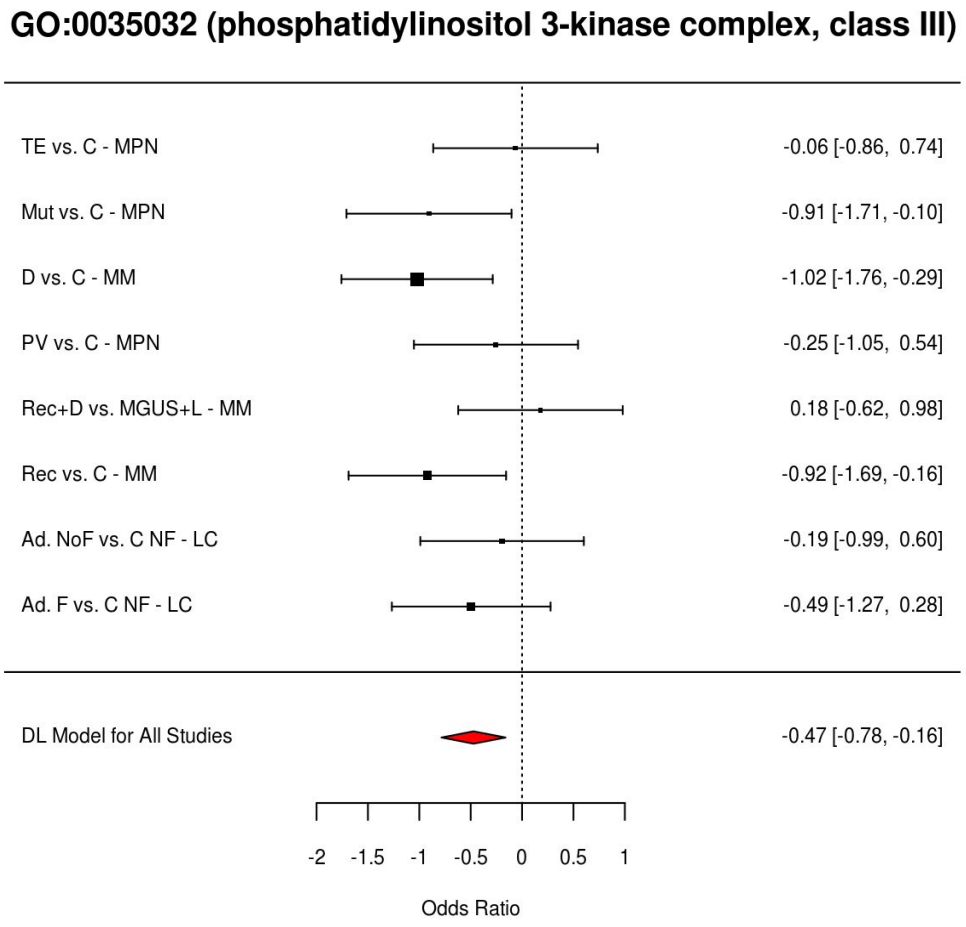

Figura 21. Gráfico de bosque. Distribución del efecto para la función GO:00035032 en el metaanálisis de transcriptómica.

1.4.4. Rutas de señalización significativas del metaanálisis funcional

En el resultado final del metaanálisis tan sólo una ruta se consideraba como sobrerrepresentada y significativa (Tabla 28). Esta ruta es la ruta de biosíntesis de valina, leucina e isoleucina (*hsa:00290*).

El resto de funciones con un LOR elevado, pero inferior a 0.5, son rutas que sintetizan metabolitos útiles para la proliferación celular y, también, funciones responsables de la estabilidad genómica y aparición de cáncer, como son la recombinación homóloga

(hsa00030) (Bishop & Schiestl, 2002) y la recombinación no homóloga (hsa00670) (Davids & Chen, 2013). Como ocurría con BRCA1, la célula podría estar sobreexpresando aquellas rutas encargadas de la reparación del DNA, en busca de corregir las modificaciones causadas por el proceso de oncogénesis.

Tabla 28. Identificadores *KEGG* de las 8 rutas más sobrerrepresentadas y estadísticos seleccionados del metaanálisis de transcriptómica.

Ruta KEGG	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
hsa00290	Biosíntesis de valina, leucina e isoleucina	0.614	3.40E-02	2.14E-01	7.70E-02
hsa03440	Recombinación homóloga	0.469	0.00E+00	6.30E-02	5.00E-03
hsa00670	Metabolismo de carbono - Folato	0.433	1.00E-03	1.00E-01	0.00E+00
hsa01230	Biosíntesis de aminoácidos	0.423	0.00E+00	6.90E-02	2.30E-02
hsa00970	Biosíntesis de aminoacil-ARNt	0.420	2.30E-02	1.34E-01	1.15E-01
hsa01210	Metabolismo del ácido 2-oxocarboxílico	0.414	2.80E-02	1.40E-01	9.10E-02
hsa03450	Recombinación no homóloga	0.391	2.30E-02	1.28E-01	5.60E-02
hsa03460	Ciclo celular	0.318	1.00E-03	7.10E-02	3.20E-02

Otras rutas de interés, pero que no se consideran significativas por tener un *LOR* próximo a 0.2, a pesar de tener un valor de efecto positivo en todos los contrastes, son rutas del metabolismo energético: la ruta de la Glicólisis/Gluconeogénesis (hsa00010), la ruta de las pentosas fosfatos (hsa00030) o las interconversiones de pentosa y gluconato. Las células cancerígenas cambian su metabolismo, haciendo que prevalezca el consumo de glucosa y glutamina así como la glicólisis frente a la respiración oxidativa, con el fin de obtener más metabolitos intermediarios y ATP (Pavlola & Thompson, 2016).

2 Metabolómica

2.1. Procesamiento de los datos

Se realizó un análisis exploratorio de los datos de metabólica para detectar la presencia de posibles anomalías. Con los diagramas de cajas se analizó la distribución de los valores de intensidad. Todas las muestras presentaban la misma distribución, confirmando que la normalización había sido correcta. En la Figura 22 se muestran los resultados del análisis para los datos de LLC.

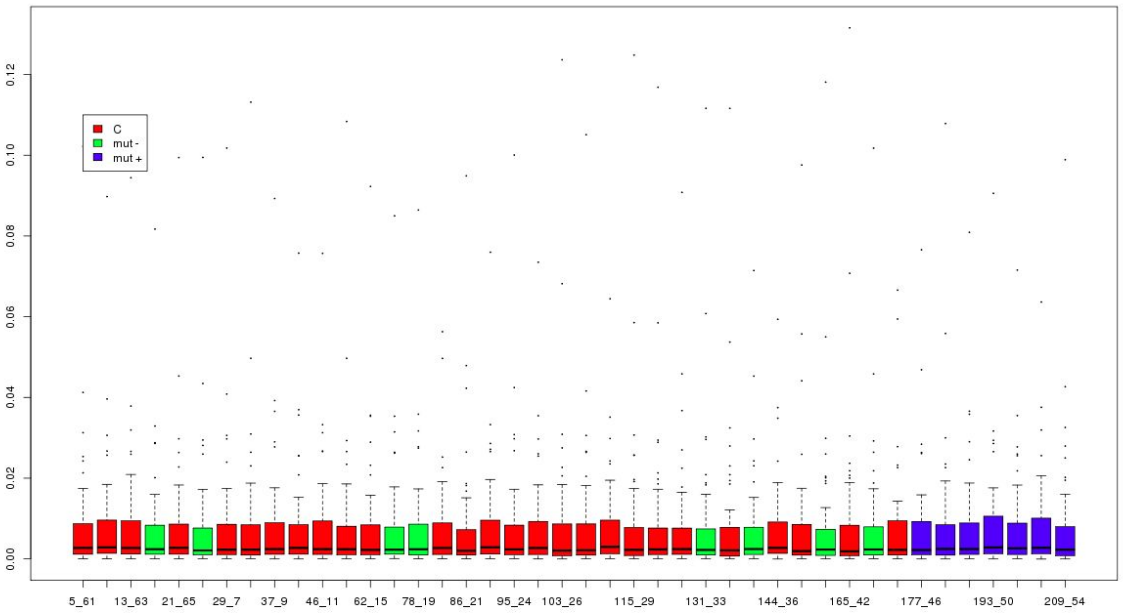

Fig 22. Diagrama de cajas de los datos de LLC de metabolómica.

La mejor forma de detectar variabilidad y agrupaciones entre las muestras de metabolómica es haciendo PCAs. En los *clusters* jerarquizados las agrupaciones no estaban bien definidas. La Figura 23 contiene el PCA de los datos de NMP, aunque la separación de los grupos no es perfecta, se aprecia separación entre los grupos de muestras, en función del estadio de la enfermedad o su condición de enfermedad benigna.

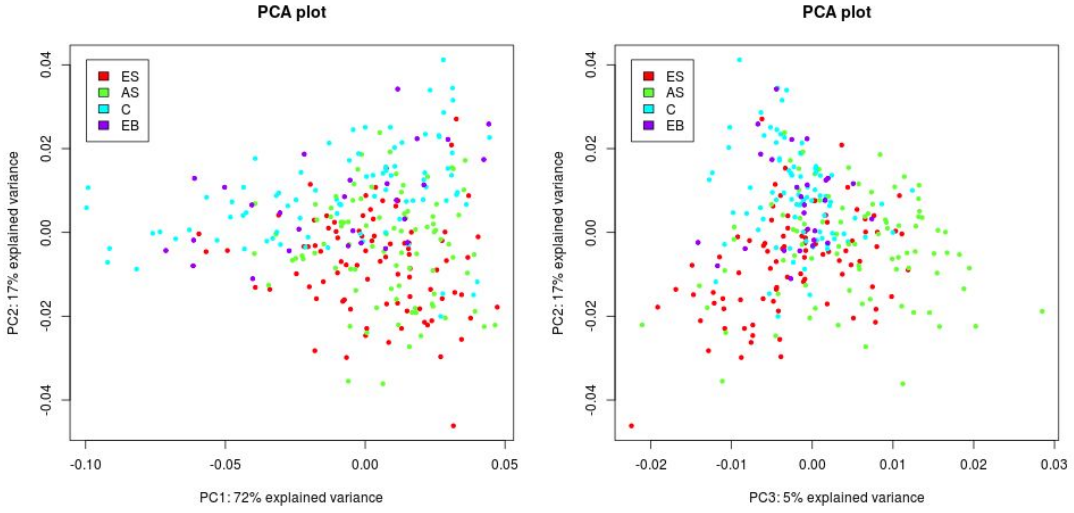

Figura 23. PCA de los datos de NMP de metabolómica.

2.2. Análisis de intensidad diferencial

Se seleccionaron en total 8 contrastes. Previamente se hicieron diferentes contrastes para determinar cuales eran informativos y compatibles con los contrastes de transcriptómica. En la Tabla 29 se encuentran los resultados. Los metabolitos

significativos, tanto infrarrepresentados como sobrerrepresentados, aquellos en los que el valor de *P* era inferior a 0.05.

Con los datos de CP y descriptores disponibles se podrían realizar además otros contrastes. Sin embargo, se observó que tanto entre distintos subtipos histológicos como en función del hábito tabáquico de los pacientes, no había una gran variabilidad en la expresión. La comparación entre pacientes de CP en *Fase Inicial* y *Avanzada* de la enfermedad frente a controles e individuos con enfermedades respiratorias benignas sí arrojaba resultados discriminantes.
Los datos de LLC que se usaron para el metaanálisis procedían de una única comparación: todos los pacientes, independientemente de que tuvieran *IGHV* mutado o no, frente a los controles; el resto de comparaciones no dio resultados concluyentes.
Los pacientes agrupados bajo el término *Diagnosis* fueron comparados con los *Controles* de MM. También se compararon con los individuos en *Remisión*, antiguos pacientes de *Diagnosis*.
Los contraste para las enfermedades agrupadas bajo el marco de NMP realizados fueron comparaciones entre *Policitemia Vera* y *Trombocitemia Esencial* (TE) frente a los controles. También se comparó TE contra trombocitemia secundaria, enfermedad hematológica no cancerosa.

Tabla 29. Nivel de intensidad diferencial entre los grupos experimentales comparados de metabolómica.

Análisis de intensidad diferencial metabolómica				
Estudio	**Comparación**	**Met. infrarrepres.**	**M. sin intensidad diferencial**	**Met. sobrerrepres.**
CP	*FI + FA* vs. *Control*	17	12	15
CP	*FI + FA* vs. *EB*	11	23	10
LLC	*IGHV- + IGHV+* vs. *Control*	0	44	0
MM	*Diagnosis* vs. *Remisión*	0	44	0
MM	*Diagnosis* vs. *Control*	7	33	4
NMP	*PV* vs. *Control*	0	43	1
NMP	*TE* vs. *Control*	5	30	9
NMP	*TE* vs. *TS*	3	34	7

A diferencia de los resultados transcriptómicos, analizando los resultados no se detecta un comportamiento concreto en la direccionalidad de los cambios de intensidad. El grupo más numeroso son los infrarrepresentados o sobrerrepresentados indistintamente, cambiando en función del contraste y enfermedad.

2.3. Análisis de enriquecimiento de grupos de metabolitos

El análisis de enriquecimiento de grupos de metabolitos se realizó para ocho contrastes, empleando dos listas de anotación. La primera lista contenía términos de la ontología *GO*, se incluyeron procesos biológicos, componentes celulares y, en su mayoría, funciones

moleculares. La segunda lista contenía rutas de señalización *KEGG* asociadas a metabolitos. Se consideraron como significativos aquellos metabolitos con un valor de *P* menor de 0.1 En las Tablas 30 y 31 se muestran los resultados del análisis de enriquecimiento de grupos.

Tabla 30. Distribución de las funciones significativas según el término *GO* en metabolómica

Análisis de Enriquecimiento. Funciones significativas *GO*					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F.sig.Casos***	***F. sig.Controles***
LC	*FI + FA* vs. *Control*	5,228	2	0	2
LC	*FI + FA* vs. *EB*	5,219	11	11	0
LCC	*IGHV- + IGHV+* vs. *Control*	5,044	186	169	17
MM	*Diagnosis* vs. *Remisión*	5,208	22	5	17
MM	*Diagnosis* vs. *Control*	5,225	5	5	0
NMP	*PV* vs. *Control*	5,229	1	1	0
NMP	*TE* vs. *Control*	5,222	8	8	0
NMP	*TE* vs. *TS*	5,224	6	6	0

Tabla 31. Distribución de las funciones significativas según las rutas *KEGG* en metabolómica.

Análisis de Enriquecimiento. Funciones significativas *KEGG*					
Estudio	**Comparación**	***No.Sig***	***Sig***	***F.sig.Casos***	***F.sig.Controles***
LC	*FI + FA* vs. *Control*	247	0	0	0
LC	*FI + FA* vs. *EB*	247	0	0	0
LCC	*IGHV- + IGHV+* vs. *Control*	235	12	10	2
MM	*Diagnosis* vs. *Remisión*	239	8	0	8
MM	*Diagnosis* vs. *Control*	237	10	2	8
NMP	*PV* vs. *Control*	247	0	0	0
NMP	*TE* vs. *Control*	245	2	2	0
NMP	*TE* vs. *TS*	247	0	0	0

Si se comparan estos resultados con los obtenidos por expresión diferencial, se observa que el contraste de LCC presenta grupos diferencialmente representados a pesar de no considerarse ningún metabolito significativo en el análisis de intensidad diferencial. Si no se hubiera realizado el análisis de enriquecimiento de grupos, no habría sido posible detectar que, aunque por sí mismos no presentan intensidad diferencial, si se consideran las agrupaciones de genes si hay variabilidad.

2.4. Metaanálisis

El metaanálisis funcional de metabolómica se hizo siguiendo el modelo de efectos aleatorios DL. En la Tabla 32 se resumen las funciones que el metaanálisis ha identificado como significativas, partiendo de los resultados de enriquecimiento de grupos para valores de *P* inferiores a 0.1.

Tabla 32. Resultado global del metaanálisis funcional de metabolómica aplicando DL.

Anotación	*Casos*	*Control*	*Sig.Casos*	*Sig.Control*	*Sig.LOR.Casos*	*Sig.LOR.Control*
Términos *GO*	979	224	88	0	83	0
KEGG	75	29	15	0	11	0

Se ha identificado un número elevado de grupos de genes sobrerrepresentados en ambos grupos. Sin embargo tan sólo se han detectado funciones significativas en el grupo de Casos. Hay un total de 83 funciones significativas anotadas con los términos *GO* y con un valor de efecto (*LOR*) superior de 0.5. El número de rutas *KEGG* significativas y con un *LOR* superior a 0.5 es de 11.

2.4.1 Procesos significativos del metaanálisis funcional

Tabla 33. Términos GO de las 8 funciones moleculares más sobrerrepresentadas y estadísticos seleccionados del metaanálisis de metabolómica.

Rutas *KEGG*	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
hsa00500	Metabolismo del almidón y sacarosa	1.113	0.021	0.324	0
hsa00524	Biosíntesis de butirosina y neomicina	1.113	0.021	0.324	0
hsa04068	Ruta de señalización FOXO	1.113	0.021	0.324	0
hsa04910	Cascada de señalización de la insulina	1.113	0.021	0.324	0
hsa04932	Esteatosis hepática no alcohólica	1.113	0.021	0.324	0
hsa04933	Productos de Glicación	1.113	0.021	0.324	0
hsa00520	Metabolismo de amino-azúcares y polisacáridos	0.85	0.038	0.263	0
hsa04973	Disgestión y absorción de carbohidratos	0.81	0.047	0.259	0

La Tabla 33 contiene las funciones significativas que tienen un *LOR* superior a 0.5. Se observa que las principales funciones están implicadas en el metabolismo y la producción de ATP. Por ejemplo, la ruta hsa04910 es la cascada de señalización de la insulina. La insulina, promueve la síntesis de lípidos, proteínas y carbohidratos así como absorción de la glucosa del medio.

La ruta de biosíntesis de butirosina y neomicina (hsa00524) también aparece como significativa en los resultados. Según la bibliografía no es esperable que ruta esté sobreexpresan en *Casos.* Al analizar los metabolitos que de la ruta, se observa que, algunos de los metabolitos analizados, forman parte del inicio de la ruta. Esto sugiere que si esta ruta es significativa y se inicia, podría ser para obtener metabolitos secundarios. También es posible que sea un error producto de la anotación, en tal caso debería probarse una nueva metodología de anotación, por ejemplo, utilizar módulos de las rutas de KEGG y no rutas completas.

La ruta de señalización FOXO (hsa04068) es un factor de transcripción que promueve la expresión de genes involucrados con la apoptosis, metabolismo de glucosa, resistencia a estrés oxidativo y la longevidad celular. Este resultado no coincide con otros estudios previos en los que se describe cómo, en condiciones de oncogénesis, FOXO se encuentra infraexpresado, salvo que las células cancerosas requieran su activación para controlar la presencia de especies reactivas de oxígeno (por ejemplo, a consecuencia de tratamientos terapéuticos) (Shukla, 2014).

2.4.2 Rutas de señalización significativas del metaanálisis funcional

Al contrario de lo que ocurría en las rutas *KEGG,* que abarcan las distintas rutas metabólicas de una manera más general, la anotación por términos *GO* ha dado en su mayoría reacciones metabólicas sencillas (Tabla 34). Es decir, la información que obtenemos de *GO* parece estar limitada por el bajo número de metabolitos de anotados.
Todos los estudios de metabolómica contaba con únicamente 44 metabolitos.
Las principales reacciones que ocurren están relacionadas con el metabolismo energético, como la glicólisis/gluconeogénesis o ruta de las pentosas fosfato. Algunas de estas reacciones sobrerrepresentadas son la actividad glucosa-6-fosfatasa (GO:0004346), actividad oligo-1,6-glucosidasa (GO:0004574) o la actividad xilosa isomerasa (GO:0009045). También hay rutas implicadas en la biosíntesis de moléculas, como por ejemplo la actividad trimetilamina N-óxido reductasa (GO:0009033) que produce un precursor de la síntesis de colina. La colina es un precursor de los fosfolípidos de membrana de las células tumorales (Bañez-Coronel et al., 2008)

Tabla 34. Términos GO de los 8 procesos más sobrerrepresentados y estadísticos seleccionados del metaanálisis de metabolómica.

Términos *GO*	Descripción	*LOR medio*	*P-ajustado*	*SE*	*tau2*
GO:0003674	Funciones elementales moleculares (catálisis, unión de proteínas...)	1.216	0.051	0.363	0.709
GO:0003824	Actividad catalítica	1.216	0.051	0.363	0.709
GO:0009033	Actividad trimetilamina N-óxido reductasa	1.122	0.040	0.324	0.000
GO:0016657	Actividad oxidorreductasa, nitrógeno aceptor	1.122	0.040	0.324	0.000
GO:0034899	Actividad trimetilamina	1.122	0.040	0.324	0.000

	monooxigenasa				
GO:0050352	Actividad trimetilamina-óxido aldolasa	1.122	0.040	0.324	0.000
GO:0000016	Actividad lactasa	1.113	0.040	0.324	0.000
GO:0004133	Actividad de la enzima desramificadora de glucógeno	1.113	0.040	0.324	0.000

3 Integración

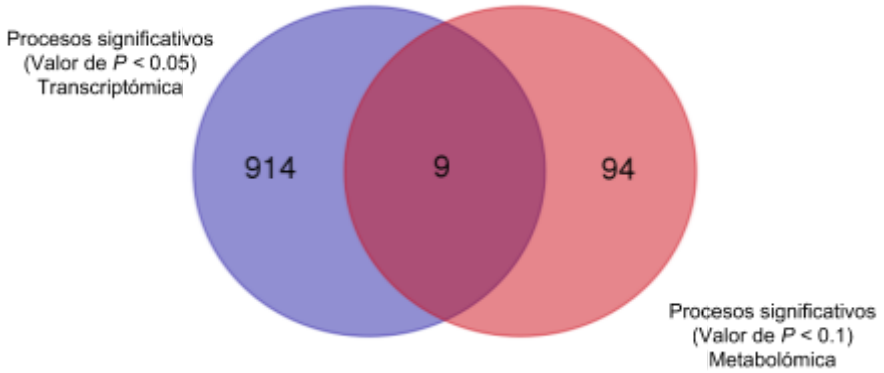

Figura 24. Diagrama de Venn. Intersección de los procesos significativos en transcriptómica y metabolómica.

El resultado de la unión de todas las rutas significativas de transcriptómica, a partir de los términos *GO* (procesos biológicos, funciones moleculares y componentes celulares) es de 923 procesos. Por su parte, los resultados de metabolómica proporcionan un total 103 procesos significativos.

La discrepancia entre los valores se debe a las características propias de las dos ómicas utilizadas. En metabolómica, en función de la técnica análitica, se pueden distinguir más de 200 metabolitos. Sin embargo, distinguir esos 200 metabolitos en una misma muestra por RMN es imposible. El solapamiento entre los picos y la elevada intensidad de los metabolitos más concentrado, hace que se pierda información de parte de los metabolitos en la muestra. Por este motivo, los datos utilizados de metabolómica constan únicamente 44 metabolitos, frente a los 12745 genes por estudio identificados en transcriptómica. Esta diferencia en el número de datos de partida causa la divergencia en los resultados, ya que, como es lógico se podrán reconocer más procesos a más información se tenga.

No se debe olvidar que la transcriptómica y la metabolómica dan información en dos regiones completamente diferentes del proceso fisiológico pero, aún así, estos cambios estarán siempre relacionados. Por ejemplo, en ambos casos los procesos mayoritariamente diferenciados son del metabolismo energético además de otros relacionados con la síntesis de biomoléculas, replicación de ADN y división celular.

Si se analizan los dos listados de procesos significativos en busca de solapamientos, obtenemos un total de 9 procesos comunes (Fig. 24). Tres de ellos corresponden a

funciones moleculares según los términos *GO*. Las restantes, son rutas recogidas en la base de datos KEGG. Tanto las rutas comunes, entre los dos metaanálisis ómicos, como las específicas de cada uno de ellos, proporcionan una información de interés y complementaria sobre todas las funciones significativas identificadas por los metaanálisis. Habrá rutas que desde un punto de vista metabolómico se considerarán activas, y otras que no, por ejemplo, por tratarse de una cascada de señalización no detectable a través de los metabolitos.

Tabla 35. Procesos biológicos comunes en ambos metaanálisis

ID	Tipo de proceso	Descripción	Ómica	*LOR medio*	*P-ajustado*	*SE*	*tau2*
GO:0016853	*GO* F. molecular	Actividad isomerasa	T	0.1340	0.0220	0.0410	0.0040
			M	0.3960	0.0890	0.1250	0.0000
GO:0016874	*GO* F. molecular	Actividad ligasa	T	0.2400	0.0400	0.0800	0.0440
			M	0.4130	0.0400	0.1200	0.0000
GO:0046527	*GO* F. molecular	Actividad glucosiltransferasa	T	0.4460	0.0010	0.0980	0.0000
			M	1.1130	0.0400	0.3240	0.0000
hsa00010	*KEGG*	Glicólisis Gluconeogénesis	T	0.1940	0.0010	0.0460	0.0000
			M	0.5660	0.0800	0.1980	0.0000
hsa00970	*KEGG*	Biosíntesis de aminoacil-tRNA	T	0.4200	0.0230	0.1340	0.1150
			M	0.3600	0.0770	0.1230	0.0000
hsa01100	*KEGG*	Rutas metabólicas	T	0.1420	0.0230	0.0460	0.0160
			M	0.6800	0.0770	0.2330	0.2210
hsa01210	*KEGG*	Metabolismo del ácido 2-oxocarboxílico	T	0.4140	0.0280	0.1400	0.0910
			M	0.3880	0.0990	0.1410	0.0000
hsa01230	*KEGG*	Biosíntesis de aminoácidos	T	0.4230	0.0000	0.0690	0.0230
			M	0.3480	0.0770	0.1200	0.0000
hsa05230	*KEGG*	Metabolismo de carbono del cáncer	T	0.1490	0.0150	0.0450	0.0000
			M	0.5530	0.0210	0.1570	0.0670

Tres de las rutas comunes están implicadas en el metabolismo energético (Tabla 35). Este resultado común entre ambas ómicas es posible gracias a que, mientras que en transcriptómica se detectan los ARNms de las enzimas participantes, en los datos de metabolómica se recogen la mayoría de metabolitos que participan. Lo mismo ocurre con la biosíntesis de aminoácidos (hsa01230), en los datos de metabolómica se incluye información de varios de ellos.

Se están detectando, a través de la integración, únicamente aquellas rutas de las que se dispone de datos procedentes de ambas ómicas. Es posible, que para aumentar el número de rutas comunes, sea necesario incrementar el número de metabolitos de los datos utilizados.

Figura 25. Diagrama de Venn. Intersección de las procesos significativos en transcriptómica y metabolómica.

Es necesario considerar también aquellas rutas que son significativas en una ómica pero no en otra (Fig. 25). La explicación de porqué ciertas rutas aparecen en transcriptómica y no en metabolómica es sencilla. Es probable que haya ciertos metabolitos que no se estén detectado por metabolómica por una limitación de la técnica analítica. Con RMN no es posible identificar metabolitos que solapan sus señales con otros, como tampoco se podrán detectar aquellos que se encuentran en concentraciones muy bajas. Otro motivo por el que la congruencia entre ambos no es total, es porque hay ciertas rutas que, para saber si están activadas, se depende de la sobreexpresión de transcritos, y no de metabolitos detectables por metabolómica.

La baja resolución hace que se pierda información sobre algunos metabolitos que pueden ser de interés. Esto podría ser lo que ocurre con la ruta de las pentosas fosfatos (hsa00030), gracias a la bibliografía sabemos que esta ruta está sobreexpresada y los datos de transcriptómica lo confirman. Sin embargo, los datos de metabolómica no aportan información sobre la mayoría de los metabolitos que forman parte de esta ruta, y por ende la ruta no se considera sobreexpresada en metabolómica. Lo mismo sugiere la ausencia de la ruta de biosíntesis de aminoácidos ramificados (hsa00290) en los resultados de metabolómica. A pesar que contamos con información de la intensidad de valina, leucina e isoleucina, falta información de los otros 20 metabolitos de la ruta, y por lo tanto, este podría ser la causa por la que no se considera como activa en el metaanálisis de metabolómica.

Al contar con muchos más genes que metabolitos como datos de partida, se obtenien más funciones significativas de transcriptómica. Los resultados sugieren que, en la integración, el factor limitante ha sido el escaso número de metabolitos identificados.

V. Conclusiones

1. Las técnicas de metaanálisis a nivel de función son una metodología adecuada para el análisis global de datos procedentes de estudios metabolómicos y transcriptómicos.

Hemos obtenido como resultado un claro desajuste en el metabolismo energético y síntesis de aminoácidos en los grupos Control. Estos cambios concuerdan con lo descrito anteriormente en estudios con células cancerosas.

2. La integración de los resultados del metaanálisis funcional proporciona una nueva perspectiva sobre la cáncer. La combinación de los resultados nos permite ver cómo los cambios a nivel de transcriptómica, por ejemplo la sobreexpresión de ciertas enzimas, se traducen en variaciones en los metabolitos con los que interacciona, cambios detectables por metabolómica.

3. A pesar de ello, existen ciertas limitaciones en los resultados obtenidos en el área de metabolómica. Probablemente, el limitado número de metabolitos de los que disponemos datos en estos estudios unido a métodos de anotación muy generales, dificulta la identificación de alteraciones relevantes mediante esta técnica.

4. Incluso con las limitaciones antes planteadas, el perfeccionamiento de este método puede ser relevante para el estudio de otras enfermedades así como también ampliar el conocimiento sobre el cáncer.

5. La perspectiva global de la integración proporciona una mejor caracterización molecular del cáncer y, tras su perfeccionamiento, facilitará el diseño de nuevos abordajes clínicos en el estudio y tratamiento de estas enfermedades.

VI. Perspectivas futuras

- Una posible vía de trabajo futuro es poner a prueba otras estrategias existentes. Se podría replicar el metaanálisis y posterior integración utilizando otras estrategias similares a las incluidas en la introducción.

- Un hecho a considerar es el equilibrio en el origen de las muestras, que proceden de tumores hematológicos y tisulares. A la hora de realizar la integración, hay dos enfoques que se deberían intentar para perfeccionar el método de integración. La primera opción es centrarse en un único tipo de cáncer. Descartar los datos de CP y utilizar únicamente las muestras procedentes de cáncer hematológico. La segunda opción sería incluir nuevas muestras procedentes de tumores hematológicos y así equiparar el número de muestras de ambos grupos.

- También se podría repetir el estudio, aprovechando los estudios de transcriptómica que se han seleccionado, pero realizar la integración de datos de manera pareada. Es decir, centrarnos en la integración de los datos procedentes de un sólo tipo de enfermedad.

- Por último, se podría probar un nuevo enfoque a la hora de anotar los grupos de genes y metabolitos. Utilizar información sobre los módulos de KEGG y no en rutas completas, así no se pierde información por falta de datos en ciertos metabolitos. Al centrarse en los módulos, se podría detectar de manera más específica las reacciones que están teniendo lugar y poder valorar su efecto en la ruta global.

VII. Anexos

A.1. Evaluación de los diferentes métodos de metaanálisis

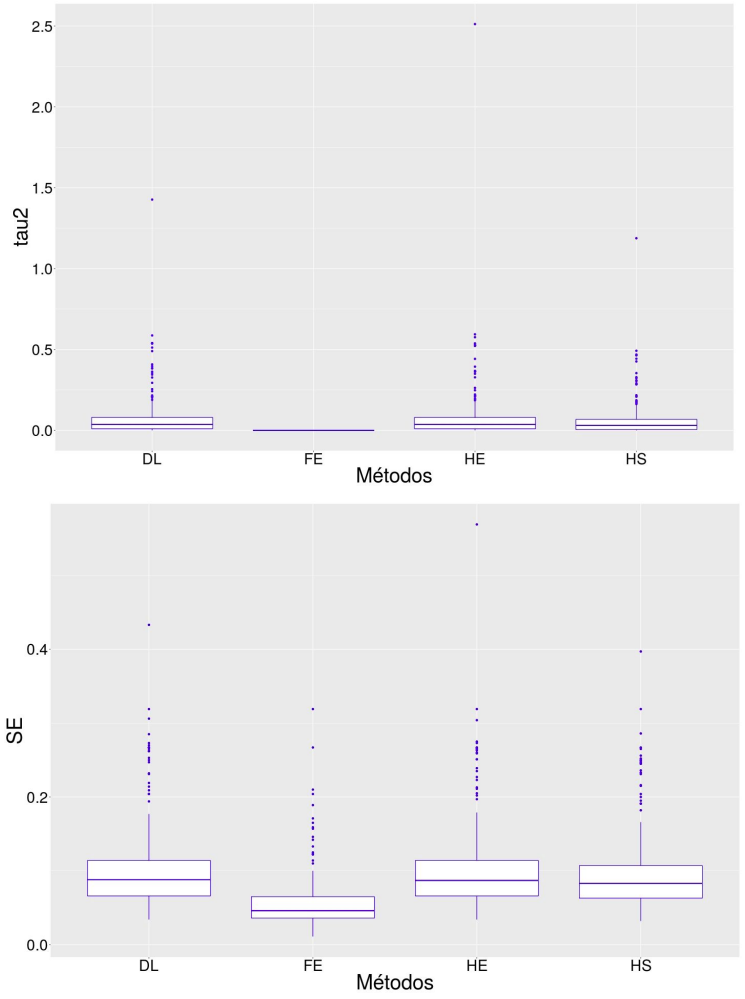

Figura 1A. Análisis global de la heterogeneidad. Distribución de (A) *tau2* (A) y *(B)* SE por métodos de estimación del efecto (*LOR*).

En el apartado de metaanálisis no se hace un único metaanálisis, sino uno por cada función o proceso de interés y modelo de efectos seleccionado. Dado que se puede aplicar sobre los datos diferentes modelos de efectos aleatorios, es necesario asegurarse que el abordaje elegido se adecúa a los datos con los que trabajamos. En este trabajo se seleccionó el modelo de efectos aleatorios DL, por ajustarse mejor a la variabilidad existente en estudios ómicos. En caso de desconocer el modelo idóneo para los datos se debe repetir el metaanálisis para cada uno de ellos y valorar, estadísticamente, la mejor opción.

La representación gráfica de los estimadores estadísticos obtenidos en el metaanálisis, de cada función o ruta de señalización, permite valorar los resultados de cada método de efectos seleccionado. En las figuras superiores (Fig. 1A-A y 1A-B) se representan en diagramas de cajas los valores de *tau2* y de *SE* de todas las rutas *KEGG* analizadas. En

ellas, se observa que al utilizar un método de estimación de efectos fijos (FE), la heterogeneidad es nula, sin embargo, sí se detecta desviación típica entre las muestras. Confirmando que el modelo de efectos correcto es uno aleatorio. Con ayuda de estas gráficas, y otras que incluyan H2 y QEp por ejemplo, es posible seleccionar el método de ajuste que mejor se adapta a los datos. En este estudio, los tres modelos aleatorios puestos a prueba daban resultados muy similares.

Los análisis de sensibilidad se basan en la repetición del metaanálisis, pero eliminando en cada repetición una de los contrastes. En la gráfica inferior, se representa la sensibilidad de los diferentes modelos de efectos (Fig. 2A). El error estándar se ha calculado a partir del valor del efecto de cada función en las sucesivas repeticiones. Se observa que el impacto de cada modelo de estimación sobre los resultados de sensibilidad son bastante similares, indicando que los tres métodos son bastante robustos.

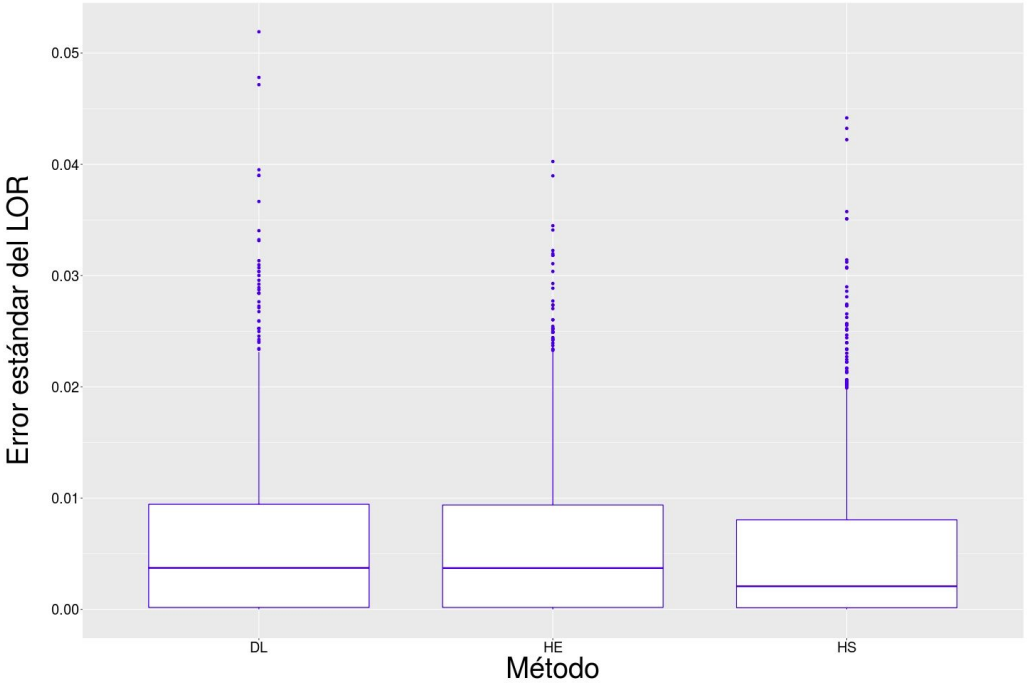

Figura 2A. Distribución de la sensibilidad de los métodos de estimación de efectos al calcular procesos biológicos .

A.2. Análisis de estudios influyentes

Este tipo de análisis es útil para identificar aquellos estudios cuya magnitud y variabilidad es muy diferente al resto teniendo por tanto una fuerte influencia sobre los resultados del metaanálisis. La función *influence* de metafor aplica modelos de regresión adaptados a los estudios de metaanálisis que permiten visualizar de forma gráfica el efecto de cada estudio en el resto.

En la Figura 3A. se representa gráficamente el resultado del análisis de los estudios influyentes de la función GO:0003887 con los datos de transcriptómica. En las gráficas se analizan estadísticos como: *rstudent* (residuos estandarizados obtenidos con al eliminar un estudio), distancias de Cook (distancia esperada con un estudio eliminado y sin

eliminar), covarianza de ratios (de los efectos observados al eliminar un estudio) o valores tau2. De esta forma se puede analizar de manera rápida y visual el efecto de los contrastes sobre la función de interés. Por ejemplo, la sexta gráfica representa el peso de ese contraste en el estudio final. Se observa que el contraste número 2, aunque tiene un peso similar al resto, es ligeramente diferente afectando al resto de valores.

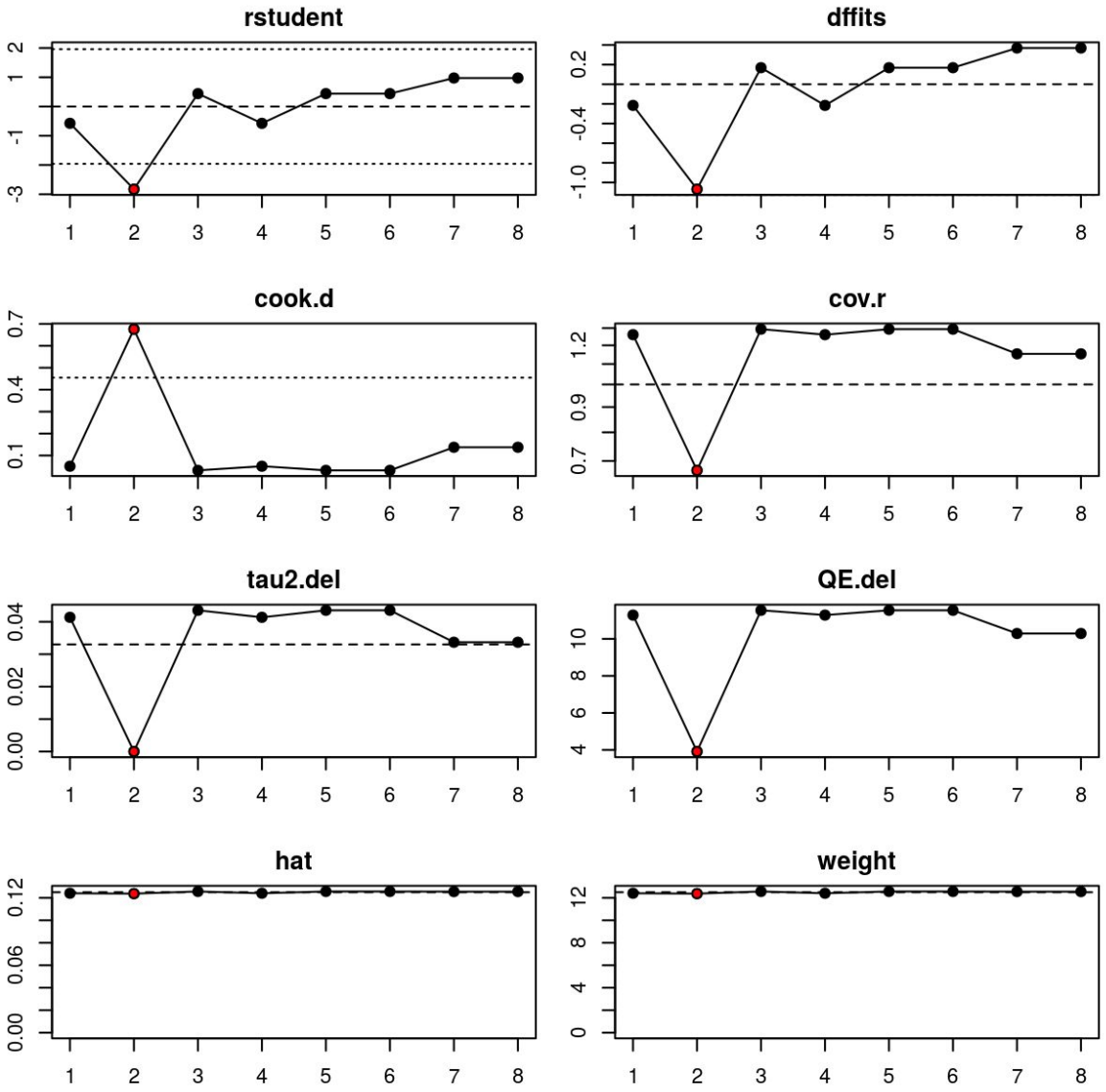

Figura 3A. Análisis de estudios influyentes para la función GO:0003887 aplicando un modelo de efectos DL.

A.3. Análisis de sensibilidad por estudios

Un análisis de sensibilidad consiste en eliminar uno de los estudios del metaanálisis y repetirlo. Utilizando la función *rma.uni* se realiza el metaanálisis, indicando el modelo de efectos deseado, y a continuación se utiliza la función *leave1out* de metafor (Viechtbauer, 2010). Esta función ajusta el modelo pero eliminando, cada una de las veces, un estudio diferente. Si los resultados obtenidos son consistentes con el resultado global, se está

garantizando la robustez del metaanálisis. Además, este método nos permite determinar la influencia que tiene cada uno de los estudios en el análisis final.

Tabla 1A. Indicadores del análisis de sensibilidad de los estudios para la función GO:0000150 aplicando un modelo de efectos DL en datos de transcriptómica.

Estudios	*LOR*	*SE*	*ZVAL*	*P-valor*
TE vs. Control	0.481	0.129	3.742	0
JAK- vs. JAK+	0.498	0.13	3.815	0
Diagnosis vs. Control	0.547	0.128	4.284	0
PV vs. Control	0.481	0.129	3.742	0
Diag + Re vs. MGUS + Latente	0.547	0.128	4.284	0
Recaída vs. Control	0.547	0.128	4.284	0
Adeno No fumador vs. No fumador	0.439	0.129	3.389	0.001
Adeno Fumador vs. No fumador	0.439	0.129	3.389	0.001

Estudios	*LI 95%*	*LS 95%*	*Q*	*Qp*	*tau2*	*I2*	*H2*
TE vs. Control	0.229	0.733	6.048	0.418	0.001	0.795	1.008
JAK- vs. JAK+	0.242	0.754	6.181	0.403	0.003	2.933	1.03
Diagnosis vs. Control	0.297	0.797	4.91	0.555	0	0	1
PV vs. Control	0.229	0.733	6.048	0.418	0.001	0.795	1.008
Diag + Re vs. MGUS + Latente	0.297	0.797	4.91	0.555	0	0	1
Recaída vs. Control	0.297	0.797	4.91	0.555	0	0	1
Adeno No fumador vs. No fumador	0.185	0.692	4.684	0.585	0	0	1
Adeno Fumador vs. No fumador	0.185	0.692	4.684	0.585	0	0	1

Los valores de la Tabla 1A son:

- LOR, coeficiente estimativo del efecto del modelo.
- SE, error estándar del coeficiente.
- ZVAL, test estadístico de los coeficientes.
- P-valor, valor de P del test estadístico.
- LI 95%, límite inferior del intervalo de confianza de los coeficientes.
- LF 95%, límite superior del intervalo de confianza de los coeficientes.

- Q, test estadística de la heterogeneidad de heterogeneidad.
- Qp, valor de P de los tests de heterogeneidad.
- tau2, indicador de heterogeneidad entre los estudios.
- I2, indicador de la heterogeneidad de los estudios.

En la tabla superior se represe el análisis de sensibilidad para la función GO:0000150. Se observa como, al eliminar los estudios indicados en la primera columna, el valor de los cálculos varían. En general, el rango de valores es similar para todos los estudios, aunque sí que se observa, con el indicador *I2*, que ciertos estudio generan heterogeneidad.

A.4. Scripts utilizados y resultados gráficos obtenidos

Los scripts utilizados y la información obtenida se encuentran accesibles en el siguiente link: https://goo.gl/JmyEHs.

A.5. Tiempo de procesado de los scripts de metaanálisis

Se crearon scripts en *R* para la obtención de los datos de transcriptómica a partir de la información disponible en *GEO*. También se usaron scripts para preparar los datos de metabolómica para su procesado en *R*.
Los scripts utilizados para el metaanálisis fueron diseñados por García-García y, ligeramente modificados, para adaptarlos al proceso de análisis de cada ómica.
En la Tabla 2A se indica el tiempo de procesado de cada uno de los scripts utilizados para el metaanálisis. El ordenador utilizado fue un ordenador portátil Lenovo Ideapad Y50-70 (coste aproximado 1,000 €). Los scripts fueron lanzados en un núcleo de los cuatro disponibles en el procesador Intel Core i7-4710HQ 2.50GHz del ordenador.

Tabla 2A. Tiempo de cálculo de los scripts de metaanálisis.

Tiempo de cálculo de los análisis		
Script	**Transcriptómica**	**Metabolómica**
s010_exploratory_analysis.r	20.240 s.	2.293 s.
s020_differential_expression.r	3.612 s.	0.949 s.
s030_anotacion_grupos.r	974.352 s.	-
s030_gsa_go.r	4,052.811 s.	62,151 s.
s030_gsa_kegg.r	300.723 s.	5.935 s.
s040_functional_meta_analysis_go.r	-	721.779 s.
s040_functional_meta_analysis_go_MF.r	638.7945 s.	-
s040_functional_meta_analysis_go_BP.r	1,358.0259 s.	-
s040_functional_meta_analysis_go_CC.r	273.008 s.	-
s040_functional_meta_analysis_kegg.r	70.604 s	53.525 s.

VIII. Bibliografía

Abcam (2017). Antibody structure and isotypes. Guide to the structural components that make up an antibody: heavy chains, light chains, F(ab)/Fc regions and isotypes [último acceso el 05/07/17]. Accesible en: http://www.abcam.com/protocols/antibody-structure-and-isotypes

AECC (2017). Asociación española contra el cáncer: Tipos de Cáncer por localización. [último acceso el 05/07/17]. Accesible en: https://www.aecc.es/SobreElCancer/CancerPorLocalizacion/Paginas/Lozalizaci%C3%B3ndelc%C3%A1ncer.aspx.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez, Banet, J., Billis, K., García, Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F.J., Murphy, D.N., Nag, R., Ruffier, M., Schuster, M., Tang, Y.A., Vogel, J.H., White, S., Zadissa, A., Flicek, P., Searle, S.M. (2016). The Ensembl gene annotation system. *Database (Oxford),* 23;2016.

Affymetrix (2017). Affymetrix Data Sheet: GeneChip® Human Genome Arrays [último acceso el 05/07/17]. Accesible en: http://media.affymetrix.com/support/technical/datasheets/human_datasheet.pdf.

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

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, AP., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25-9.

Ayala, G. (2017). Capítulo 4: Microarrays. Bioinformática Estadística: Análisis Estadístico de Datos Ómicos con R/Bioconductor. Accesible en: http://www.uv.es/ayala/docencia/tami/tami13.pdf.

Bañez-Coronel, M., Ramírez de Molina, A., Rodríguez-González, A., Sarmentero, J.,

Ramos, M. A., García-Cabezas, M. A., García-Oroz, L., & Lacal, J. C. (2008). Choline kinase alpha depletion selectively kills tumoral cells. *Curr Cancer Drug Targets*, 8(8): 709–19.

Barrett, T. (2013). Gene Expression Omnibus (GEO). The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Accesible en: https://www.ncbi.nlm.nih.gov/books/NBK159736/.

Beckonert, O., Keun, H.C., Ebbels, T.M., Bundy, J., Holmes, E., Lindon, J.C., Nicholson, J.K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc,* 2(11):2692-703.

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

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W. (2013). GenBan. *Nucleic Acids Res,* 41(Database issue):D36-42.

Bhola, N.E., Jansen, V.M., Koch, JP., Li, H., Formisano, L., Williams, J.A., Grandis, J.R., Arteaga, C.L.

Bishop, A.J.R., Schiestl, R.H. (2002). Homologous recombination and its role in carcinogenesis. J *Biomed Biotechnol*, 2(2): 75–85.

Bonneau, E., Tétreault, N., Robitaille, R., Boucher, A., De Guire, V. (2016). Metabolomics: Perspectives on potential biomarkers in organ transplantation and immunosuppressant toxicity. *Clin Biochem*, 49(4-5):377-84.

Braun, R.D, Bissig, D, North, R., Vistisen, K.S., Berkowitz, B.A.(2012). Human Tumor Cell Proliferation Evaluated Using Manganese-Enhanced MRI. *PLos One,* 7(2): e30572.

Brown, T.A. (2002). Chapter 3, Transcriptomes and Proteomes. Genomes. 2nd edition. Oxford: Wiley-Liss.. Accesible en: https://www.ncbi.nlm.nih.gov/books/NBK21121/.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 23;455(7216):1061-8.

Cañizares, J. (2010). BioMur modulo3 v1.1 documentation. Microarrays para el análisis de expresión [último acceso el 05/07/17]. Accesible en: http://personales.upv.es/jcanizar/modulo_3/aspectos_tecnicos_1.html.

Carreras (2017). Fundación Josep Carreras contra la Leucemia: Tipos de enfermedades hematológicas. [último acceso el 05/07/17]. Accesible en: http://www.fcarreras.org/es/tipos-de-enfermedades-hematologicas_357013.

Carvalho, B.S., Irizarry, R.A. (2010). A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics*, 26(19):2363-7.

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

Cavill, R., Jennen, D., Kleinjans, J., Briedé, J.J. (2016). Transcriptomic and metabolomic data integration. *Brief Bioinform*, 17(5):891-901.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A.P., Sander, C., Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*, 2(5):401-4.

Corre, J., Munshi, N., Avet-Loiseau, H. (2015). Genetics of multiple myeloma: another heterogeneity level? *Blood*, *125*(12), 1870–1876.

DerSimonian, R. Laird, N., 1986. Meta-analysis in clinical trials. *Control clin trials,* 7(3):177–188.

Dilley, R.L., Verma, P., Cho, N.W., Winters, H.D., Wondisford, A.R., Greenberg, R.A. (2016). Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature*, 3;359(7627):54-58.

Dumas, M.E. (2012). Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. *Mol Biosyst,* 8(10):2494-502.

Eales, K.L, Hollinshead, K.E.R., Tennant, D.A. (2016). Hypoxia and metabolic adaptation of cancer cells. *Oncogenesis*, 5(1):e190.

Ebbels, T., Cavill, R. (2009). Bioinformatic methods in NMR-based metabolic profiling. *Prog* Nucl Magn Reson Spectrosc, 55(4):361-374.

Edgar, R., Domrachev, M., Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207-10

Elfilali, A., Lair, S., Verbeke, C., La Rosa, P., Radvanyi, F., Barillot, E. (2006). ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. *Nucleic Acids Res,* 34:D613-6.

Emwas, A.H. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. En Metabomics: Methods and protocols pp. 161-93. New York, USA. Springer New York.

Evans, T.G. (2015). Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. J Exp Biol, 218:1925-35.

Fonseca, R., Bergsagel, P. L., Drach, J., Shaughnessy, J., Gutierrez, N., Stewart, A. K., Morgan, G., Van Ness, B., Chesi, M., Minvielle, S., Neri, A., Barlogie, B., Kuehl, W. M., Liebisch, P., Davies, F., Chen-Kiang, S., Durie, B. G. M., Carrasco, R., Sezer, O., Reiman, T., Pilarski, L., Avet-Loiseau, H. (2009). International Myeloma Working Group molecular classification of multiple myeloma: spotlight review. Leukemia, 23(12): 2210–21.

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

García-Alcalde, F., García-López, F., Dopazo, J., Conesa, A. (2011). Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 1;27(1): 137–139

García-García, F. (2016). Métodos de análisis de enriquecimiento funcional en estudios genómicos. Tesis Doctoral. Universidad de Valencia, Valencia (España).

Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315.

Geistlinger L, Csaba G and Zimmer R (2016). Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics*, **17:**45.

Gene Expression Omnibus (GEO). 2013 May 19. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK159736/.

GLOBOCAN (2012). GLOBOCAN: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. ARCI: OMS; [último acceso el 05/07/17]. Accesible en: http://globocan.iarc.fr/Default.aspx.

Gómez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., Tegner, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst Biol*, 8(2):I1.

Gupta, S., Chawla, K. (2013). Oncometabolomics in cancer research. *Expert Rev Proteomics,* 10(4):325-36.

Hanahan, D., Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell,* 4;144(5):646-74.

Hasty, P., Montagna, C. (2014). Chromosomal rearrangements in cancer: Detection and potential causal mechanisms. *Mol Cell Oncol*, 1(1):e29904.

Haug, K., Salek, R.M., Conesa, P., Hastings, J., de, Matos, P., Rijnbeek, M., Mahendrake, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S.A., Griffin, JL., Steinbeck, C. (2013). MetaboLights -- an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41:D781-6.

Hedges, L.V., Gurevitch, J.,Curtis, P.S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, 80:1150-1156.

Horgan, R.P, Kenny, L.C. (2011). SAC review 'Omic' technologies: genomics, transcriptomics,proteomics and metabolomics. The Obstetrician & Gynaecologist, 13:189–195.

Huang, S., Chaudhary, K., Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet*, 8:84.

Hubbell, E., Liu, W.M., Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics,* 18(12):1585-92.

International Cancer Genome Consortium (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249-64.

Jantus-Lewintre, E., Usó, M., Sanmartín, E., Camps, C. (2012). Update on biomarkers for the detection of lung cancer. Lung Cancer (Auckl), 11(3),21-29.

Kamburov, A., Cavill, R., Ebbels, T.M., Herwig, R., Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917-8.

Kamburov, A., Wierling, C., Lehrach, H., Herwig, R. (2009). ConsensusPathDB -- a database for integrating human functional interaction networks. *Nucleic Acids Res,* 37:D623-8.

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

Kenfield, S. A., Wei, E. K., Stampfer, M. J., Rosner, B. A., Colditz, G. A. (2008). Comparison of Aspects of Smoking Among Four Histologic Types of Lung Cancer. *Tob Control*, *17*(3):198–204.

Kutmon, M., van Iersel, M.P., Bohler, A., Kelder, T., Nunes, N., Pico, A.R., Evelo, C.T. (2015). PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLoS Computational Biology,* 11(2): e1004085

Kyle, R. A., Buadi, F. I., Rajkumar S. V. E. (2011). Management of Monoclonal Gammopathy of Undetermined Significance (MGUS) and Smoldering Multiple Myeloma (SMM). *Oncology (Williston Park, N.Y.)*, *25*(7):578–586.

Landgren, O., Kyle, R. A., Rajkumar, S. V. (2011). From myeloma precursor disease to multiple myeloma: new diagnostic concepts and opportunities for early intervention. *Clin Cancer Res*, 17(6):1243–52.

Law, V., Knox, C., Djoumbou, Y., Jewison, T.,Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res,* 1;42(1):D1091-7. 24203711

Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., Storey, J.D., Zhang, Y., Torres, L.C. (2017). sva: Surrogate Variable Analysis. R package version 3.24.4.

Leinonen, R., Sugawara, H., & on behalf of the International Nucleotide Sequence Database Collaboration, M. (2011). The Sequence Read Archive. *Nucleic Acids Res*, *39*:D19–D21.

Ligtenberg W (2017). *reactome.db: A set of annotation maps for reactome*. R package version 1.59.1.

Tenenbaum D (2017). *KEGGREST: Client-side REST access to KEGG*. R package version 1.16.1.

MacIntyre D.A., Jiménez, B., Lewintre EJ, Martín, C.R., Schäfer, H., Ballesteros, C.G., Mayans, J.R., Spraul, M., García-Conde, J., Pineda-Lucena, A. (2010). Serum metabolome analysis by 1H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. *Leukemia*, 24(4):788-97.

Mamas, M., Dunn, W. B., Neyses, L., Goodacre, R. (2011). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch Toxicol,* 85(1):5–17.

Mannoor, K., Laio, J., Jiang, F. (2012). Small nucleolar RNAs in cancer. *Biochim Biophys,* 182(1):121-8.

Markley, J.L., Brüschweiler, R., Edison, A.S., Eghbalnia, H.R., Powers, R., Raftery, D., Wishart, D.S. (2017). The future of NMR-based metabolomics. *Curr Opin Biotechnol*. 43:34-40.

Montaner, D., Minguez P, Al-Shahrour, F., Dopazo, J. (2009). Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10:197.

Montaner, D., Dopazo, J. (2010). Multidimensional gene set analysis of genomic data. *PLoS One* 27;5(4):e10348.

Mosteller, F., Tukey, J.W. (1977). Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley Publishing Company, Reading (MA), USA.

NHGRI (2015). National Human Genome Research Institute: Transcriptoma. [último acceso el 10/07/17]. Accesible en: https://www.genome.gov/27562853/transcriptoma/.

Pagès, H., Carlson, M., Falcon, S., Li, N. (2017). AnnotationDbi: Annotation Database Interface. R package version 1.38.2.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., Brazma, A. (2005). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res,* 1;33:D553-5.

Pavlova, N.N.,& Thompson, C.B. (2017). The emerging hallmarks of cáncer metabolism. *Cell Metab,* 12;23(1):27-47.

Puchades-Carrasco, L. (2013). Aplicaciones de la RMN a la identificación de nuevos biomarcadores de utilidad clínica en oncología. Tesis Doctoral. Universidad de Valencia, Valencia (España).

Puchades-Carrasco, L., Jantus-Lewintre, E., Pérez-Rambla, C., García-García, F., Lucas, R., Calabuig, S., Blasco, A. Dopazo, J., Camps, Carlos., Pineda-Lucena, A. (2016). Serum metabolomic profiling facilitates the non-invasive identification of metabolic biomarkers associated with the onset and progression of non-small cell lung cancer. *Oncotarget*, *7*(11), 12904–12916.

Puchades-Carrasco, L., Lecumberri R, Martínez-López J, Lahuerta JJ, Mateos MV, Prósper F, San-Miguel JF, Pineda-Lucena A. (2013). Multiple myeloma patients have a specific serum metabolomic profile that changes after achieving complete remission. *Clin Cancer Res*, 19(17), 4770-9.

Puchades-Carrasco, L., Palomino-Schätzlein, M., Pérez-Rambla, C., Pineda-Lucena, A. (2016). Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. *Brief Bioinform*, 17 (3):541-552.

Rampal, R., Al-Shahrour, F., Abdel-Wahab, O., Patel, J.P., Brunel, J.P., Mermel, C.H., Bass, AJ., Pretz, J., Ahn, J., Hricik, T., Kilpivaara, O., Wadleigh, M., Busque, L., Gilliland, D.G., Golub, T.R., Ebert, B.L., Levine, R.L (2014). Integrated genomic analysis illustrates the central role of

JAK-STAT pathway activation in myeloproliferative neoplasm pathogenesis. *Blood*, 29;123(22):e123-33.

Min, B., Park, B., Kim, K.H., K.H, Choi, I.G. (2015). RbioRXN: Process Rhea, KEGG, MetaCyc, Unipathway Biochemical Reaction. Data. R package version 1.5.1.

REDECAN (2014). Red Española de Registros de Cáncer. Estimaciones de la incidencia y la supervivencia del cáncer en España y su situación en Europa. [último acceso el 05/07/17]. Accesible en: http://.www.redecan.org/es/download_file.cfm?file=257&area=196.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, CW., Shi, W, Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res,* 43(7):e47.

Ritchie, M.E., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet,* 16(2):85-97.

Sartor, M.A., Leikauf, G.D., Medvedovic, M. (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics,* 15;25(2):211-7.

Schroder, M.S., Gusenleitner, D., Quackenbush, J., Culhane, A.C, Haibe-Kains, B. (2013). RamiGO: an R/Bioconductor package providing an AmiGO Visualize interface. *Bioinformatics*, 29:666-668.

SEOM (2017). Sociedad española de oncología médica: Las cifras del Cáncer en España. [último acceso el 05/07/17]. Accesible en: http://www.seom.org/seomcms/images/stories/recursos/Las_cifras_del_cancer_en_Esp_2017.pdf.

Singh, P., Yang, M., Dai, H., Yu, D., Huang, Q., Tan, W., Kernstine, KH., Lin, D., Shen, B. (2008). Overexpression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. *Mol Cancer Res*, 6(11):1710-7.

Shen, R., Olshen, A.B., Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics,* 26(2):292-3.

Shukla, S. (2014). FOXO3a: A Potential Target in Prostate Cancer. *Austin J Urol*, 1(1): 4.

Smith, C.A., O'Maille, G, Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Ther Drug Monit*, 27(6):747–51.

Steele, E., Tucker, A. (2008). Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. *J Biomed Inform*, 41(6):914-26.

Sterne, J.A., Egger, M., Smith, G.D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ,* 14;323(7304):101-5.

Strausberg RL, Feingold EA, Klausner RD, Collins FS. (1999). The Mammalian Gene Collection. *Science*. 286:455-457.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545-50.

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fieh, O., Higashi, R., Nair, K.S. (2016). Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res,* 44:D463–70.

Tautenhahn, R., Patti, G.J., Rinehart, D., Siuzdak, G. (2012). XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal Chem,* 84(11):5035–5039.

Tenenbaum, D. (2017). KEGGREST: Client-side REST access to KEGG. R package version 1.16.1.

The GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet,* 45(6):580-5.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 45: D158-D169.

Travis, W. D. (2002). Pathology of lung cancer. *Clin Chest Med*, 23(1):65– 81, viii.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Markley, J. L. (2008). BioMagResBank. *Nucleic Acids Res*, *36*:D402–D408.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods,* 7:55-79.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *J Stat Softw,* 36(3):1–48.

Vogt, P.K., Hart, J.R., Gymnopoulos, M., Jiang, H., Kang, S., Bader, A.G., Zhao, L., Denley, A. (2013). Phosphatidylinositol 3-kinase (PI3K): The Oncoprotein. *Curr Top Microbiol Immuno*l, 2011; 347: 79–104.

von Hipper, P.T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Med Res Methodo.*, 5: 35.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, USA.

Wang, H., Zhang, X., Teng, L., Legerski, R.J. (2015). DNA damage checkpoint recovery and cancer development. *Exp Cell Res,* 10;334(2):350-8.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M., Spencer, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc,* 99:909-917.

Xia, J., Fjell, C.D., Mayer, M.L., Pena, O.M., Wishart, D.S., Hancock, R.E. (2013). INMEX-- a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res,* 41(Web Server issue):W63-70.

Yauk, C.L., Berndt, M.L., Williams, A., Douglas, G.R. (2004). Comprehensive comparison of six microarray technologies. *Nucleic Acids Res,* 32(15):e124.

Zimmermann,K., Leser, U. (2010). Analysis of Affymetrix Exon Arrays [último acceso el 05/07/17]. Accesible en:
https://edoc.hu-berlin.de/bitstream/handle/18452/3142/235.pdf?sequence=1&isAllowed=y.