UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE BIOTECNOLOGÍA

ABORDAJES BIOINFORMÁTICOS PARA LA INTEGRACIÓN FUNCIONAL DE DATOS TRANSCRIPTÓMICOS EN BIOMEDICINA

TRABAJO FIN DE MÁSTER EN BIOTECNOLOGÍA BIOMÉDICA

ALUMNA: SANDRA ALANDES ESTEVE

TUTOR: FRANCISCO GARCÍA GARCÍA

Curso Académico: 2016-2017

VALENCIA, 6 DE JULIO DE 2017

Resumen

La aplicación de técnicas de alto rendimiento como los microarrays y la secuenciación masiva han potenciado los estudios transcriptómicos como abordajes de interés en las áreas biomédicas. Su uso ha generado un gran volumen de información de tipo biológico, accesible en los repositorios públicos de datos ómicos, como GEO (*Gene Expression Omnibus*) o SRA (*Sequence Read Archive*), pudiéndose emplear y combinar en diversas aproximaciones analíticas con el objetivo de responder a nuevas preguntas de investigación. En este trabajo se presentan dos abordajes *in silico* donde la integración de datos procedentes de estudios transcriptómicos proporciona nuevo conocimiento científico capaz de solucionar problemas biológicos o clínicos. En el primero de ellos se integran a nivel funcional los niveles de expresión génica y miARN de datos procedentes de un estudio en cáncer de colon, considerando el efecto inhibidor de los miARN sobre los ARNm. El segundo abordaje permite la creación de un panel de genes de respuesta a un fármaco a partir de la integración de los resultados del enriquecimiento funcional de tres estudios de expresión donde se evaluó la misma respuesta farmacológica. Ambas propuestas confirman las aplicaciones transcriptómicas *in silico* como potentes enfoques para resolver problemas clínicos o biológicos mediante procedimientos computacionales.

Índice

1 Introducción

Actualmente nos encontramos en un momento histórico a nivel científico y tecnológico. Gracias a la actividad de la comunidad científica volcada en el estudio del ADN, en pocas décadas se ha conseguido conocer cómo se almacena la información biológica y la vida en sí misma, y cómo se estructura, funciona y se regula esta información, lo que ha permitido planificar y desarrollar numerosos estudios de investigación en los que se emplea este gran volumen de datos públicos y accesibles.

Desde que en 1869 Miesher descubrió la molécula del ADN fueron muchos los científicos que se interesaron por ella. En los años '20 Levene describió la base bioquímica del ADN y los experimentos de Hershey-Chase probaron que el ADN era la molécula que poseía la información genética y la transmitía a las células hijas. Cuando en 1953, Watson y Crick describieron la estructura del ADN los avances en el conocimiento se sucedieron rápidamente llegando a lo que se llamó el “Dogma central de la biología molecular”.

El dogma central de la biología molecular describe cómo un gen que forma parte de una secuencia de ADN, es la información necesaria para acabar formando una proteína. Los procesos que lo hacen posible son la transcripción y la traducción. Durante la transcripción el ADN es copiado a ARN y éste se traduce más tarde a proteínas. Este hecho cobra especial relevancia con la aparición de las tecnologías de secuenciación, ya que nos han permitido estudiar el genoma a gran escala con el objetivo de encontrar qué proteína codifica cada gen y cuál es su función, al mismo tiempo que estudiamos cuál la aportación funcional de todas esas secuencias que no son genes.

Por otra parte, también ha habido grandes avances tecnológicos y la creación de nuevas técnicas para el estudio del ADN. En 1975 Sanger desarrolló un método de secuenciación que fue la base sobre la que se construyeron posteriormente las tecnologías NGS (*Next Generation Sequencing)*.

En la actualidad, la que muchos han llamado la "era ómica", tenemos amplios conocimientos sobre las bases de la vida y tecnologías de alto rendimiento a nuestro alcance que hacen posible obtener gran cantidad de datos ómicos. La limitación principal se encuentra en el almacenamiento y análisis de esos datos, es por ello que la Bioinformática se ha convertido en una herramienta imprescindible en investigación biológica y biomédica. Gracias a los repositorios públicos de datos ómicos somos capaces de realizar nuevas aproximaciones *in silico* con nuevos abordajes bioinformáticos a través de la integración de datos, aportando así nuevo conocimiento científico.

1.1 Bioinformática

Las tecnologías de alto rendimiento generan un gran volumen de datos "*ómicos*" que han ido aumentando exponencialmente en las últimas décadas. Gracias a la Bioinformática es posible analizar estos datos cuyos resultados nos ofrecen la posibilidad de generar nuevas hipótesis comprobables.

La Bioinformática es útil en muchos campos de estudio, desde la investigación básica hasta los estudios traslacionales y ofrece gran cantidad de posibilidades (tabla 1).

Tal y como la define el NCBI (*National Center for Biotechnology Information*) (Kans *et al*., 2001): "La Bioinformática es el campo de la ciencia en el que la Biología, la Ciencia y Tecnología Informática se unen para formar una única disciplina. El objetivo final de este campo es tanto permitir el descubrimiento de nuevas visiones biológicas, como crear una perspectiva global que unifique los principios que forman parte de la Biología."

	ómica	Material de estudio
Genética molecular	Genómica	Genes
	Epigenómica	Modificaciones epigenéticas
	Exposómica	Factores ambientales causantes de enfermedad
	Exómica	Exones del genoma
	ORFeómica	Open Reading Frames (ORF)
	Fenómica	Fenotipos
	Farmacogenómica	Impacto de los genes en la respuesta a fármacos
	Farmacogenética	SNPs y su impacto en farmacodinámica y farmacocinética
	Toxicogenómica	Respuesta génica a sustancias tóxicas
Biología Molecular	Proteómica	Proteínas y aminoácidos
	Metabolómica	Metabolitos
	Transcriptómica	Transcritos (rRNA, mRNA, tRNA y miRNA)
	Ionómica	Biomoléculas inorgánicas
	Kinómica	Protein-quinasas
	Metagenómica	Material genético de múltples organismos
	Regulómica	Factores de transcripción y otras biomoléculas envueltas en la regulación de la expresión génica
	Toponómica	Estructura celular y tisular

Tabla 1. Campos de investigación biológica en los que se requieren análisis bioinformáticos.

1.1.1 Estudios *in silico*

En Genómica y Transcriptómica existen repositorios de datos disponibles a partir de los cuales se pueden realizar diferentes abordajes computacionales con el objetivo de encontrar diferencias tanto a nivel genómico como transcriptómico, que puedan ayudar a los científicos a centrar sus investigaciones para confirmar los resultados con experimentos realizados *in vitro* e *in vivo.* Este tipo de estudios que incluyen la utilización de datos procedentes de otros experimentos y que emplean métodos bioinformáticos para su tratamiento y análisis son denominados estudios *in silico*.

Hasta la llegada de la Bioinformática, los científicos debían realizar una minería de datos basada en lectura de artículos en los que se ponía de relieve unos pocos genes, actualmente esta minería de datos se ha convertido en un estudio computacional en el que se extrae información a partir de grandes volúmenes de conjuntos de datos (Raja *et al.* 2017).

Las ventajas de ese tipo de estudios son su bajo coste, ya que no es necesario crear nuevos datos en esta fase, y la posibilidad de realizar análisis masivos e integrar datos procedentes de diferentes experimentos o estudios. Además, el hecho de emplear datos obtenidos a partir de publicaciones científicas otorga calidad y rigor científico a este tipo de análisis.

1.2 Transcriptómica

La Transcriptómica, según la revista *Nature,* es el "estudio del set completo de transcritos de ARN producidos por el genoma, bajo circunstancias específicas o en una célula específica, usando métodos de alto rendimiento, como los análisis de microarrays". La comparación del transcriptoma permite la identificación de genes que se encuentran diferencialmente expresados en distintas poblaciones celulares o en diferentes circunstancias.

Los transcritos son, tanto moléculas de ARNm que codifican proteínas, como cualquier otra molécula de ARN no codificante, como los microARN (miARN). Los miARN son pequeños fragmentos de ARN no codificante implicados en la regulación génica post-transcripcional, impidiendo la traducción del ARNm a proteína. Se trata de un método natural de autorregulación de la expresión génica.

Debido a la relación directa que tienen los miARN con la regulación de la expresión génica, la integración de los datos de expresión de miARN junto con la expresión de ARNm ofrece una visión más aproximada de la compleja realidad biológica, lo que puede mejorar la descripción del desarrollo y funcionamiento de las enfermedades, permitiendo un mejor conocimiento de los mecanismos de regulación implicado, así como la posible detección de biomarcadores de interés y de nuevas dianas para fármacos.

El valor de los estudios transcriptómicos reside en que lo que dirige el desarrollo y el funcionamiento celular es la expresión génica y su regulación, que hace que cada grupo celular sea diferente y esté especializado en diferentes funciones, teniendo la misma información genética. Existen innumerables estudios de búsqueda de expresión

génica diferencial entre grupos de pacientes, en diferentes condiciones y entre diferentes tejidos del mismo organismo. En Biomedicina, este tipo de estudios son empleados para la caracterización de enfermedades y la obtención de biomarcadores. Los estudios transcriptómicos han aportado a la comunidad científica una ingente cantidad de datos, gracias a su bajo coste respecto a otros abordajes ómicos y a sus diversas aplicaciones.

1.2.1 Repositorios públicos de datos ómicos

El gran volumen de datos genómicos generados gracias a las tecnologías de alto rendimiento, ha hecho que la comunidad científica demande la creación de repositorios donde almacenar estos datos procedentes de diversos estudios de manera organizada, con un formato común y accesible. Existen múltiples repositorios como *Gene Expression Omnibus* (*GEO*, www.ncbi.nlm.nih.gov), *Sequence Read Archive* (*SRA,* www.ncbi.nlm.nih.gov/sra) y *Array Express* (www.ebi.ac.uk/microarray-as/ae), e incluso algunas organizaciones publican sus datos para su uso público por parte de investigadores de todo el mundo.

En este trabajo se han utilizado datos obtenidos de la plataforma GEO, la mayor base de datos pública de expresión génica y funcional, propiedad del NCBI (*National Center for Biotechnology Information*, www.ncbi.nlm.nih.gov). Se encuentra activa desde el año 2000 y en ella se pueden encontrar tanto datos de expresión obtenidos por arrays como por secuenciación. Además cuenta con herramientas para consultar y descargar experimentos y perfiles de expresión génica curados, ya que todos ellos forman parte de publicaciones, por lo que ofrece un alto nivel de fiabilidad de los datos almacenados. La gran ventaja de este repositorio es que todos los datos aquí publicados presentan un formato compatible con MIAME (*Minimum Information About a Microarray Experiment*), lo que simplifica su utilización.

1.2.2 Análisis de datos transcriptómicos

1.2.2.1 Estrategias de análisis

Las estrategias de análisis dependen del tipo de datos y los objetivos establecidos. En cualquier análisis de datos transcriptómicos se debe realizar un preprocesamiento específico dependiendo de la tecnología de alto rendimiento empleada. Tras este preprocesamiento se obtiene una matriz en la que aparece la cuantificación de la expresión de cada transcrito y muestra biológica del estudio, sobre la que se aplicará la estrategia de análisis seleccionada, siendo las más comunes:

- Análisis de expresión diferencial: permite la detección de diferencias de expresión entre las condiciones del experimento estudiado. Se puede realizar en diferentes niveles: gen, transcrito, miARN… y los resultados identifican grupos de elementos biológicos con diferencia de expresión o bien ordenan todos ellos según su nivel de expresión diferencial.

- Predicción de clases: se emplea para la localización de un elemento biológico, a partir del entrenamiento de muestras iniciales, que sea capaz de clasificar nuevas muestras en los diferentes grupos del ensayo (Lee *et al.,* 2005).

- Análisis clúster: agrupa muestras o elementos biológicos con un patrón común de expresión.

1.2.2.2 Caracterización funcional

Tras el estudio de expresión diferencial nos encontramos ante una lista de elementos biológicos (genes o miARN) ordenados según su expresión diferencial siguiendo criterios estadísticos y biológicos. Mediante el uso de la información disponible en las bases de datos biológicos, podemos conocer qué funciones están caracterizando a estos genes de interés. Algunas de las bases de datos en las que podemos encontrar anotaciones funcionales son:

- GO, *Gene Ontology* (Ashburner & others 2000): clasifica las anotaciones funcionales en procesos biológicos, funciones moleculares o componentes celulares.

- KEGG (Kanehisa & Goto 2000): trabaja con redes de interacción molecular.

Para combinar la información procedente de nuestros experimentos y la información biológica disponible en las bases de datos, existen diversos métodos para la caracterización o enriquecimiento funcional de los estudios transcriptómicos, que se pueden agrupar en métodos de análisis de sobrerrepresentación y métodos de grupos de genes.

En los métodos de análisis de sobrerrepresentación, AS (Al-Shahrour et al. 2004; Al-Shahrour et al. 2007), se caracteriza funcionalmente un grupo de genes de interés. Para ello se determina si cada una de las funciones evaluadas aparece más sobrerrepresentada en este grupo frente a una referencia (resto del genoma u otra lista de genes). Las funciones estadísticamente significativas son detectadas mediante la aplicación de test estadísticos de asociación (hipergeométrico, Fisher…)

El AS es el método más ampliamente utilizado para la determinación del perfil funcional de genes o miARN. Sin embargo, estos métodos presentan varias limitaciones: trato igualitario de los genes o miARN significativos, dependencia del punto de corte elegido en la selección de elementos biológicos de interés…. Además, en el caso de los miARN no se tiene en cuenta que su efecto puede ser aditivo, por lo que pequeñas desregulaciones pueden tener una importante relevancia. Por todo ello, en este trabajo se ha realizado un análisis de enriquecimiento de grupo de genes, GSEA (*Gene Set Enrichment Analysis*), (Subramanian *et al*. 2005) que, a diferencia de los AS, considera el conjunto de todos los genes que forman parte del estudio en su enriquecimiento funcional. El objetivo es averiguar si los genes pertenecientes a un mismo grupo funcional tienden a agruparse en la parte superior o inferior de la lista de todos los genes ordenada según su nivel de expresión diferencial, lo que se interpreta como una sobrerrepresentación significativa de una función. Para emplear los métodos de análisis GSEA en estudios de expresión de miARN es necesario

transferir la evidencia de expresión de miARN a expresión génica, ya que las anotaciones se encuentran en las bases de datos a nivel de gen.

1.2.2.3 Integración de datos

La integración de datos transcriptómicos procedentes de varios estudios o bien obtenidos desde distintos elementos biológicos para un mismo estudio (ARNm y miARN), mejoran los resultados obtenidos con la incorporación de diferentes niveles de información y generan un nuevo conocimiento científico capaz de dar respuesta a preguntas de investigación en diversos escenarios científicos. Habitualmente esta integración se realiza a nivel de gen, miARN…, sin embargo cuando la evaluación conjunta de estos datos se produce a nivel de función, nos dirigimos directamente al fenotipo producido por estos elementos biológicos, proporcionando una detallada descripción del desarrollo y de los mecanismos de regulación de las enfermedades, mejorando el conocimiento sobre éstas y haciendo posible la detección de biomarcadores de interés.

Las estrategias de integración combinan diferentes niveles de información biológica en un mismo análisis. En ellas se evalúa el efecto global sobre las funciones biológicas, afinando los resultados tras considerar la interacción entre los distintos niveles de información. En este trabajo se ha realizado la integración de datos transcriptómicos de ARNm y miARN de muestras de pacientes con cáncer de colon, con el objetivo de mejorar el conocimiento de los mecanismos moleculares de esta enfermedad.

Otra estrategia de integración, consiste en la combinación de un mismo tipo de datos genómicos que proceden de diferentes estudios, en los que se evalúa una cuestión común, por ejemplo la detección de diferencias transcriptómicas entre casos y controles. Este abordaje conjunto, permite ampliar el tamaño muestral obteniendo una mayor potencia en los procedimientos estadísticos utilizados. En este trabajo se ha realizado la integración de tres conjuntos de datos transcriptómicos de ARNm de pacientes con la misma enfermedad y que respondieron o no a un tratamiento con el

objetivo de encontrar aquellas funciones diferencialmente representadas que puedan ser objeto de estudio en un panel de genes de predicción de respuesta a fármaco.

2 Objetivos

El objetivo principal de este trabajo es la resolución de problemas biológicos y clínicos con abordajes bioinformáticos en estudios transcriptómicos.

Para ello se llevarán a cabo dos abordajes *in silico* en los que la integración funcional de datos aportará una mejor interpretación biológica y clínica a los problemas tratados.

2.1 Integración de datos de ARNm y miARN en un mismo estudio

En este ejemplo de integración funcional se integrarán datos transcriptómicos de ARNm y miARN de muestras de pacientes con adenocarcinoma de colon en los diferentes estadios de la enfermedad con el objetivo de interpretar su desarrollo observando qué funciones tienen una representación diferencial dependiendo del estadio de la enfermedad.

2.2 Integración de datos transcriptómicos que evalúan una misma situación

En el segundo ejemplo se integrarán tres conjuntos de datos de ARNm procedentes de tres estudios independientes en los que se administró a pacientes con artritis reumatoide, un anticuerpo monoclonal dirigido a IL-6. Los pacientes se dividieron en dos grupos, los que respondieron al tratamiento y los que no lo hicieron. El objetivo de este estudio es identificar biomarcadores genéticos predictores de respuesta a fármaco que permitan seleccionar a aquellos pacientes que se beneficiarán del tratamiento evaluado.

3 Integración funcional de datos de ARNm y miARN.

En Biomedicina se han realizado multitud de estudios de expresión diferencial, tanto a nivel de ARNm, como de miARN. Los estudios pareados de datos de ARNm y miARN presentan un mayor valor analítico dado que existe una fuerte interacción entre ambos elementos. La integración funcional de estos elementos biológicos en un único abordaje proporciona un mejor conocimiento de la situación biológica que se estudia, mejorando la interpretación de los resultados.

Con esta propuesta se pretende dar una visión más amplia de la compleja realidad biológica de las enfermedades a través de la integración de datos de ARNm y miARN.

3.1 Material y métodos

3.1.1 Datos

Se descargaron del repositorio GEO los datos de expresión sin procesar de los estudios GSE29621 y GSE29622 correspondientes a ARNm y miARN respectivamente. Se trata de datos de expresión obtenidos con microarrays de Affymetrix a partir de tejido tumoral de 65 pacientes con tumores primarios de adenocarcinoma de colon en sus cuatro estadios (tabla 2).

El cáncer de colon es el tumor maligno de mayor incidencia en España, siendo el adenocarcinoma el tipo tumoral de mayor prevalencia en este órgano. Cuando la lesión neoplásica es de carácter infiltrante se dividen en cuatro estadios, correspondiendo el estadio I a una fase temprana de la enfermedad y el estadio IV a fase avanzada en la que el paciente presenta lesiones metastásicas.

Este estudio resulta especialmente interesante ya que se ha realizado un análisis pareado de expresión génica de ARNm y de miARN sobre los mismos tejidos, lo que no es habitual en los estudios de expresión, fundamentalmente por el coste y la disponibilidad de muestras biológicas.

Muestras				
Total	Estadio I	Estadio II	Estadio III	Estadio IV
65	7	22	18	18

Tabla 2. Número de muestras analizadas de cada uno de los cuatro estadios de adenocarcinoma de colon primario.

3.1.2 Métodos de análisis

En este trabajo se ha realizado un triple abordaje del estudio transcriptómico descrito: en el primero se evaluaron los cambios de expresión a nivel de ARNm, en el segundo a nivel de miARN y, finalmente, se combinaron ambos niveles en un mismo enfoque. En cada uno de los tres escenarios, se consideraron seis comparaciones de interés (tabla 3) que permitieron un mejor conocimiento, a nivel transcriptómico y funcional, de los diferentes estadios en tumores de adenocarcinoma de colon.

Comparaciones
Estadio I vs II
Estadio I vs III
Estadio I vs IV
Estadio II vs III
Estadio II vs IV
Estadio III vs IV

Tabla 3. Comparaciones estudiadas entre los diferentes estadios de adenocarcinoma de colon

El *pipeline* de la figura 1 detalla los pasos realizados en cada uno de los tres enfoques. En los dos primeros, se procesaron los datos, seguido de un análisis de expresión diferencial y a continuación se caracterizaron funcionalmente los resultados obtenidos. En la tercera fase se integraron los niveles de ARNm y miARN a nivel funcional con el objetivo de conseguir una mejor interpretación biológica.

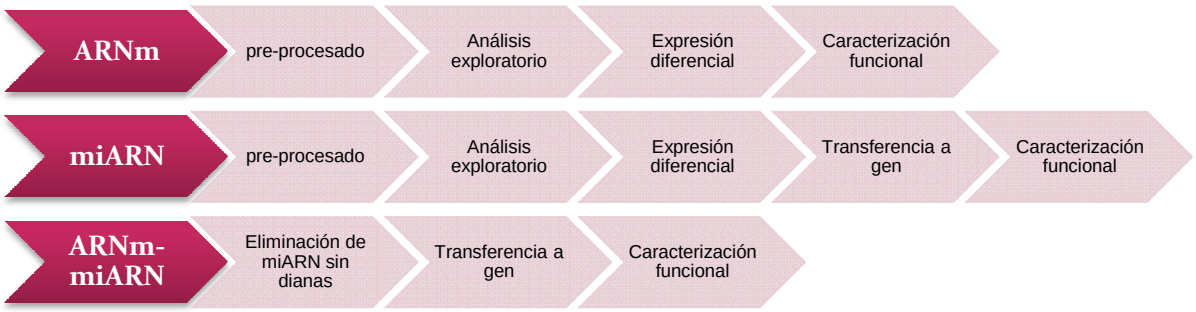


Figura 1. *Pipeline* empleado en el análisis funcional integrado.

3.1.2.1 Análisis de ARNm

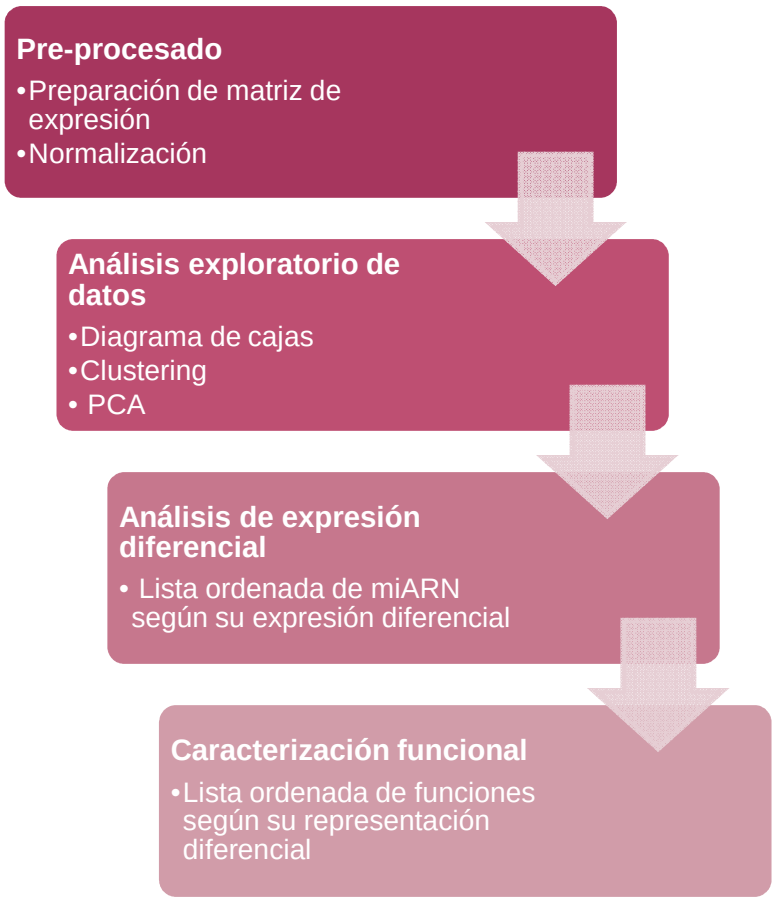


Figura 2. *Pipeline* empleado en el enriquecimiento funcional de ARNm

- Pre-procesado de datos: en primer lugar se preparó la matriz de expresión y los subgrupos de pacientes según el estadio de su enfermedad. En la matriz de expresión las columnas corresponden a las muestras y las filas a los transcritos. Los niveles de expresión se normalizaron utilizando el método RMA (Irizarry *et al.* 2003), implementado en el paquete *affy* de Bioconductor (Gentleman *et al.*

2004). Los identificadores de los transcritos se transfirieron a identificadores *Gene Symbol*, promediando la expresión de los genes que aparecen repetidos con la función *avereps* del paquete *limma* (Smyth, 2005).

- Análisis exploratorio mediante diagrama de cajas, clústers y análisis de componentes principales (PCA) para detectar comportamientos anómalos de muestras y comprobar si los niveles de expresión de los diferentes estadios de la enfermedad se agrupan mostrando similitud de expresión según al grupo de muestras al que pertenecen.

- Análisis de expresión diferencial: se realizó un análisis para cada comparación entre los diferentes estadios de adenocarcinoma de colon. Se llevó a cabo empleando el paquete *limma*, seguido de una corrección *BH* (Benjamini & Hochberg, 1995) del valor de *p* para controlar los falsos positivos**.**

- Caracterización funcional: a partir de las anotaciones disponibles en las bases de datos GO y KEGG se realizó un enriquecimiento funcional de la expresión génica diferencial obtenida en el apartado anterior. Se empleó el método GSEA que utiliza toda la lista de genes ordenada según su expresión diferencial con el objetivo de no sesgar el estudio al escoger pequeños grupos de genes. El método estadístico empleado en este trabajo es la regresión logística (Montaner & Dopazo, 2010), que estudia la dependencia de una variable binaria y una variable continua. En este caso la variable binaria es la anotación funcional, es decir si un gen está anotado o no para una función; la variable continua es el índice que resume el nivel de expresión diferencial. Este método presenta la ventaja de poder combinar diferentes tipos de datos ofreciendo la posibilidad de realizar abordajes genómicos multidimensionales, con la anotación funcional como variable binaria y añadiendo diferentes variables continuas simultáneamente, como por ejemplo, nivel de expresión génica, nivel de metilación…

3.1.2.2 Análisis de miARN

Figura 3. *Pipeline* empleado en el enriquecimiento funcional de miARN

De forma análoga a los métodos empleados en el análisis de datos de expresión de ARNm, se realizó el análisis de miARN. La diferencia entre estos *pipelines* reside en la necesidad de transferir la expresión diferencial de miARN a inhibición diferencial de la expresión génica. Para ello se empleó el método GSEA, con el que a partir de toda la lista de miARN ordenada según su expresión diferencial se creó una lista de genes ordenada según su inhibición diferencial. Esta lista de genes es la que se utilizó para realizar el enriquecimiento funcional. La necesidad de transferir la expresión diferencial de miARN a inhibición diferencial de genes se debe a que las anotaciones funcionales aparecen en las bases de datos en relación a los genes, lo que hace que sea necesario este análisis GSEA en dos pasos para obtener una lista de las funciones según su representación diferencial (Garcia-Garcia *et al.*, 2016).

3.1.2.3 Integración de datos ARNm-miARN

Los miARN tienen un efecto autorregulatorio de los transcritos de ARNm, es por ello que en esta última fase se integraron ambos estudios, pudiendo así aproximarnos a los niveles reales de representación diferencial de las funciones biológicas.

La integración funcional de datos transcriptómicos se realizó en el paso de conversión de la expresión de miARN a gen con el método GSEA modificado. La expresión de miARN se convirtió en genes empleando identificadores *Gene Symbol*, pero únicamente se recogieron para el enriquecimiento funcional aquellos genes que aparecían expresados en los datos de ARNm. De este modo se eliminaron de la lista de genes todos aquellos que no se expresaban en los diferentes escenarios estudiados, ya que el efecto de la expresión de miARN no está causando un efecto inhibitorio sobre ellos. Seguidamente se realizó el enriquecimiento funcional para detectar aquellas funciones representadas diferencialmente en las diferentes comparaciones.

En la figura 4 se indica el diseño del *pipeline* para la integración funcional de ARNm-miARN.

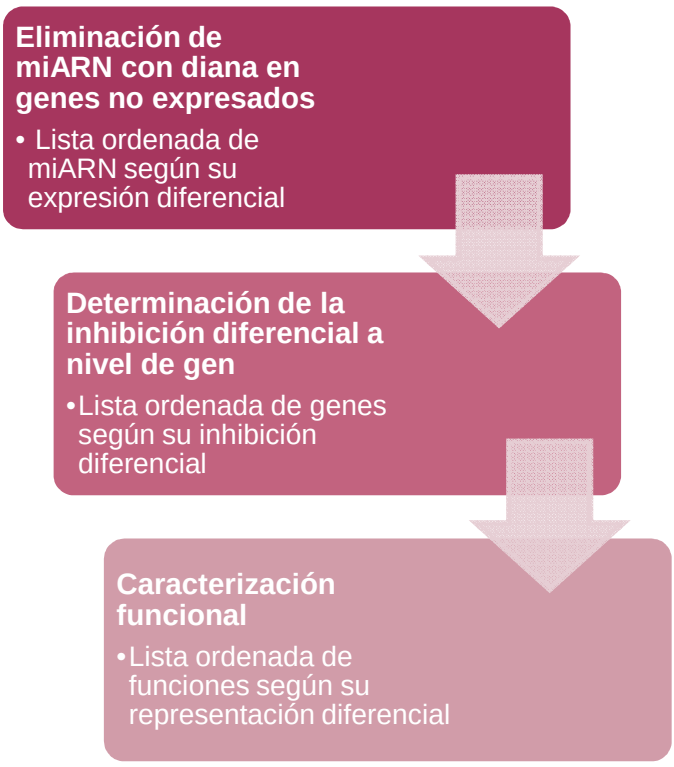


Figura 4. *Pipeline* empleado en la integración funcional de ARNm-miARN

3.2 Resultados

3.2.1 Resultados del análisis de ARNm

- Análisis exploratorio de datos: tras el preprocesamiento y la normalización de los datos se evaluó la distribución de la expresión de las muestras mediante diagramas de cajas (figura 5) en el que no se detectaron muestras con comportamientos anómalos. Con el fin de buscar similitudes de expresión entre las muestras de cada grupo se realizaron dos análisis de clúster con diferentes tipos de distancia, correlación y euclídea (figura 6), en ambos casos las diferentes muestras aparecen mezcladas no separándose bien las que pertenecen a cada estadio de la enfermedad. Asimismo, el PCA mostró el mismo patrón, no consiguiendo separarse las muestras de los diferentes estadios (figura 7). Esto se debe a que todas las muestras son tumorales y, por lo tanto, con pocas diferencias entre ellas al no existir un control negativo de tejido sano. Estos resultados anticipan que no se encontrará una diferencia de expresión génica importante entre los diferentes estadios.

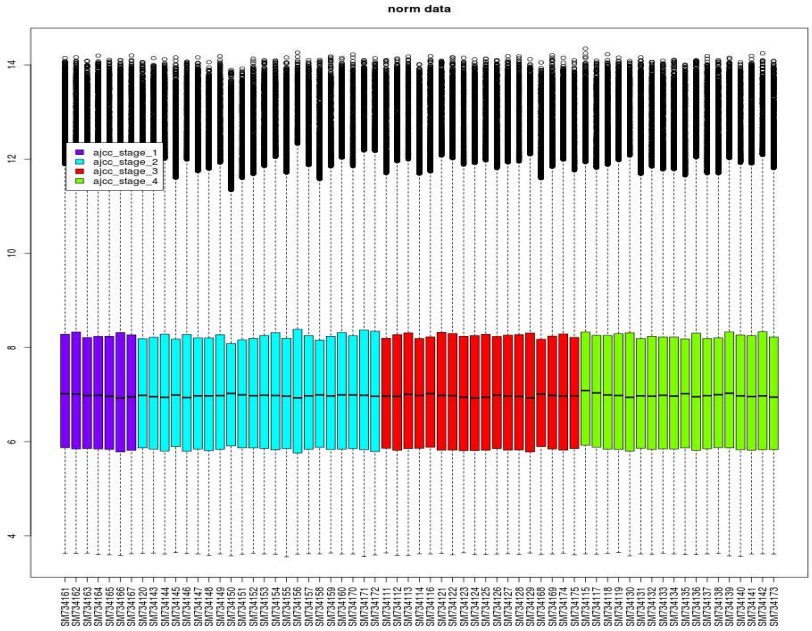


Figura 5. Distribución de los niveles de expresión normalizados a nivel de transcrito. Cada caja hace referencia a los datos de un paciente y cada color representa un estadio de adenocarcinoma de colon.

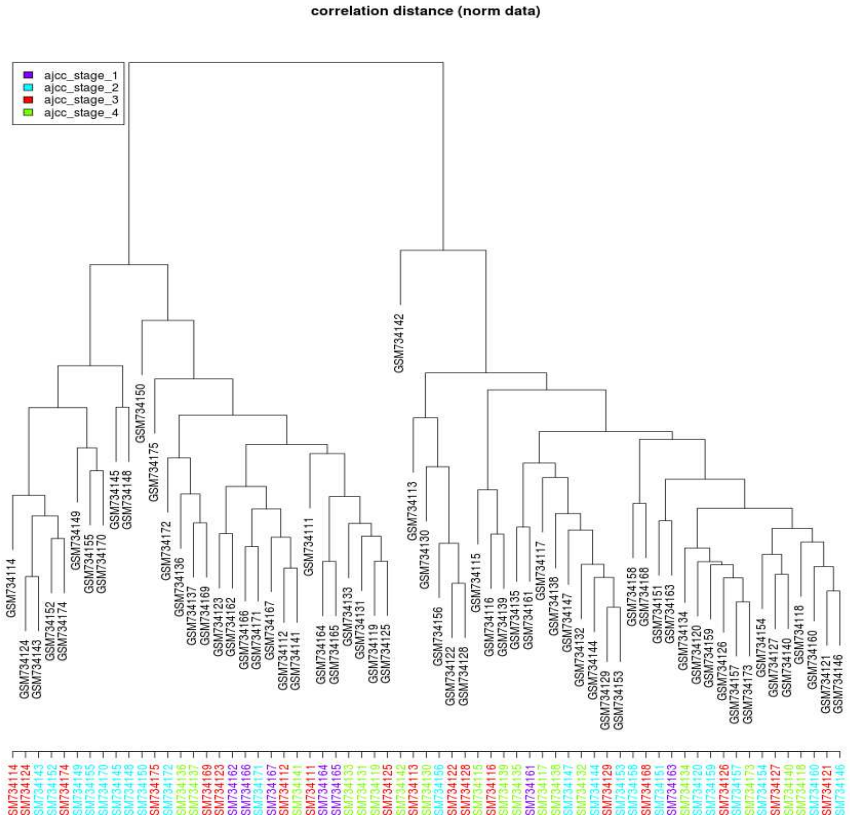


Figura 6. Análisis de clustering de expresión de ARNm. Método de distancia: correlación

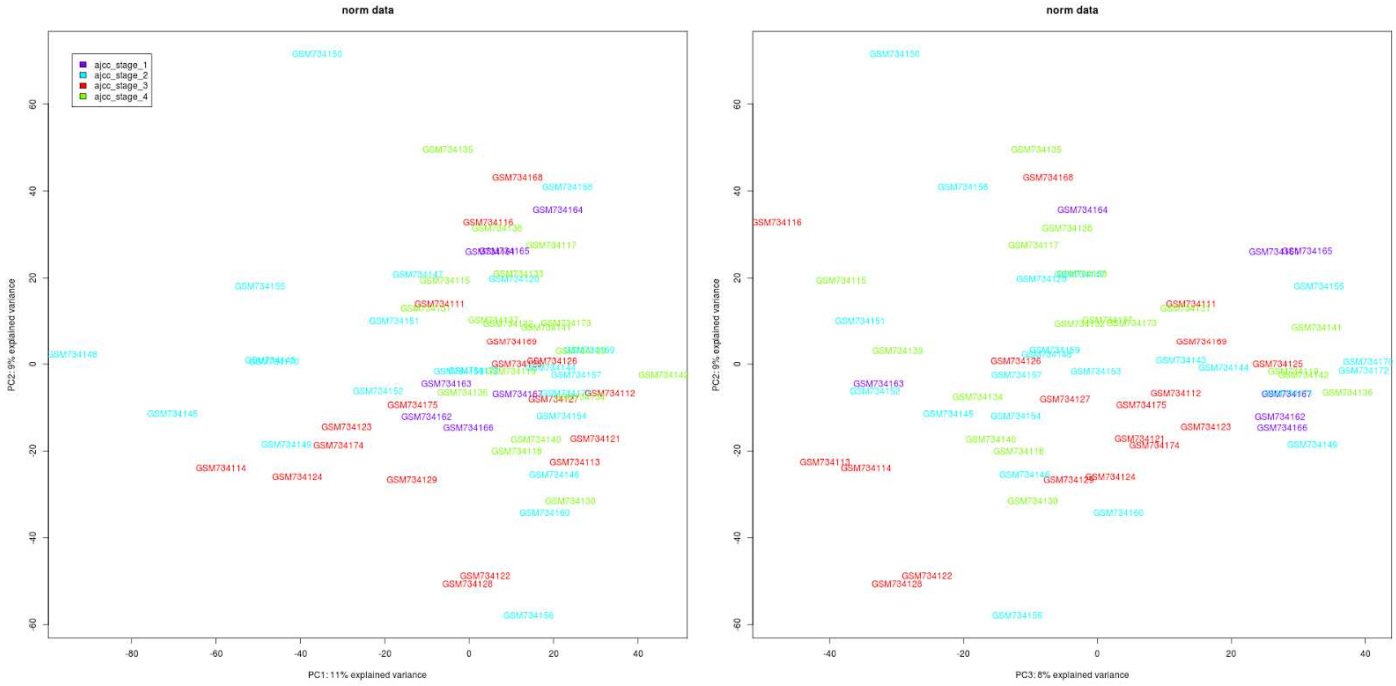


Figura 7. Análisis de componentes principales de la expresión de ARNm en pacientes con adenocarcinoma de colon. Los cuatro estadios de la enfermedad aparecen representados en diferentes colores.

- Expresión diferencial: No se observó expresión diferencial significativa de los genes, en ninguna de las comparaciones descritas (tabla 4).

Expresión diferencial de ARNm				
Comparación	**infra**	**no dif**	**sobre**	**total**
estadio I vs II	0	20639	0	20639
estadio I vs III	0	20639	0	20639
estadio I vs IV	0	20639	0	20639
estadio II vs III	0	20639	0	20639
estadio II vs IV	0	20639	0	20639
estadio III vs IV	0	20639	0	20639

Tabla 4. Resultado del análisis de expresión diferencial de genes en las diferentes comparaciones estudiadas.

- Enriquecimiento funcional: el análisis de enriquecimiento funcional (GSEA) se realizó con el paquete *mdgsa* (Montaner & Dopazo 2010) de Bioconductor. Pese a no observarse expresión diferencial significativa en el patrón de expresión génica de las distintas comparaciones, al realizar el análisis de enriquecimiento funcional con los niveles de diferencia de expresión ordenada, detectamos la presencia de algunas funciones diferencialmente representadas. En las tablas 5 a 8 se muestran el número de términos GO y KEGG asociados a genes que aparecen diferencialmente representados en cada comparación. Por ejemplo, en el enriquecimiento funcional de términos GO relacionados con procesos biológicos, concretamente, en la comparación del estadio I frente al estadio II observamos un total de 11243 términos funcionales, de ellos 240 aparecen significativamente sobrerrepresentados y 287 están significativamente infrarrepresentados, de modo que existe una diferencia fenotípica debido a que estas funciones son más activas en un estadio que en el otro. Teniendo en cuenta los resultados que presentan un log OR mayor, es decir que tienen un mayor nivel de sobrerrepresentación, detectamos que aparecen 232 en el estadio I (columna “sig.o.sobre”) y 259 en el estadio II

(columna “sig.o.infra”). La interpretación del resto de resultados se realizaría de forma análoga.

GO procesos biológicos								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	5342	5901	240	287	232	259	10716	11243
estadio I vs III	5185	6058	228	376	215	340	10639	11243
estadio I vs IV	6406	4837	216	134	196	130	10893	11243
estadio II vs III	5134	6109	130	215	123	192	10898	11243
estadio II vs IV	6884	4359	420	107	395	103	10716	11243
estadio III vs IV	7299	3944	434	108	387	106	10701	11243

Tabla 5. Número de términos GO “procesos biológicos” significativos de las diferentes comparaciones estudiadas en el análisis de ARNm.

GO componentes celulares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	857	675	51	36	39	22	1445	1532
estadio I vs III	813	719	63	66	55	45	103	1532
estadio I vs IV	947	585	53	26	33	19	1453	1532
estadio II vs III	711	821	19	50	15	38	1463	1532
estadio II vs IV	900	632	64	28	42	25	1440	1532
estadio III vs IV	984	548	79	19	45	16	1434	1532

Tabla 6. Número de términos GO “componentes celulares” significativos de las diferentes comparaciones estudiadas en el análisis de ARNm.

GO funciones moleculares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	2167	1807	140	94	127	84	3740	3974
estadio I vs III	2155	1819	149	92	142	87	3733	3974
estadio I vs IV	2310	1664	112	48	103	45	3814	3974
estadio II vs III	1962	2012	71	62	69	51	3841	3974
estadio II vs IV	2226	1748	125	48	117	46	3801	3974
estadio III vs IV	2212	1762	124	53	108	52	3797	3974

Tabla 7. Número de términos GO "funciones moleculares" significativos de las diferentes comparaciones estudiadas en el análisis de ARNm.

KEGG								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	59	240	2	25	1	4	272	299
estadio I vs III	74	225	11	34	1	4	254	299
estadio I vs IV	173	126	9	1	1	0	289	299
estadio II vs III	148	151	8	7	0	1	284	299
estadio II vs IV	261	38	54	1	17	0	244	299
estadio III vs IV	256	43	61	1	10	1	237	299

Tabla 8. Número de términos KEGG significativos de las diferentes comparaciones estudiadas en el análisis de ARNm.

3.2.2 Resultados del análisis de miARN

- Análisis exploratorio de datos: al igual que con los datos de mARN, tras el pre-procesamiento y la normalización, se realizó una evaluación de los niveles de expresión mediante diagrama de cajas en los que no se observaron patrones anómalos en ninguna de las muestras (figura 8). Se llevaron a cabo dos análisis de clustering con diferentes tipos de distancia: correlación y euclídea (figura 9).

En este caso, aunque bien es cierto que las muestras correspondientes al estadio I aparecen agrupadas mayoritariamente en la parte izquierda de los árboles, el resto de estadios no muestran diferencias de agrupación, lo que sugiere que se encontrarán pocas diferencias significativas en la expresión de los miARN, probablemente con mayor nivel diferencial en las comparaciones que implican al grupo del estadio I. El análisis de componentes principales no presenta diferencias de agrupación de las muestras según el estadio de la enfermedad (figura 10).

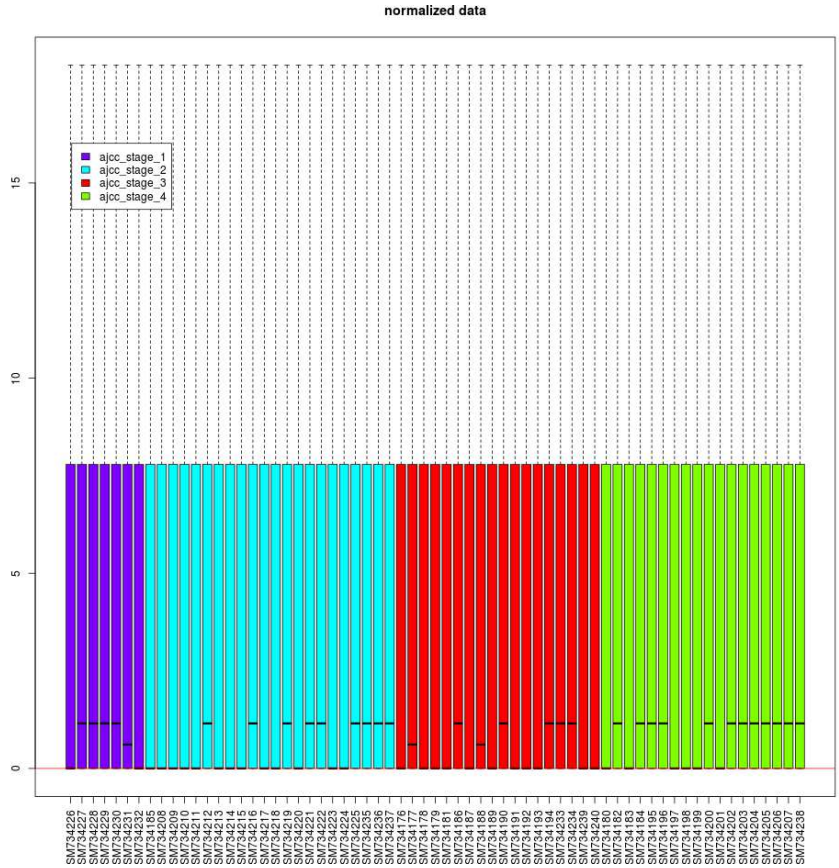


Figura 8. Distribución de los niveles de expresión tras la normalización. Cada caja corresponde a una muestra y cada color a un estadio de adenocarcinoma de colon.

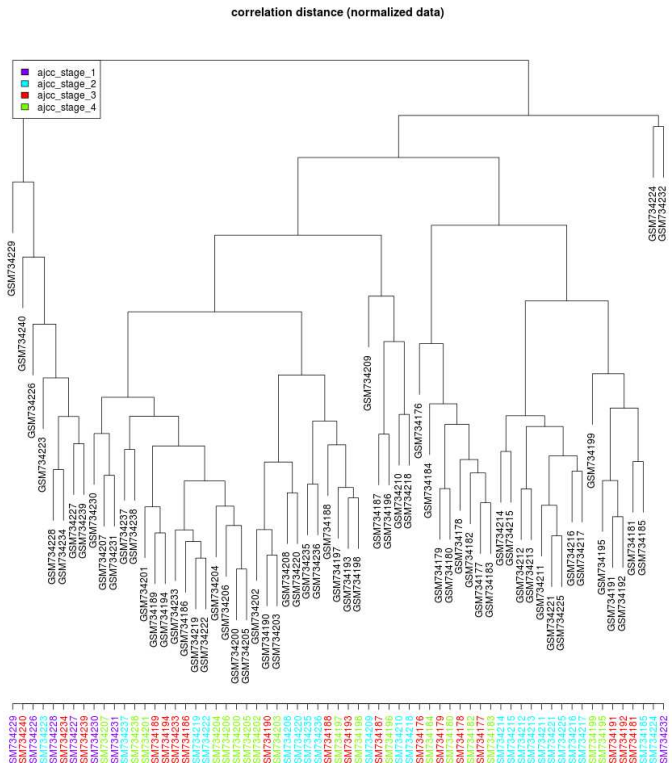


Figura 9. Análisis de clustering de expresión de miARN. Método de distancia: correlación

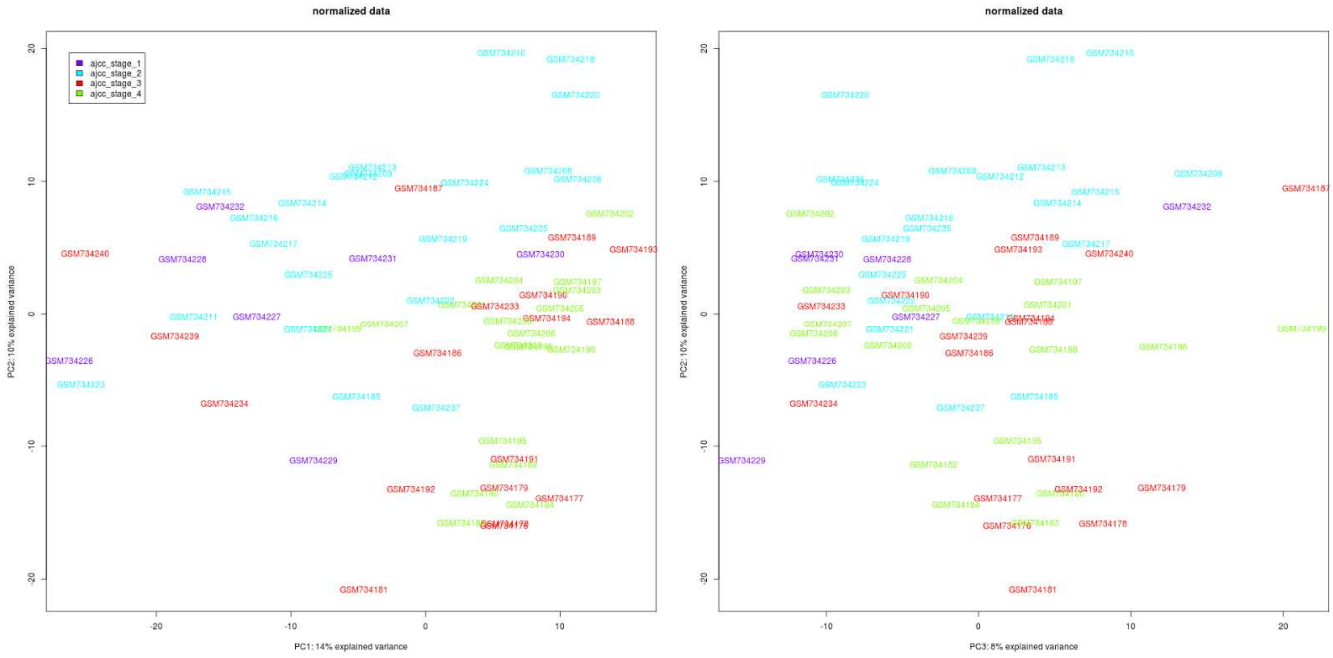


Figura 10. Análisis de componentes principales de la expresión de miARN en pacientes con adenocarcinoma de colon. Los cuatro estadios de la enfermedad aparecen representados en diferentes colores.

- Expresión diferencial: se realizó el análisis de expresión diferencial para las seis comparaciones descritas. La tabla 9 muestra el número de miARN sobre, infra y no diferencialmente expresados en cada comparación. Se encontraron

diferencias significativas de expresión de miARN en cinco de las seis comparaciones, no observándose diferencias entre los estadios III y IV. Las mayores diferencias de expresión se localizaron especialmente en las comparaciones de los estadios I vs III y I vs IV. Los miARN significativos constituyen potenciales biomarcadores para distinguir los estadios de la enfermedad más tempranos de los más tardíos. Además aparecen muchos miARN que se encuentran sobre e infraexpresados, aunque no de una manera significativa, siendo posible que tengan un significado en la inhibición génica dado el carácter aditivo de la regulación de los miARN sobre los genes. Este aspecto también será evaluado en el siguiente paso de la estrategia de análisis.

Expresión diferencial de miARN

Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no dif	total
estadio I vs II	205	242	4	1	4	1	659	664
estadio I vs III	209	232	9	1	9	1	654	664
estadio I vs IV	210	245	9	2	9	2	653	664
estadio II vs III	237	209	1	1	1	1	662	664
estadio II vs IV	216	242	1	0	1	0	663	664
estadio III vs IV	188	263	0	0	0	0	664	664

Tabla 9. Número de miARN expresados diferencialmente en las comparaciones estudiadas.

- Enriquecimiento funcional: pese a ser pocos los miARN que se expresan diferencialmente de modo significativo, gracias al análisis de enriquecimiento funcional, detectamos funciones sobrerrepresentadas en los diferentes estadios de manera significativa, en todas las ontologías y en todas las comparaciones, exceptuando el análisis de los términos KEGG donde no aparece representación diferencial de funciones (tablas 10-13). En la comparación entre los estadios I y II, si observamos los términos GO referidos a los procesos biológicos, se han detectado 171 funciones sobrerrepresentadas y 163 funciones infrarrepresentadas, de las 11243 anotaciones funcionales,

además 170 tienen un nivel de sobrerrepresentación alto en el estadio I (columna "sig.o.sobre") y 160 en el estadio II (columna "sig.o.infra"). Del mismo modo se realizaría la interpretación del resto de resultados.

GO procesos biológicos								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	5136	6107	171	163	170	160	10909	11243
estadio I vs III	5307	5936	201	154	201	154	10888	11243
estadio I vs IV	5220	6023	168	138	168	138	10937	11243
estadio II vs III	6086	5157	210	163	204	163	10870	11243
estadio II vs IV	6138	5105	176	163	176	163	10904	11243
estadio III vs IV	5584	5659	174	191	174	188	10878	11243

Tabla 10. Número de términos GO "procesos biológicos" significativos de las diferentes comparaciones estudiadas en el análisis de miARN.

GO componentes celulares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	694	838	17	29	17	25	1486	1532
estadio I vs III	722	810	20	20	20	20	1492	1532
estadio I vs IV	702	830	14	14	14	13	1504	1532
estadio II vs III	812	720	24	17	22	16	1491	1532
estadio II vs IV	786	746	22	21	22	21	1489	1532
estadio III vs IV	718	814	22	34	21	26	1476	1532

Tabla 11. Número de términos GO "componentes celulares" significativos de las diferentes comparaciones estudiadas en el análisis de miARN.

GO funciones moleculares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	1717	2257	59	56	59	53	3859	3974
estadio I vs III	1700	2274	42	50	42	50	3882	3974
estadio I vs IV	1779	2195	53	64	53	61	3857	3974
estadio II vs III	2111	1863	53	70	49	70	3851	3974
estadio II vs IV	2154	1820	47	59	46	59	3868	3974
estadio III vs IV	2132	1842	66	72	66	67	3836	3974

Tabla 12. Número de términos GO "funciones moleculares" significativos de las diferentes comparaciones estudiadas en el análisis de miARN.

KEGG								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	115	184	0	0	0	0	299	299
estadio I vs III	158	141	0	0	0	0	299	299
estadio I vs IV	138	161	0	0	0	0	299	299
estadio II vs III	178	121	0	0	0	0	299	299
estadio II vs IV	167	132	0	0	0	0	299	299
estadio III vs IV	77	222	0	0	0	0	299	299

Tabla 13. Número de términos KEGG significativos de las diferentes comparaciones estudiadas en el análisis de miARN.

3.2.3 Resultados del análisis de la integración funcional de ARNm y miARN

Para lograr la integración funcional de datos de miARN y ARNm se realizó la trasnferencia de los niveles de expresión de los miARN a ARNm para los genes que sí estaban expresados (presentaban un nivel de expresión mínimo en todas las muestras de todos los grupos evaluados). A continuación se completó el análisis de enriquecimiento funcional sobre la nueva lista de genes ordenados por su nivel de inhibición.

- Enriquecimiento funcional: se detectaron algunas variaciones en las funciones diferencialmente representadas en los distintos grupos. Por ejemplo, en la comparación entre los estadios I y II, los términos GO referidos a procesos biológicos aparecen 173 funciones sobrerrepresentadas y 164 infrarrepresentadas, afinando así el resultado obtenido en el análisis de miARN. Del mismo modo se interpretan el resto de resultados.

GO procesos biológicos								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	5152	6091	173	164	172	161	10906	11243
estadio I vs III	5315	5930	202	153	202	153	10888	11243
estadio I vs IV	5232	6011	169	137	169	137	10937	11243
estadio II vs III	6145	5098	205	152	199	152	10886	11243
estadio II vs IV	6172	5071	173	161	173	161	10909	11243
estadio III vs IV	5647	5596	166	188	166	185	10889	11243

Tabla 14. Número de términos GO “procesos biológicos” significativos de las diferentes comparaciones estudiadas en la integración miARN-ARNm.

GO componentes celulares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	697	835	17	26	17	25	1489	1532
estadio I vs III	718	814	19	20	19	20	1493	1532
estadio I vs IV	698	834	14	13	14	13	1505	1532
estadio II vs III	820	712	23	16	20	15	1493	1532
estadio II vs IV	795	737	22	21	22	21	1489	1532
estadio III vs IV	726	806	20	31	20	25	1481	1532

Tabla 15. Número de términos GO "componentes celulares" significativos de las diferentes comparaciones estudiadas en la integración miARN-ARNm.

GO funciones moleculares								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	1707	2267	57	55	57	52	3862	3974
estadio I vs III	1688	2286	43	48	43	48	3883	3974
estadio I vs IV	1780	2194	55	62	55	60	3857	3974
estadio II vs III	2140	1834	51	69	46	69	3854	3974
estadio II vs IV	2183	1791	47	59	46	59	3868	3974
estadio III vs IV	2154	1820	65	67	65	62	3842	3974

Tabla 16. Número de términos GO "funciones moleculares" significativos de las diferentes comparaciones estudiadas en la integración miARN-ARNm.

KEGG								
Comparación	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra	no enriquecido	total
estadio I vs II	120	179	0	0	0	0	299	299
estadio I vs III	167	132	0	0	0	0	299	299
estadio I vs IV	149	150	0	0	0	0	299	299
estadio II vs III	182	117	0	0	0	0	299	299
estadio II vs IV	177	122	0	0	0	0	299	299
estadio III vs IV	82	217	0	0	0	0	299	299

Tabla 17. Número de términos KEGG significativos de las diferentes comparaciones estudiadas en la integración miARN-ARNm

3.3 Discusión

El enriquecimiento funcional ofrece la posibilidad de estudiar los fenotipos desde un enfoque de la actividad biológica, tanto si se encuentran diferencias de expresión significativas como si no se hubieran detectado cambios de expresión, tal como ocurrió en el análisis de ARNm. Pequeñas diferencias de expresión de varios genes con una misma función son capaces de aumentar/disminuir la representación de esa función de un modo significativo, pudiendo ser evaluadas con este abordaje.

En la expresión diferencial de miARN aparecen algunos miARN diferencialmente expresados de modo significado entre el estadio I y el resto de estadios, por lo que podrían emplearse como biomarcadores para diferenciar los estadios de la enfermedad más tempranos de los más tardíos (tabla 18).

Comparaciones	sig.o.sobre	sig.o.infra
estadio I vs II	4	1
estadio I vs III	9	1
estadio I vs IV	9	2

Tabla 18. Número de miARNs expresados diferencialmente entre los estadios de adenocarconoma I y II, III y IV.

Tal y como ocurre con la expresión de los ARNm, pequeñas diferencias de expresión de los miARN pueden tener un significado importante en la inhibición de la expresión génica dado su carácter aditivo, dando lugar a una diferencia importante de la representación funcional.

La información obtenida a partir de la integración aporta mayor robustez a los análisis de datos de ARNm y miARN que cuando éstos se analizan de manera individual. Esto se debe a que ambos elementos biológicos están estrechamente relacionados regulando la expresión génica por lo que la integración de ambos datos ofrece una visión global y una mejor interpretación biológica de la enfermedad.

El análisis de ARNm referido a la comparación del estadio I vs II, ofrece una lista de funciones sobrerrepresentadas en el estadio I que cuando se compara con los

resultados obtenidos del análisis de miARN para la misma comparación parece completamente diferente. La integración de de ambos datos, al realizarse a partir de la expresión de miARN, eliminando los genes no expresados según los datos de ARNm, se parece mucho a los resultados del análisis de miARN, pero cuando solapamos los resultados, se observa cómo aparecen funciones diferencialmente representadas que con el análisis individual habían pasado inadvertidas (en este caso 7 funciones) y cómo desaparecen algunas funciones que podrían ser falsos positivos (5 funciones), logrando una mayor solidez en los resultados (figura 11).

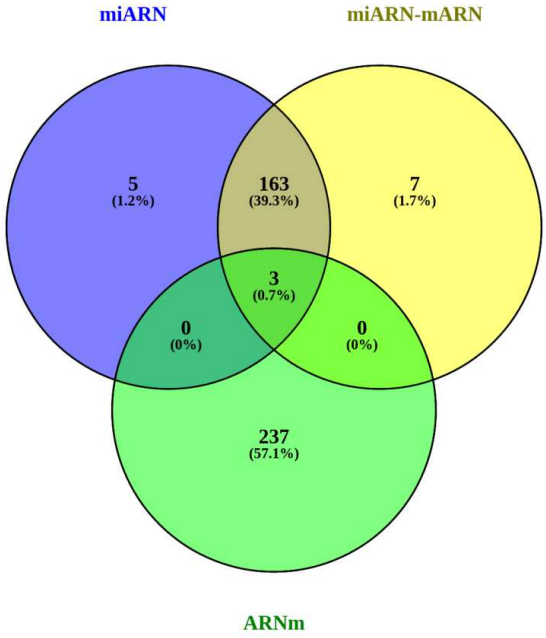


Figura 11. Número de términos GO "procesos biológicos" coincidentes en los tres análisis de enriquecimiento funcional (ARNm, miARN y miARN-ARNm), en la comparación del estadio I frente al estadio II de adenocarcinoma de colon.

Gracias a la integración de datos se detectaron, en la comparación descrita, las siguientes funciones de interés: segregación de cromátidas hermanas, regulación del pH celular, regulación negativa de la muerte celular inducida por estrés oxidativo.

Las imágenes tipo *treemaps* ayudan a visualizar los resultados y a detectar de una manera rápida las diferencias que aportan los distintos tipos de datos. En el anejo 1 se adjuntan los *treemaps* de la comparación descrita con los datos de ARNm, miARN y la integración de ambos elementos. Las funciones diferencialmente representadas en el análisis de datos de ARNm son diferentes a las funciones diferencialmente

representadas en el análisis de datos de miARN, en cambio éstas presentan pocas diferencias con los resultados de la integración. Estos se debe a las características de los datos y ambos resultados son complementarios, empleándose la integración para mejorar la solidez de los estudios.

La integración funcional de los datos de ARNm y miARN ayuda a afinar los resultados. En este análisis las diferencias han sido sutiles, probablemente debido a que se trata únicamente de análisis de tejido neoplásico con el que se pretende comparar las funciones que aparecen diferencialmente representadas según el estadio del cáncer y sin que se haya tomado muestras de tejido intestinal sano de los pacientes.

4 Selección de genes candidatos para el diseño de un panel genético predictor de respuesta a fármaco: Metaanálisis Funcional

Un panel de genes es una prueba genética que analiza varios genes de forma simultánea gracias a las tecnologías de NGS. La secuenciación del genoma o exoma completo genera un gran volumen de información difícil de abordar en la clínica, produciendo habitualmente hallazgos fortuitos y de significación desconocida. La secuenciación de paneles de genes presenta muchas ventajas: se abaratan los costes, todos los hallazgos están relacionados con la patología de estudio, la cobertura es mayor por lo que se evitan falsos positivos, y se obtiene una menor cantidad de datos, lo que facilita su análisis.

La selección de genes que integran un panel diagnóstico resulta una tarea compleja. El metaanálisis de estudios *in silico* permite una aproximación en la selección y priorización de genes candidatos para configurar un panel genético.

El metaanálisis es una herramienta útil para la obtención de información a partir de estudios empíricos individuales relacionados con un tema de interés específico. Debido a los todavía altos costes de las tecnologías de alto rendimiento, la mayoría de los estudios que encontramos incluyen un número reducido de muestras. El metaanálisis ofrece un abordaje global e integrador, ya que permite aumentar el tamaño muestral, consiguiendo así una mayor potencia estadística. No se trata únicamente de aumentar la N, se debe tener presente cuál es la pregunta que se quiere responder y escoger muy bien los datos que nos acerquen a la respuesta, siendo muy importante la selección de los métodos estadísticos y la interpretación de los resultados, teniendo en cuenta la variabilidad de los estudios individuales (Catalá-López & Tobías 2014). Gracias a los metaanálisis es posible combinar múltiples estudios realizados sobre un mismo rasgo biológico con el propósito de responder a una nueva pregunta de interés científico, determinando marcadores comunes a los diferentes estudios.

La mayor parte de los estudios de metaanálisis genómicos intentan localizar variantes génicas, no obstante, en muchas ocasiones la respuesta es más compleja y

en biología existen multitud de factores que interfieren dando lugar a diferencias funcionales. Es por ello, que el metaanálisis a nivel de función nos permite interpretaciones más acordes con los sistemas biológicos y nos orientan para la toma de decisiones en futuras investigaciones.

4.1 Material y métodos

Para poner en valor los estudios *in silico* se ha realizado un metaanálisis funcional en el que se pretende ofrecer al investigador una serie de funciones diferencialmente representadas entre grupos de muestras de pacientes que responden o no a un fármaco, con el propósito de detectar genes candidatos en la creación de un panel predictor de respuesta a fármaco.

La metodología aquí empleada permite utilizar cualquier base de datos públicos o anotaciones propias, además se puede aplicar a datos de expresión génica obtenidos por microarrays o secuenciación masiva, así como en otros escenarios ómicos como metabolómica o proteómica. Se trata de un abordaje funcional flexible con la finalidad de poder ofrecer a los investigadores una mejor interpretación de los resultados, combinando diferentes tipos de datos.

4.1.1 Revisión sistemática y selección de estudios

La selección de los estudios individuales es fundamental para que el metaanálisis sea capaz de responder al objetivo del estudio con calidad y robustez, integrando los resultados. Para ello se deben definir claramente cuáles serán los criterios de inclusión y exclusión, crear un diseño muestral adecuado y valorar la calidad de las publicaciones a las que pertenecen los datos. Existen unas indicaciones consensuadas para evaluar estos aspectos en *PRISMA* (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, www.prisma-statement.org).

En la aplicación de este método al escenario descrito, se han escogido tres conjuntos de datos que provienen de tres experimentos en los que se estudió la expresión génica de pacientes con artritis reumatoide a los que se les trató con un

anticuerpo monoclonal dirigido contra la IL-6. Los pacientes se dividieron en dos grupos: respondedores y no respondedores al tratamiento.

La artritis reumatoide es una enfermedad inflamatoria crónica, de carácter autoinmune, que afecta especialmente a las articulaciones. Afecta aproximadamente al 0,5% de la población y en España cada año se diagnostican entre 10000 y 20000 nuevos casos. El tratamiento con anticuerpo monoclonal dirigido a IL-6 ofrece buenos resultados en algunos de estos pacientes, pero existen otros pacientes que no responden al tratamiento, es por ello que la creación de un panel genético de respuesta a fármaco podría ayudar a los facultativos a proporcionar un tratamiento personalizado según el perfil genético de cada paciente.

Debido a que el objetivo de este metaanálisis es proponer al investigador genes candidatos para la creación de un panel que evalúe la respuesta a un fármaco, los datos escogidos pertenecen únicamente a las muestras obtenidas de los pacientes antes del tratamiento.

Los criterios empleados para la selección de datos fueron:

- Mismo diseño experimental
- Pacientes con artritis reumatoide tratados con anticuerpo anti-IL-6
- Grupos de pacientes que responden al tratamiento y pacientes que no responden al tratamiento.
- Muestras de los pacientes previas al tratamiento

Datos GEO	Origen	Tecnología	Material de estudio	Nº muestras	Respondedores	No respondedores
GSE25160	Hungría	Microarray Affymetrix	Sangre periférica	13	8	5
GSE45867	Bélgica	Microarray Affymetrix	Tejido sinovial	12	7	5
GSE78068	Japón	Microarray Agilent	Sangre periférica	38	8	30

Tabla 19. Características de los estudios seleccionados para su integración funcional en el metaanálisis.

Los datos pertenecientes a estos tres experimentos se descargaron del repositorio GEO, donde se encuentran publicados. Todos ellos consisten en estudios

transcriptómicos por microarrays, tomándose muestras de sangre periférica en dos de ellos y de tejido sinovial en el tercero, antes del tratamiento con el anticuerpo monoclonal (tabla 19).

4.1.2 Métodos de análisis

Los métodos basados en metaanálisis de estudios genómicos a nivel de función se suelen dividir en dos etapas: en la primera se realiza un análisis de expresión diferencial con el que se obtienen estadísticos a nivel de gen para cada conjunto de muestras. En la segunda etapa se realiza el análisis de enriquecimiento funcional y el metaanálisis, momento en el que se integran todos los conjuntos de datos. Algunos métodos optan por integrar el resultado de la expresión diferencial para después realizar el enriquecimiento funcional, mientras que otros integran el resultado del enriquecimiento funcional a nivel de grupos de genes.

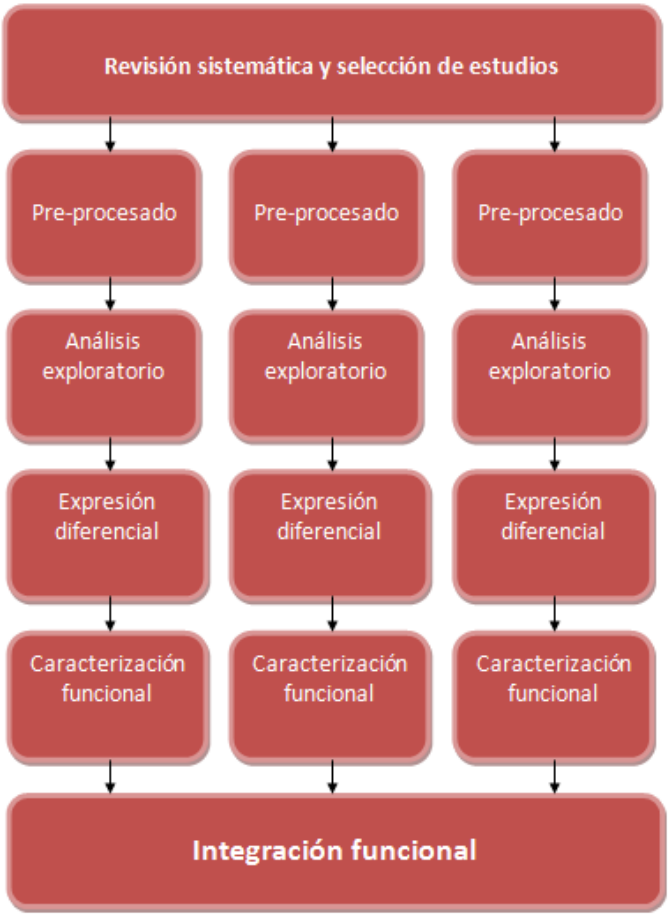


Figura 12. *Pipeline* de integración funcional de datos en un modelo de metaanálisis.

El *pipeline* de la estrategia de análisis utilizada se divide en tres fases:

1. Revisión sistemática y selección de estudios
2. Análisis primario para cada estudio:
    - Preprocesamiento y análisis exploratorio
    - Análisis de expresión diferencial
    - Análisis de enriquecimiento funcional
3. Metaanálisis a nivel de función: integración funcional de todos los estudios

4.1.2.1 Análisis primario

- Preprocesamiento: Los datos de los tres estudios se descargaron como matrices de expresión normalizadas en las que los identificadores se corresponden con el código del fragmento transcrito. Para su posterior análisis de expresión como ARNm se tradujeron a identificadores Gene Symbol, realizándose un promedio de las sondas repetidas con la función *avereps* del paquete *limma*. Finalmente se obtuvieron tres matrices, una por cada estudio individual, en las que las filas corresponden a los genes y las columnas a las muestras (tabla 20).

	Expresión génica				
	GSM206619	GSM2066120	GSM2066121	GSM2066122	GSM2066123
FAM1748	4.27739079	3.6515644	4.01152530	4.87912954	3.77439026
AP352	5.51323491	5.5793266	4.76177048	6.10249229	5.94698157
SV28	2.76275027	2.7706438	0.44999231	0.97444671	2.05054833
RBPMS2	3.93799169	2.0081328	2.42378765	1.23047757	2.77382645
AVEN	6.29085063	6.9040524	6.40980840	6.63258142	6.57106933
ZSCAN29	5.46065627	5.5426277	5.42514752	5.09314117	5.02048124

Figura 20. Ejemplo de matriz de expresión génica normalizada. La columnas corresponden a las muestras y las filas a los genes.

- Análisis exploratorio: tras la preparación de las matrices de expresión de todos los estudios se procedió a su exploración. Para ello se realizaron diagramas de cajas, análisis de clustering y PCAs con el objetivo de detectar un comportamiento anómalo de algunas de las muestras y evaluar las relaciones entre los grupos experimentales definidos.

- Análisis de expresión diferencial: se utilizaron modelos lineales implementados en el paquete estadístico *limma*, comparando los datos de los pacientes respondedores frente a los no respondedores. El *p-valor* se ajustó con el método *BH*.

- Análisis de enriquecimiento funcional: se configuró una lista ordenada a partir del estadístico de contraste obtenido en el análisis de expresión diferencial entre respondedores y no respondedores. En la anotación funcional de los genes se utilizó la base de datos GO (Gene Ontology) y KEGG. La anotación requiso de la exploración de los datos, propagación y filtrado de los términos funcionales. El método de enriquecimiento funcional se basó en modelos de regresión logística (Montaner et al. 2009; Sartor & others 2009; Montaner & Dopazo 2010).

4.1.2.2 Metaanálisis a nivel de función

De cada uno de los estudios de enriquecimiento funcional se obtuvieron matrices de resultados, en las que las filas corresponden a funciones y las columnas son indicadores estadísticos de la sobrerrepresentación funcional obtenida en cada estudio. Para configurar las matrices que incluyen una mayor presencia funcional en cada estudio, se emplearon los valores de *odds ratio,* que mide el efecto entre pacientes respondedores frente a no respondedores, y también las varianzas de los *odds ratio*. Con estos elementos se configuraron tres matrices en las que las filas son las funciones y las columnas hacen referencia a los resultados de cada uno de los estudios individuales. La primera de ellas presenta los resultados de *odds ratio*, mientras la que la segunda matriz muestra la varianza.

log *odds ratio*			
	go_bp_25160_R.vs.NR	go_bp_45867_R.vs.NR	go_bp_78068_R.vs.NR
GO:0002262	-0.206265914	-0.25755403	-0.048125734
GO:0002263	-0.086579202	-0.43296022	0.022760930
GO:0002274	-0.299532921	-0.49759162	-0.051217326
GO:0002275	-0.401737665	-0.52919492	-0.184052174
GO:0002278	-0.939303104	-0.60612489	-0.158789441
GO:0002279	-0.230679507	-0.49957691	-0.206216798

Tabla 21. Niveles de sobrerrepresentación funcional en cada estudio individual, para los términos GO "procesos biológicos".

sd			
	go_bp_25160_R.vs.NR	go_bp_45867_R.vs.NR	go_bp_78068_R.vs.NR
GO:0000002	0.19584808	0.19544440	0.20801094
GO:0000012	0.34030145	0.35087526	0.35333590
GO:0000018	0.12843198	0.12746745	0.13031070
GO:0000022	0.33169124	0.35319884	0.37808459
GO:0000027	0.21305555	0.22238554	0.19705102
GO:0000028	0.23454446	0.24926172	0.24175957

Tabla 22. Niveles de varianza obtenidos del enriquecimiento funcional de cada estudio individual, para los términos GO "procesos biológicos"

Para el metaanálisis de cada término funcional se emplearon las funciones del paquete *metafor* (Viechtbauer 2010) de R, evaluando diferentes métodos de metaanálisis. Se evaluaron el modelo de efectos fijo (FE) y los modelos de efectos aleatorios DL (DerSimonian & Laird, 1986), HE (Hedges *et al.* 2008) y HS (Schmidt & Hunter, 2014). En el modelo FE existe únicamente un efecto en la población sin considerar la variabilidad de los resultados de los estudios individuales, considerando que las diferencias se deben únicamente al azar; en cambio, los modelos de efectos aleatorios incorporan la variabilidad entre los estudios e intraestudio. Dado que el

estudio presenta diferentes fuentes de variabilidad se empleó el modelo de efectos aleatorios DL.

Es necesario realizar una estimación de la medida del efecto para cada una de las funciones, acompañado de indicadores de heterogeneidad del metaanálisis y análisis de estudios influyentes, para asegurar que la variabilidad entre estudios y su influencia no sesga los resultados obtenidos.

Existen diferentes representaciones gráficas para resumir y mostrar los resultados del metaanálisis facilitando su interpretación:

- Gráficos volcán: diagramas de puntos útiles para describir los resultados globales de análisis ómicos u otros análisis con gran volumen de datos. Permiten la representación de dos informaciones de interés para cada uno de los elementos biológicos valorados: en el eje X se indican los logaritmos de *odds ratio* y en el eje Y el nivel de significación.

- Gráficos bosque: muestran el peso de cada estudio individual en el conjunto de resultados. En estos gráficos se indica la estimación de la medida resumen individual de cada estudio y su intervalo de confianza (95%). La medida del efecto aparece representada por el tamaño del cuadrado central que es proporcional a la precisión de las estimaciones (a mayor variabilidad menor tamaño del cuadrado). El segmento en el cual se encuentra el cuadrado indica el intervalo de confianza. EL resultado del metaanálisis aparece con una forma romboidal en la que la anchura indica su precisión y su posición indica su significación estadística.

- Gráficos de embudo: se emplean como indicador de la heterogeneidad del metaanálisis y representa la variabilidad de los estudios. Cada punto representa un estudio individual y en su interpretación se valora la nube de puntos (Sterne & Egger 2001). En el eje X aparece el valor del logaritmo de *odds ratio* y en el eje Y una medida de precisión (desviación estándar o varianza). Si no hay sesgos los puntos deben distribuirse en forma de embudo.

- Gráficos radiales: es un indicador de la heterogeneidad que evalúa la consistencia de los efectos según su nivel de precisión (Galbraith 1988b, Galbraith 1988a, Galbraith 1994). En el eje X se indica la inversa del error estándar y en el eje Y el valor de los efectos estandarizado según su error estándar. A la derecha se indica el valor de los efectos previa estandarización.

4.2 Resultados

Esta metodología proporciona dos tipos de resultados, unos resultados globales en los que presentamos el conjunto global de las funciones que aparecen representadas diferencialmente y unos resultados específicos en los que se evalúa la integración de cada estudio para cada función diferencialmente representada.

La revisión sistemática dio lugar a la selección de tres estudios con diseños experimentales muy similares con los que se obtuvieron un total de 63 muestras de pacientes con artritis reumatoide, de los cuales 23 pertenecen al grupo de los pacientes respondedores al tratamiento con anti-IL-6, y 40 lo hacen al grupo de no respondedores al tratamiento. Los datos de estos tres estudios individuales se descargaron de la plataforma GEO para su integración en el metaanálisis.

4.2.1 Resultados globales

Los resultados del metaanálisis proporcionan funciones que aparecen sobrerrepresentadas o infrarrepresentadas de modo significativo entre todo el conjunto de muestras cuando se comparan el grupo de pacientes respondedores frente a los no respondedores. Se emplearon cuatro métodos de metaanálisis, tres de ellos de efectos aleatorios y uno de ellos de efectos fijos.

En la tabla 23 se muestra el resultado de los términos funcionales GO referidos a procesos biológicos, donde se describe el número de funciones con representación diferencial según el método empleado. Las columnas “sobre” e “infra” hacen referencia al número de funciones que muestran representación diferencial; las columnas “sig.sobre” y “sig.infra” indican las funciones con representación diferencial significativa; las columnas “sig.o.sobre” y “sig.o.infra” señalan el número de funciones

que, además de mostrar representación diferencial, tienen un efecto de mayor magnitud entre los grupos.

De este modo, con el método DL se observa que, cuando comparamos los pacientes que responden al tratamiento frente a los pacientes que no lo hacen, aparecen 2769 funciones sobrerrepresentadas y 5223 funciones infrarrepresentadas, de las cuales 27 se encuentran sobrerrepresentadas de modo significativo y 402 infrarrepresentadas de modo significativo. Las funciones en las que aparece una mayor diferencia de representación son 12 sobrerrepresentadas y 83 infrarrepresentadas, del total de 11243 funciones estudiadas en relación con los términos GO referidos a procesos biológicos.

Los métodos de efectos variables obtienen resultados similares, especialmente entre los métodos DL y HE, mientras que el método FE, de efectos fijos, presenta una mayor discrepancia con el resto. Sin embargo, las características específicas de los estudios y la posible presencia de fuentes de variabilidad diversas apuntan a la idoneidad de los modelos de efectos variables.

GO Procesos Biológicos						
Método	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra
DL	2769	5223	27	402	12	83
HE	2766	5223	27	404	12	85
HS	2756	5235	46	528	23	109
FE	2748	5247	270	987	65	170

Tabla 23. Número de términos GO “procesos biológicos” significativos en el metaanálisis con diferentes métodos.

GO Componentes Celulares						
Método	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra
DL	490	451	27	35	14	5
HE	490	451	28	34	15	5
HS	490	450	33	49	16	12
FE	494	448	108	93	34	23

Tabla 24. Número de términos GO "componentes celulares" significativos en el metaanálisis con diferentes métodos.

GO Funciones Moleculares						
Método	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra
DL	675	984	1	35	1	18
HE	675	984	1	36	1	20
HS	674	984	4	45	4	24
FE	671	989	66	119	26	37

Tabla 25. Número de términos GO "funciones moleculares" significativos en el metaanálisis con diferentes métodos.

KEGG						
Método	sobre	infra	sig.sobre	sig.infra	sig.o.sobre	sig.o.infra
DL	78	224	1	39	0	1
HE	78	224	1	39	0	1
HS	77	224	2	50	0	1
FE	77	225	7	75	1	1

Tabla 26. Número de términos KEGG significativos en el metaanálisis con diferentes métodos.

En la figura 13 se muestra el gráfico volcán con los resultados globales del metaanálisis, en el que cada punto representa una función. En este ejemplo, las funciones que superan un valor de 1.3 (-log10(0.05)) en el eje Y son resultados significativos. Se emplearon colores para facilitar la visualización, siendo los puntos verdes los que corresponden a las funciones infrarrepresentadas en el grupo de los pacientes respondedores y los puntos rojos las funciones sobrerrepresentadas en este grupo.

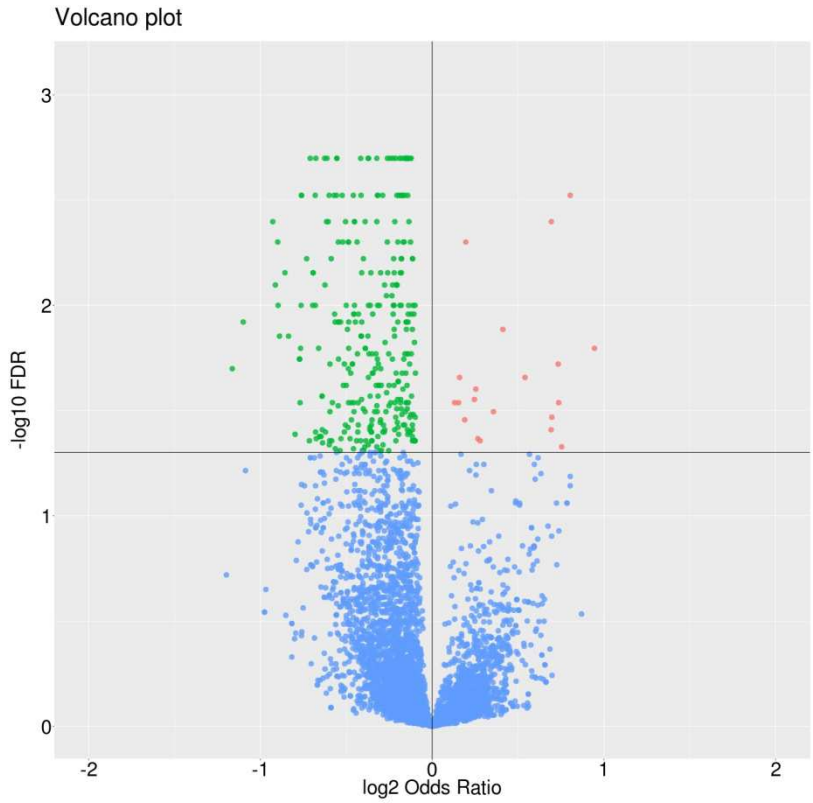


Figura 13: resultados del metaanálisis funcional de términos GO ”procesos biológicos”.

La visualización gráfica de las funciones diferencialmente representadas proporciona una visión global de las funciones significativas obtenidas en el análisis. Para las funciones GO se emplean gráficos en los que, además de detallar las relaciones entre las funciones, se indican cuáles de ellas se encuentran representadas de modo significativo. La figura 14 muestra una de estas gráfica de funciones GO en red, donde los nodos coloreados corresponden a las funciones sobrerrepresentadas en el grupo de pacientes que no responden al tratamiento. La figura Y es una ampliación de la anterior donde podemos observar que muchas de las funciones significativas

están relacionadas con la unión a receptores del sistema inmunológico como *Toll-like receptor binding*, *Interleukin receptor binding* o *interferon receptor activity*.

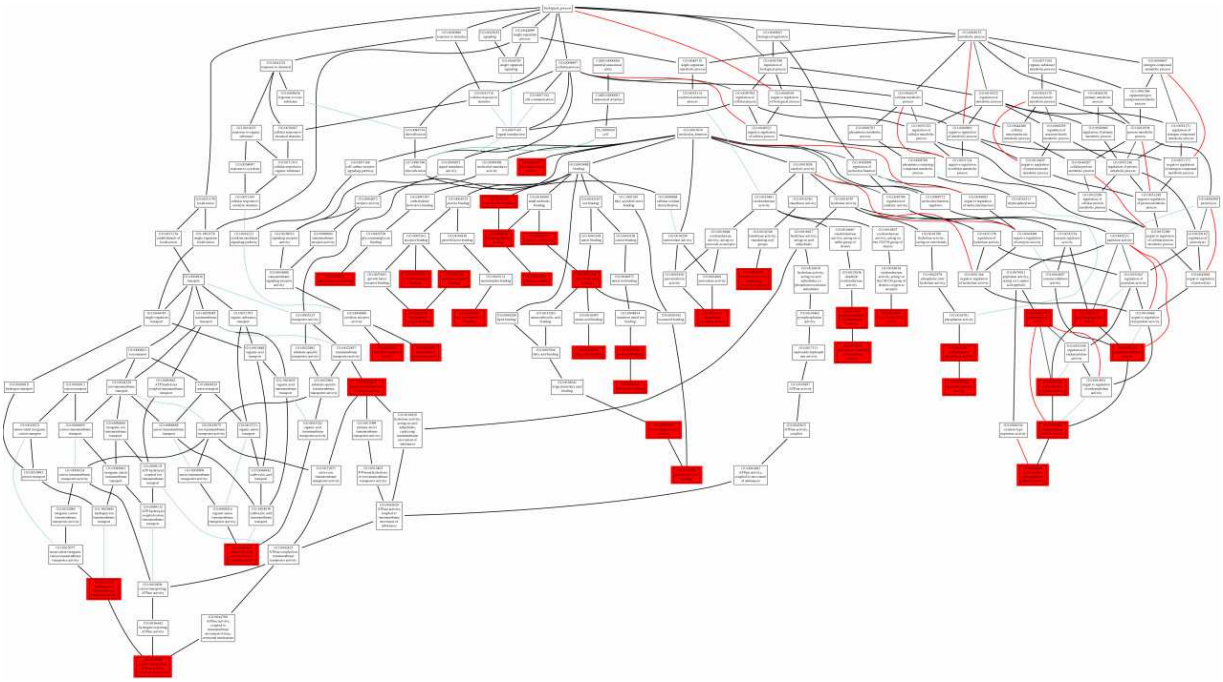


Figura 14. Procesos biológicos significativos con sobrerrepresentación en el grupo de pacientes no respondedores al tratamiento.

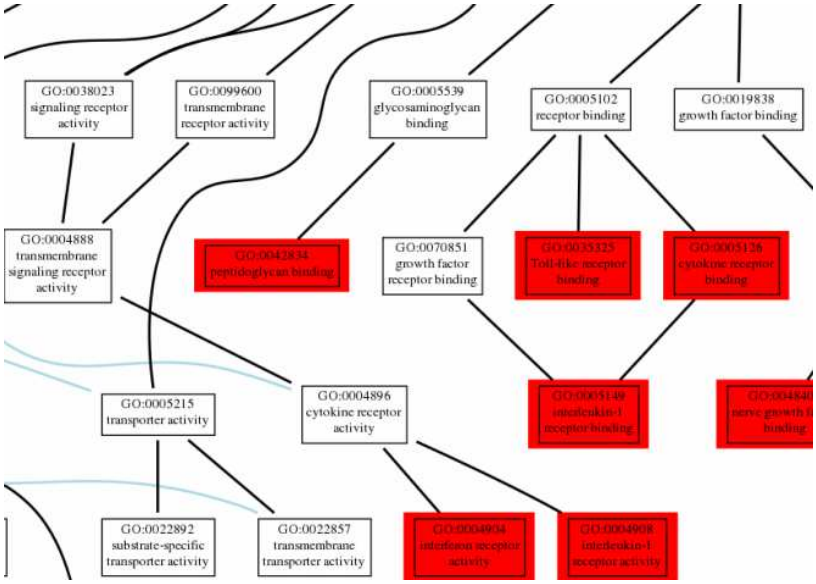


Figura 15. Detalle de la figura X, procesos biológicos significativos con sobrerrepresentación en el grupo de pacientes no respondedores al tratamiento.

4.2.2 Resultados específicos

Para cada función representada diferencialmente de modo significativo se realizó la evaluación del efecto a nivel de estudio y la heterogeneidad del metaanálisis que se representaron gráficamente para facilitar su interpretación.

Cada estudio aporta un efecto diferente en la estimación del efecto global para cada función. Para visualizar este efecto a nivel de estudio se realizaron gŕaficos de bosque, donde se muestra el peso de cada estudio en el conjunto de resultados. En la figura 16 se muestra el gráfico de bosque para el término GO:0002251 (*organ or tissue specific immune response*) donde se observa que la función GO:0002251 aparece diferencialmente representada de modo significativo en los tres estudio. El valor negativo de *odds ratio* indica que está función se encuentra sobrerrepresentada en el grupo experimental de pacientes no respondedores al tratamiento. Asimismo, se observa que tiene un mayor efecto sobre el metaanálisis el estudio GO_BP_25160, es decir, los datos de este estudio presentan un peso mayor en la estimación del efecto global.

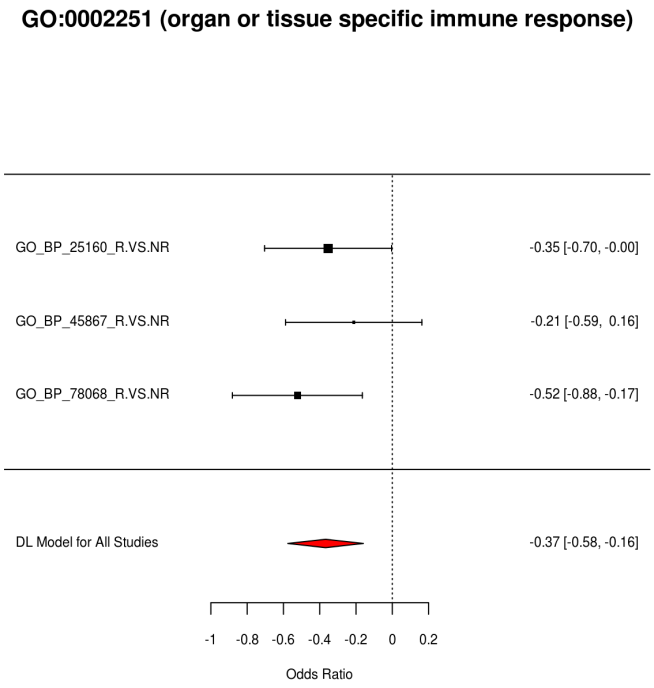


Figura 16. Distribución del efecto para la función GO:0002251

4.2.3 Estimación de la medida del efecto y análisis de heterogeneidad

Para cada uno de los métodos se recuperó una tabla con los estimadores de la medida del efecto en cada función estudiada, junto a indicadores de la heterogeneidad de los estudios incluidos en el metaanálisis. Las tablas 27 y 28 muestran estos indicadores para el término GO:0002251 tras su análisis con el método DL.

ID	nombre	QE	QEp	límite inferior	LOR	límite superior	SE	pvalue	p.adjust
GO:0002251	organ or tissue specific immune response	1.385	0.5	-0.576	-0.368	-0.159	0.106	0.001	0.017

Tabla 27. Estimadores de la medida del efecto para la función GO:0002251.

ID	nombre	tau^2	H^2	I^2
GO:0002251	organ or tissue specific immune response	0	0	1

Tabla 28. Indicadores de heterogeneidad para la función GO:0002251

- QE y QEp: son el estadístico de contraste y el *valor p* en el método DL. Se emplea para detectar heterogeneidad entre los estudios, en el que la hipótesis nula es la homogeneidad de los estudios, con un nivel de confianza de 95%. La detección de heterogeneidad confirma la necesidad de emplear un modelo de efectos aleatorios.

- LOR: estimación del efecto combinado de todos los estudios. Un valor positivo indica mayor cantidad de genes con un nivel alto de expresión en la primera clase experimental, en este caso los pacientes respondedores al tratamiento. La magnitud indica la sobrerrepresentación de la función.

- SE: intervalo de confianza, va desde el límite inferior hasta el límite superior. Si su valor no es 0 confirma la significatividad de LOR.

- p-valor: indica el nivel de significación pero no contempla la multiplicidad del metanálisis.

- p-valor ajustado: con una corrección del p-valor se tiene en cuenta esta multiplicidad y se evitan falsos positivos.

- Estimadores de la heterogeneidad: tau^2 es 0 cuando se emplea un método de efectos fijos, ya que estima la heterogeneidad entre los estudios individuales. I^2 es la relación entre la variabilidad de los estudios y la variabilidad total. H^2 hace referencia al cociente entre la variabilidad total y la variabilidad en el muestreo.

Los diagramas de cajas son muy útiles para conocer la distribución de los indicadores de todas las funciones para cada método en solo un vistazo, y así escoger el método que mejor se ajuste a los datos estudiados. En el anejo 2 se presentan los diagramas de cajas de los resultados de los indicadores de heterogeneidad del análisis de funciones GO "Procesos Biológicos".

Gráficos indicadores de la heterogeneidad

- Gráfico de embudo: la figura 17 muestra gráficos de embudo para la función GO:0002251. En todos ellos se observa que los estudios individuales se encuentran dentro del embudo por lo que no hay sesgos y la heterogeneidad es pequeña, no despuntando en ningún estudio.

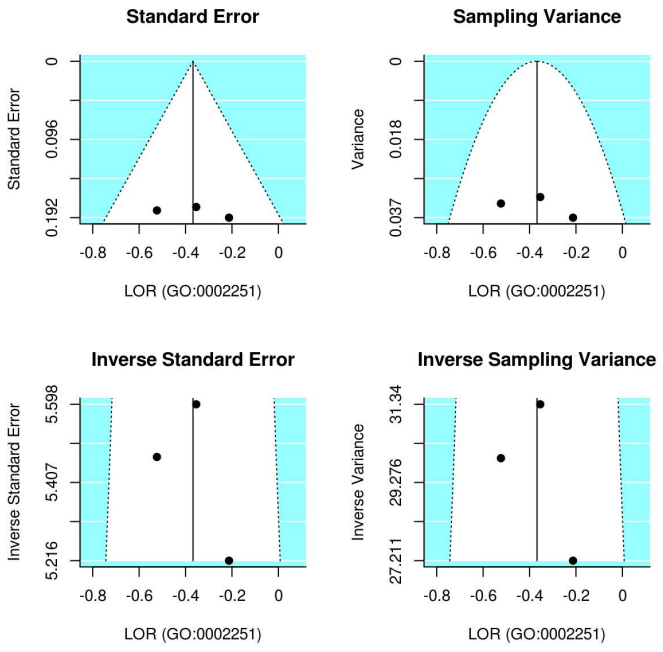


Figura 17. Variabilidad del efecto estudiado de la función GO:0002251

- Gráfico radial: la figura 18 muestra el gráfico radial correspondiente al análisis de la función GO:0002251 en el que se observa que los tres estudios no muestran sesgos y su heterogeneidad no impide su análisis conjunto.

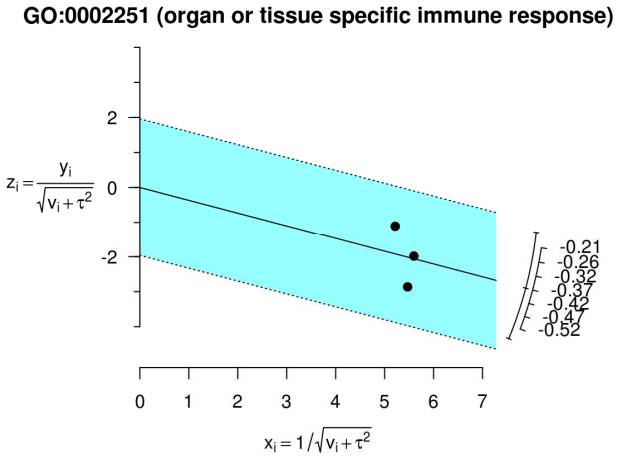


Figura 18. Variabilidad del efecto estudiado en la función GO:0002251

Análisis de estudios influyentes

Algunos de los estudios individuales pueden tener un mayor peso y producir una fuerte influencia en el metaanálisis. La detección de datos influyentes se realiza junto la regresión logística (Belsey *et al.* 1980; Cook & Weisberg 1982). Existen diferentes medidas de detección de datos influyentes. En la figura 19 se muestran las medidas de detección de datos influyentes para la función GO:0002251.

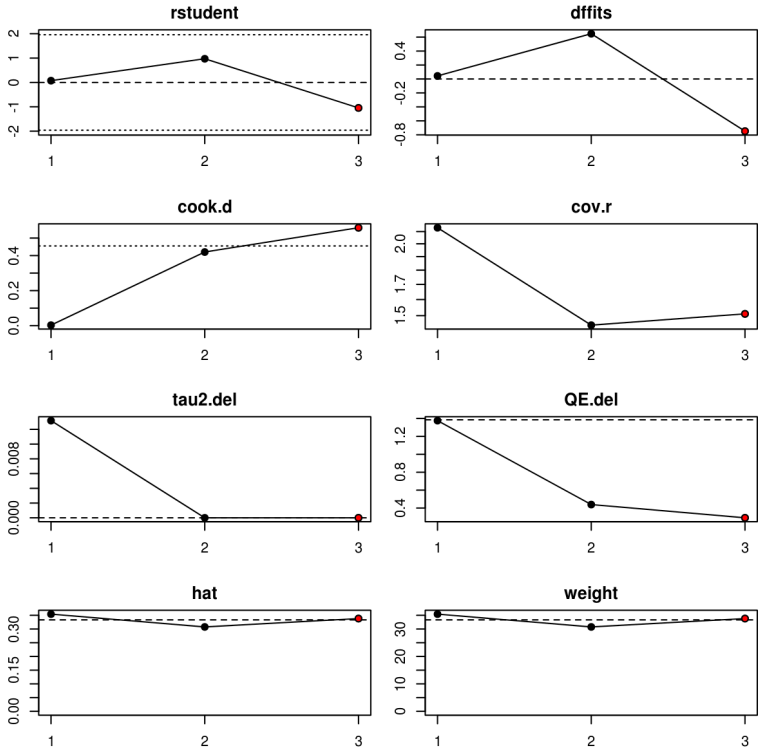


Figura 19. Análisis de estudios influyentes para la función GO:0002251

Análisis de sensibilidad

Para cada función representada diferencialmente de modo significativo se realizó un análisis de sensibilidad con el fin de evaluar la influencia de cada estudio individual. El método consiste en reproducir el metaanálisis excluyendo el estudio que se revisa. La sensibilidad se determinó a través de indicadores de la estimación del efecto y su variabilidad. En la tabla 29 se muestran los resultados para la función GO:0002251. En el anejo 3 se presentan los diagramas de cajas que representan la medida del efecto de cada estudio según su *odds ratio* y su error estándar.

estudios	lor	se	zval	pval	ci.lb	ci.ub	Q	Qp	tau2	I2	H2
GSE25160	-0.274	0.202	-1.356	0.175	-0.671	0.122	0.696	0.404	0	0	1
GSE45867	-0.351	0.202	-1.737	0.082	-0.747	0.045	0.199	0.656	0	0	1
GSE78068	-0.181	0.204	-0.885	0.376	-0.581	0.22	0.148	0.7	0	0	1

Tabla 29. Estimadores de la medida del efecto e indicadores de heterogeneidad en el metaanálisis funcional con procesos biológicos.

4.3 Discusión

La selección de genes para el diseño de un panel génico es una tarea compleja que además está sujeta a la optimización de los recursos del estudio. Los análisis *in silico* ayudan a seleccionar y priorizar genes de interés con mayor robustez, sin un coste económico importante y disminuyendo el tiempo de obtención de resultados, en contraposición a los estudios *in vivo*. Los resultados obtenidos a partir de la integración funcional de datos de diferentes estudios con similares diseños experimentales, aporta nueva información útil en la selección de genes de interés ya que el número muestral es mayor y los resultados más rigurosos, disminuyendo la tasa de falsos positivos.

La artritis reumatoide es una enfermedad de carácter autoinmune, inflamatoria y crónica en la que se ha observado un aumento importante de IL-6 que es una citocina proinflamatoria. El tratamiento con anticuerpos anti-IL-6 va dirigido a reducir la inflamación por lo que cabe esperar que las funciones diferencialmente representadas entre los pacientes que responden al tratamiento y los que no lo hacen, deban estar relacionadas con la inflamación. Este abordaje bioinformático demuestra que las funciones que aparecen representadas de manera diferencial entre ambos grupos de pacientes están fuertemente relacionadas con el sistema inmunológico, lo que indica que en cada grupo de pacientes la inflamación está sostenida por diferentes ambientes inflamatorios en los que actúan diferentes proteínas. Los genes relacionados con estas funciones son los candidatos seleccionados para el diseño del panel de genes de respuesta a este fármaco.

ID	Nombre	LogOR	p.ajustado
GO:0032747	positive regulation of interleukin-23 production	-0.889	0.014
GO:0072672	neutrophil extravasation	-0.857	0.007
GO:1900017	positive regulation of citokyne production involved in inflammatory response	-0.680	0.010
GO:0034154	toll-like receptor 7 signaling pathway	-0.766	0.016

Tabla 30. Funciones sobrerrepresentadas en el grupo de pacientes no respondedores al tratamiento con anti-IL-6.

En la tabla 30 se indican funciones sobrerrepresentadas en el grupo de pacientes no respondedores al tratamiento con anti-IL-6. Son funciones relacionadas con la

inflamación que es la causante de la enfermedad. Estas funciones son las que se utilizarían en la selección de genes candidatos para el diseño de un panel genético de respuesta a anti-IL-6 (tabla 31).

Relación funciones - genes			
GO:0032747 IL17RA, CSF2, IFNG, MYD88	GO:0072672 TREM1, PIK3CD, JAML, PIK3CG	GO:1900017 IL17RA, IL17A, IL17B, NOD2, IL17F, GBP5, CD6, IL17RC, GPSM3	GO:0034154 HAVCR2, TLR7, UNC93B1, PIK3AP1

Tabla 31. Lista de genes involucrados en cada función.

Tras la priorización y selección de genes candidatos es necesaria su valoración por parte de los investigadores implicados en el estudio y su validación a través de estudios *in vivo*, a partir de los cuales se llegará al diseño final del panel genético.

La creación de un panel genético de respuesta al fármaco anti-IL-6, también podría emplearse en otras enfermedades crónicas asociadas a inflamación, como la enfermedad de Castleman, enfermedad inflamatoria intestinal, diabetes…, siendo capaces de detectar nuevos pacientes que puedan beneficiarse de este fármaco.

5 Conclusiones

Los repositorios de datos públicos nos ofrecen un gran volumen de información transcriptómica a partir de la cual se pueden realizar nuevos abordajes y combinaciones con el objetivo de responder preguntas de investigación de interés y aportar nuevo conocimiento científico

Los análisis de expresión diferencial no siempre producen resultados significativos, sin embargo esta situación no supone una limitación en la aplicación de abordajes como el enriquecimiento funcional o la integración a nivel de función.

El enriquecimiento funcional es una herramienta muy útil para conocer qué funciones se encuentran diferencialmente representadas entre dos grupos de modo significativo, independientemente de si el análisis de expresión diferencial ha aportado resultados significativos.

La integración de distintos niveles de información biológica mejora la interpretación de resultados y la potencia del estudio.

La integración de datos de ARNm y miARN procedente de muestras de adenocarcinoma de colon en sus cuatro estadios, ha permitido una mejor caracterización molecular de esta enfermedad, proporcionado resultados funcionales ajustados por la inclusión de ambos niveles de información biológica

El metaanálisis funcional de datos transcriptómicos de pacientes con artritis reumatoide ha proporcionado la relación de funciones más activas en los pacientes que responden al tratamiento con anti-IL-6, a partir de las cuales se han obtenido y priorizado genes candidatos a formar parte de un panel genético de respuesta a tratamiento

6 Bibliografía

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

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

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

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

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

Chen, D.T. *et al.,* "Complementary strand microRNAs mediate acquisition of metastatic potential in colonic adenocarcinoma.", *J Gastrointest Surg*, 2012 Feb 24;16(5):905-12; discussion 912-3

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

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

Ducreux, J. et al., 2014. Global Molecular Effects of Tocilizumab Therapy in Rheumatoid Arthritis Synovium. Arthritis & Rheumatology. Vol. 66, pp 15–23. DOI 10.1002/art.38202

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

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

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

Garcia-Garcia F, Panadero J, Dopazo J, Montaner D. Integrated gene set analysis for microRNA studies. Bioinformatics. 2016 Sep 15;32(18):2809-16. doi: 10.1093/bioinformatics/btw334. Epub 2016 Jun 20.

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

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

Irizarry, RA; Hobbs, B; Collin, F; Beazer-Barclay, YD; Antonellis, KJ; Scherf, U; Speed, TP (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data.". Biostatistics. 4 (2): 249–64. doi:10.1093/biostatistics/4.2.249.PMID 12925520.

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

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

Mesko, B. *et al*., 2012. Peripheral Blood Gene Expression and IgG Glycosylation Profiles as Markers of Tocilizumab Treatment in Rheumatoid Arthritis. The Journal of Rheumatology 2012; 39:5; doi:3899/jrheum.110961

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

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

Nakamura *et al*., 2016. Identification of baseline gene expression signatures predicting therapeutic responses to three biologic agents in rheumatoid arthritis: a retrospective

observational study. Arthritis Research & Therapy (2016) 18:159 DOI 10.1186/s13075-016-1052-8

Raja *et al*., 2017. A Review of Recent Advancement in Integration Omics Data with Literature Mining towards Biomedical Discoveries. International Journal of Genomics. ID 6213474

Sartor, M.A. & others, 2009. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics, 25(2), pp.211–217.

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

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

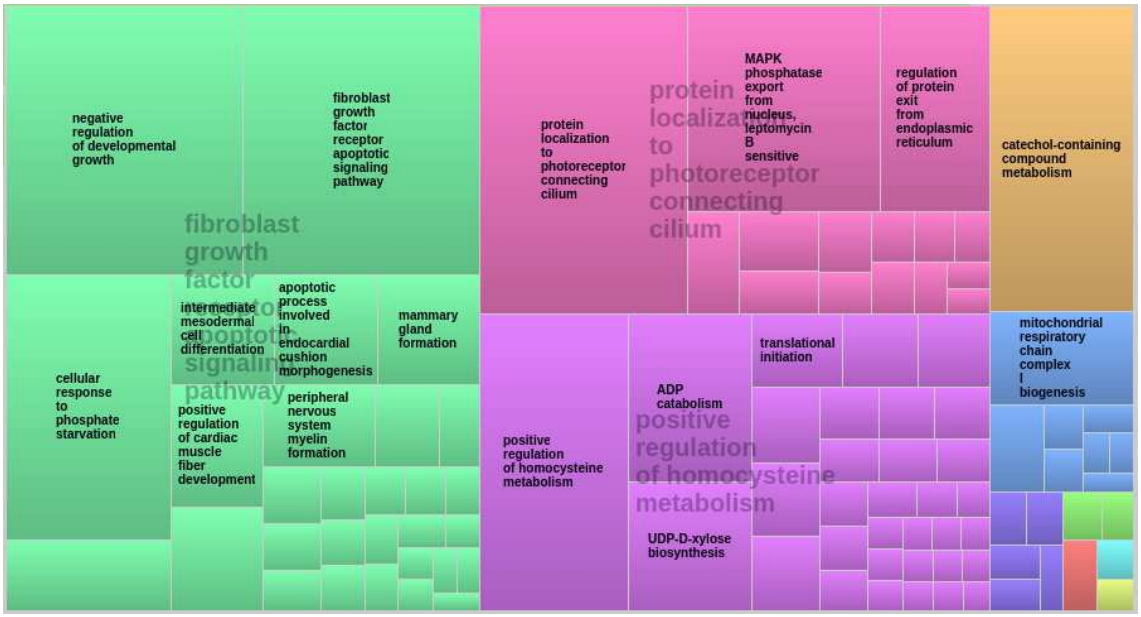
Sterne, J.A. & Egger, M., 2001. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. Journal of clinical epidemiology, 54(10), pp.1046–1055.

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

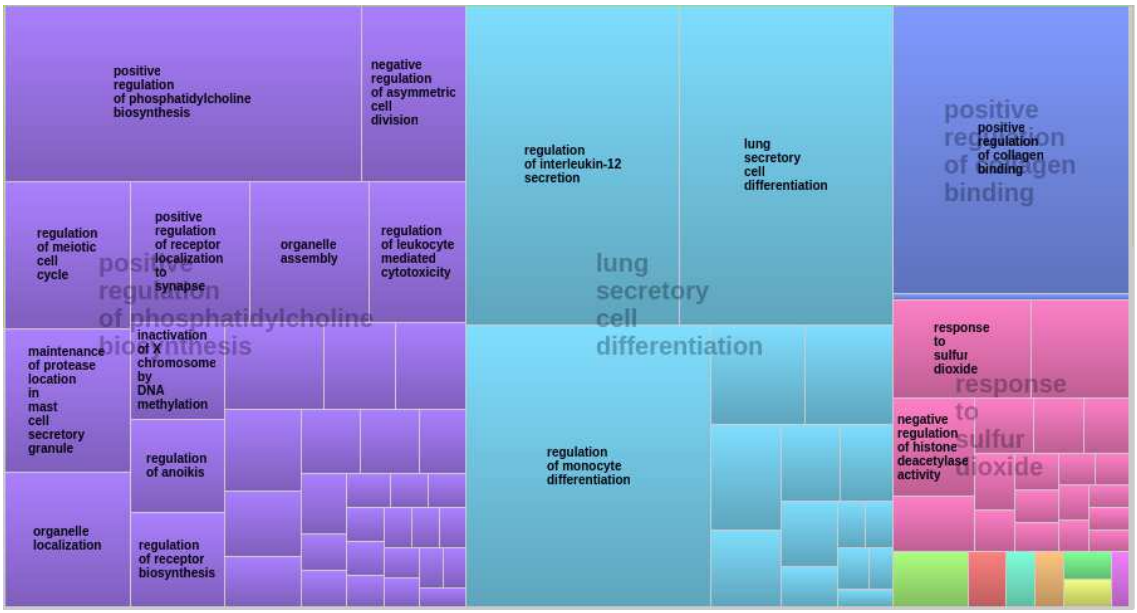
7 Anejos

ANEJO 1: *Treemaps* de la comparación a nivel funcional de adenocarcinoma de colon en estadio I frente a estadio II, para datos de ARNm, miARN e integración ARNm-miARN.

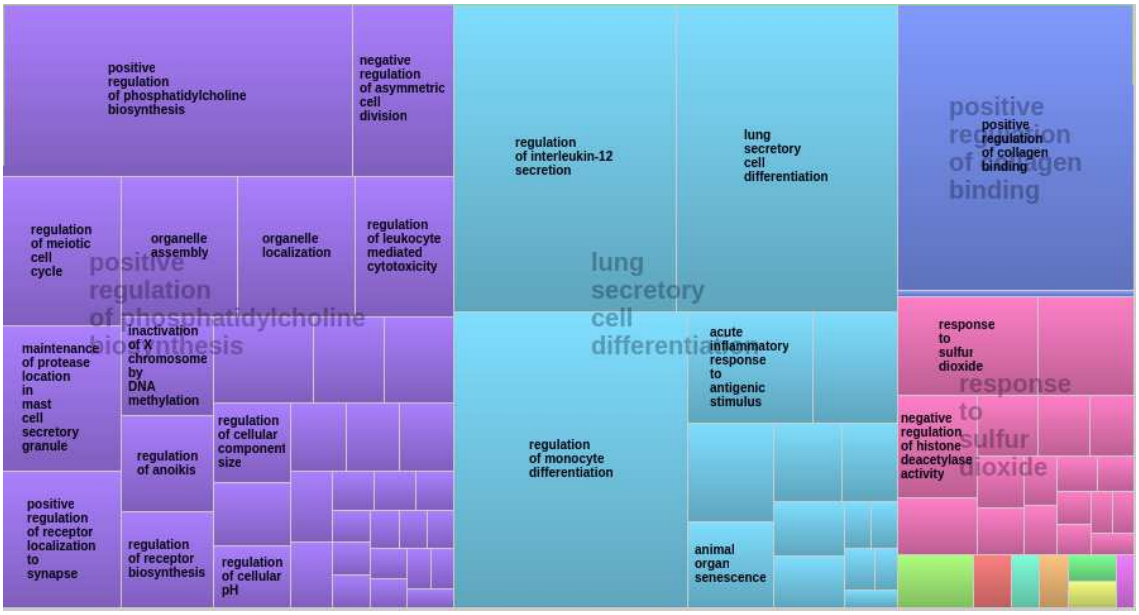
ARNm: funciones sobrerrepresentadas en el estadio I



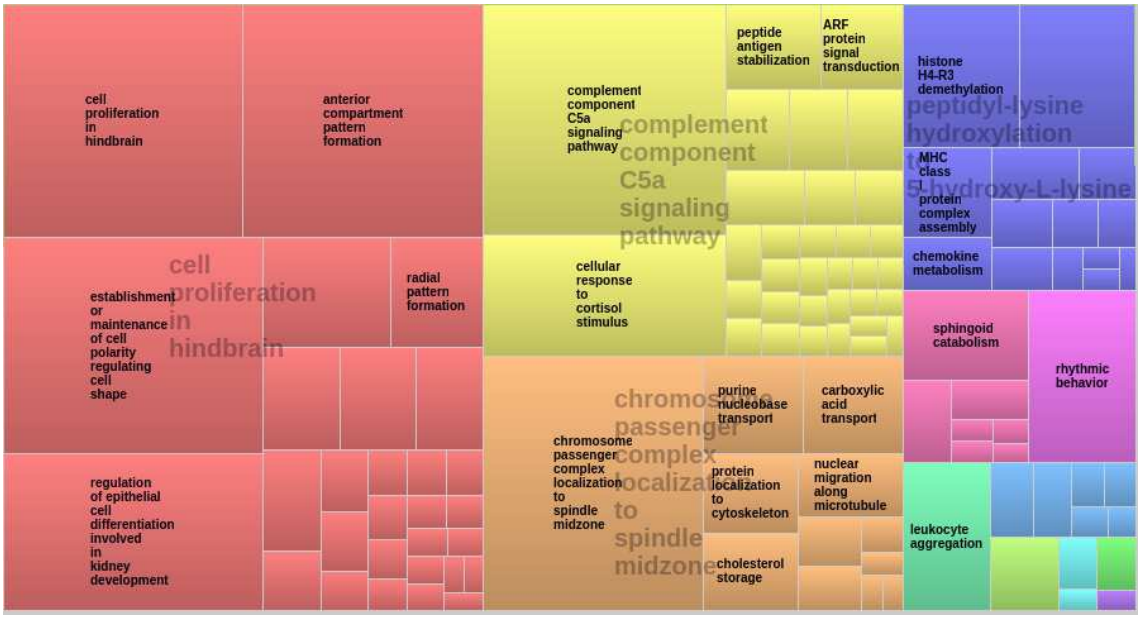
miARN: funciones sobrerrepresentadas en el estadio I



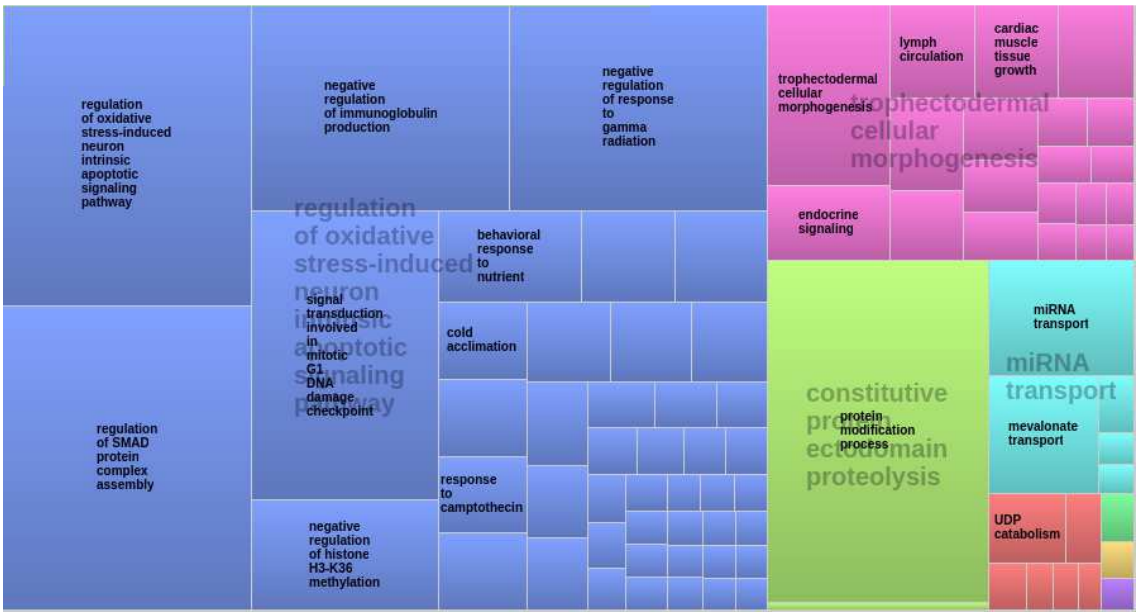
Integración ARNm-miARN: funciones sobrerrepresentadas en el estadio I



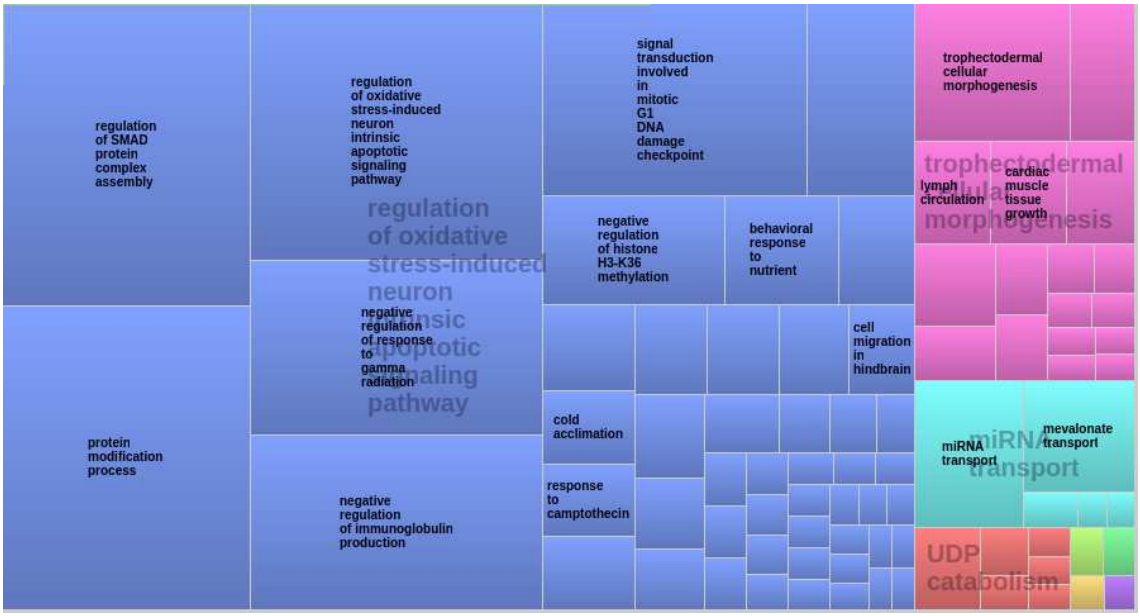
ARNm: funciones sobrerrepresentadas en el estadio II



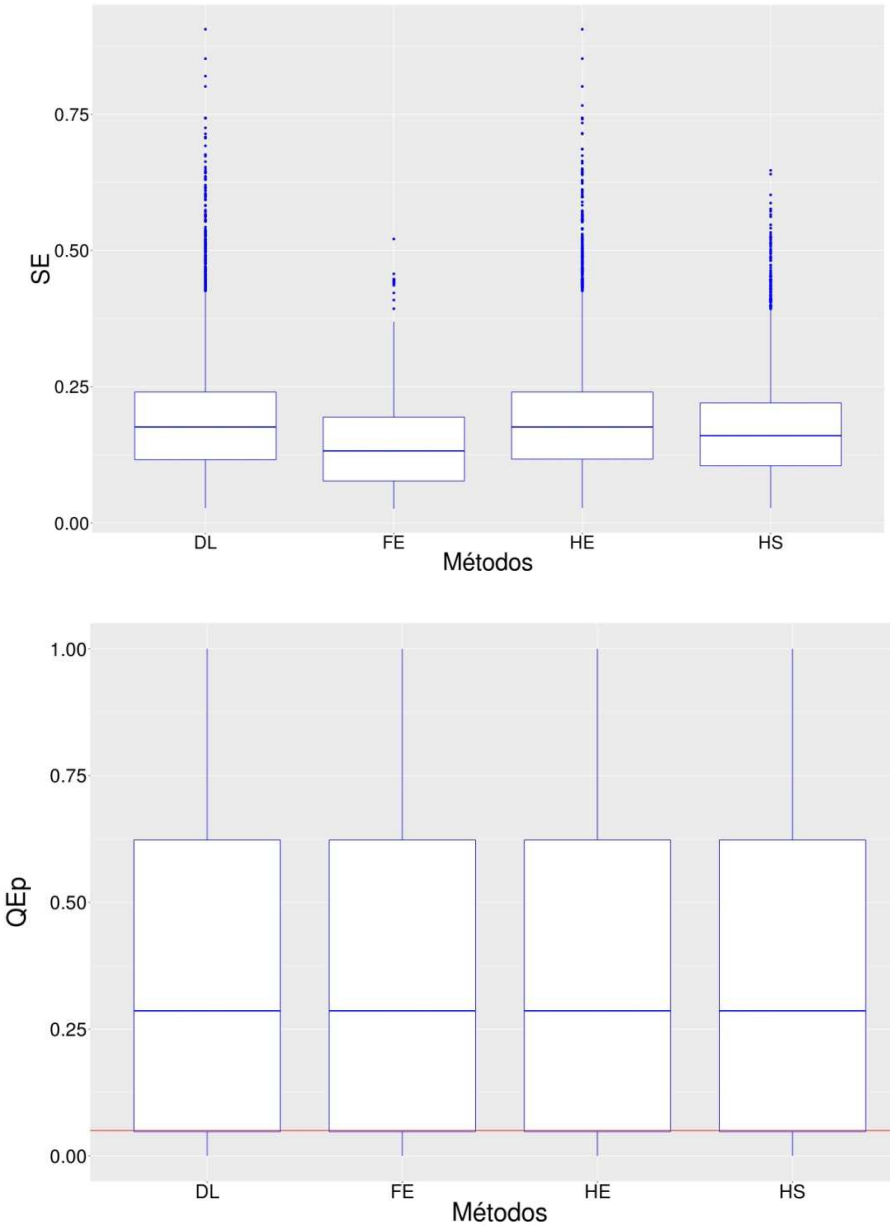
miARN: funciones sobrerrepresentadas en el estadio II

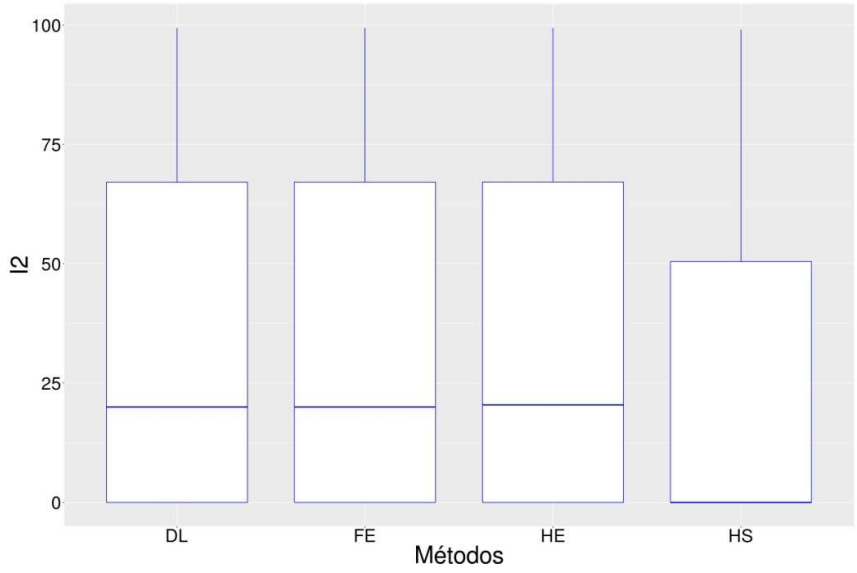


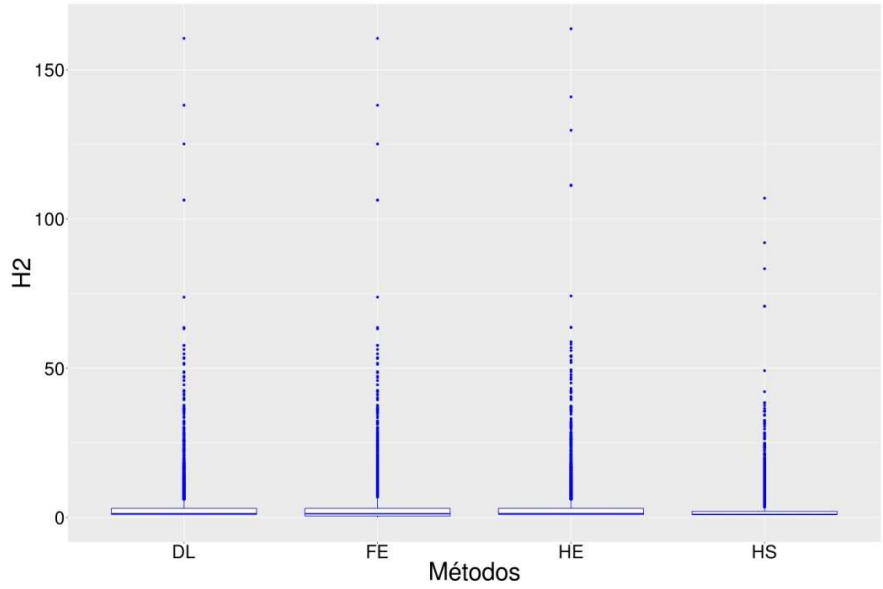
Integración ARNm-miARN: funciones sobrerrepresentadas en el estadio II

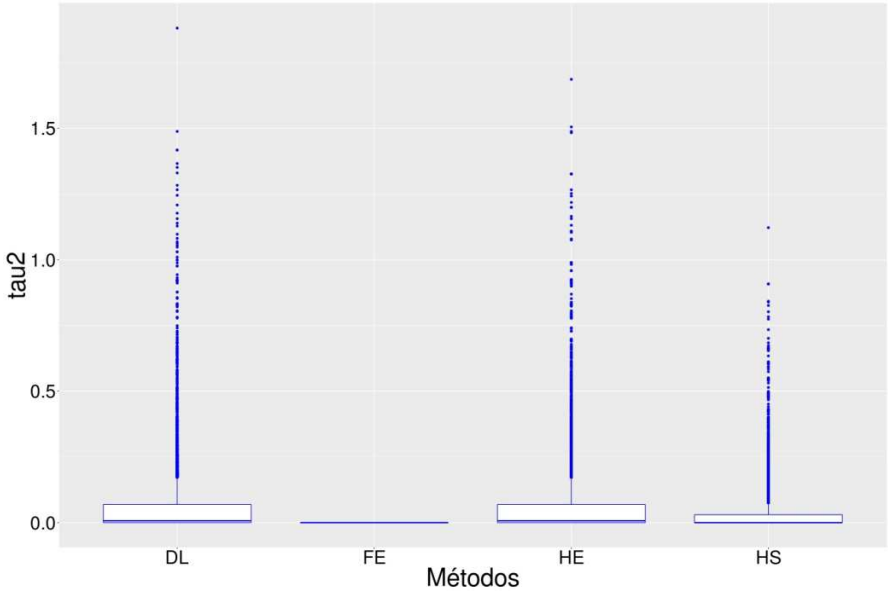


ANEJO 2: Diagramas de cajas de indicadores de heterogeneidad para los metaaanálisis referidos a términos GO “Procesos Biológicos”.




100
75
50
25
0
I2
DL
FE
HE
HS
Métodos


150
100
50
0
H2
DL
FE
HE
HS
Métodos


1.5
1.0
0.5
0.0
tau2
DL
FE
HE
HS
Métodos

ANEJO 3: Diagramas de cajas para la comparación del metaanálisis referido a los términos GO “Procesos Biológicos” según su *Odds Ratio* y su error estándar.