

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA



VNIVERSITAT
ID VALÈNCIA

TRABAJO DE FIN DE MÁSTER

METAANÁLISIS FUNCIONAL DE LAS DIFERENCIAS DE SEXO EN ESTUDIOS ÓMICOS DE ADENOCARCINOMA DE PULMÓN

**AUTORA:
IRENE PÉREZ DÍEZ**

**TUTORES:
FRANCISCO GARCÍA GARCÍA
MARTA HIDALGO GARCÍA
GUILLERMO AYALA GALLEGO**

MAYO, 2019

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

METAANÁLISIS FUNCIONAL DE LAS DIFERENCIAS DE SEXO EN ESTUDIOS ÓMICOS DE ADENOCARCINOMA DE PULMÓN

**AUTORA:
IRENE PÉREZ DÍEZ**

**TUTORES:
FRANCISCO GARCÍA GARCÍA
MARTA HIDALGO GARCÍA
GUILLERMO AYALA GALLEGO**

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

Resumen

El cáncer de pulmón es el tipo de cáncer más frecuente y el más mortal, y dentro de esta patología el subtipo más frecuente es el adenocarcinoma de pulmón. Al igual que en otros grupos de enfermedades, como en las enfermedades autoinmunes, el adenocarcinoma de pulmón presenta diferencias entre hombres y mujeres a nivel epidemiológico, con una mayor incidencia en mujeres y una mayor mortalidad en hombres.

Entender los mecanismos moleculares subyacentes a estas diferencias podría abrir la puerta a un tratamiento más personalizado de los pacientes. Con este objetivo en mente, en este trabajo hemos realizado un metaanálisis combinando la información de una selección de estudios transcriptómicos. Los resultados obtenidos apuntan a una sobreactivación de la inflamación y la respuesta inmunitaria en mujeres, mientras que en hombres parece haber un mayor crecimiento celular.

Palabras clave: adenocarcinoma de pulmón, sexo, transcriptómica, metaanálisis.

Abstract

Lung cancer is the type of cancer with the highest rates of incidence and mortality. The most frequent subtype of lung cancer is lung adenocarcinoma. The same way as other pathologies, such as autoimmune diseases, lung adenocarcinoma displays epidemiological differences between male and female, showing a higher incidence rate in females and a higher mortality rate in males.

Understanding the molecular mechanisms underlying these differences could open the door to a precision medicine approach to treat the patients. With this aim, we have carried out a meta-analysis combining information from a selection of transcriptomics studies. Our results point out to a superactivation of inflammation and immune response in females, while in males there seems to be a higher cellular growth.

Keywords: lung adenocarcinoma, sex, transcriptomics, meta-analysis.

Índice general

1. Introducción	1
1.1. Cáncer de pulmón	1
1.1.1. Epidemiología y diferencias entre sexos	1
1.1.2. Bases moleculares del adenocarcinoma de pulmón	4
1.1.3. Microambiente tumoral y respuesta inmunitaria	6
1.2. Datos ómicos y tecnologías de alto rendimiento	7
1.2.1. Transcriptómica	9
1.2.2. Análisis transcriptómicos	9
1.2.3. Caracterización funcional	10
1.2.4. Metaanálisis funcional	11
2. Hipótesis	13
3. Objetivos	15
4. Material y métodos	17
4.1. Revisión sistemática y selección de estudios	17
4.2. Análisis individual de los estudios	18
4.2.1. Procesamiento de los datos	18
4.2.2. Análisis de expresión diferencial	20
4.2.3. Análisis de enriquecimiento funcional	21
4.2.4. Análisis de rutas de señalización	21
4.3. Metaanálisis	22
4.3.1. Determinación de la medida combinada del efecto	22
4.3.2. Evaluación de la heterogeneidad	24
4.3.3. Representación de resultados	24
5. Resultados	27
5.1. Revisión sistemática y selección de estudios	27
5.2. Análisis individual de los estudios	28
5.2.1. Procesamiento de los datos	28
5.2.2. Análisis de expresión diferencial	36
5.2.3. Análisis de enriquecimiento funcional	37
5.2.4. Análisis de rutas de señalización	40

5.3. Metaanálisis	41
5.3.1. Términos GO para procesos biológicos	41
5.3.2. Términos GO para funciones moleculares	49
5.3.3. Rutas KEGG	53
6. Discusión	57
7. Conclusiones	63
Bibliografía	65
A. Anexo I - Figuras	73
B. Anexo B - Tablas	87

Índice de figuras

1.1.	Incidencia y mortalidad del cáncer en ambos sexos	2
1.2.	Incidencia del cáncer de pulmón por sexo y área geográfica	3
1.3.	Alteraciones en las rutas de señalización en adenocarcinoma de pulmón	5
1.4.	El microambiente en cáncer de pulmón	7
1.5.	Dogma central de la biología molecular ampliado	8
4.1.	Flujo de información a través de las diferentes fases de una revisión sistemática	18
4.2.	Ejemplo de un gráfico de bosque	25
5.1.	Diagrama de flujo de la revisión sistemática, siguiendo la declaración PRISMA	28
5.2.	Número de pacientes en cada estudio	29
5.3.	Clustering exploratorio de las muestras del estudio GSE19188 mediante distancia de correlación	31
5.4.	Análisis de componentes principales de las muestras del estudio GSE19188 .	32
5.5.	Clustering exploratorio de las muestras del estudio GSE31210 mediante distancia de correlación	33
5.6.	Análisis de componentes principales de las muestras del estudio GSE31210 .	34
5.7.	Diagrama UpSet de los términos GO BP y MF	39
5.8.	Diagrama UpSet de los términos GO CC y las rutas KEGG	40
5.9.	Número de funciones y rutas significativas	41
5.10.	Gráfico de bosque para el término GO:0002438	44
5.11.	Gráfico de bosque para el término GO:0002863	44
5.12.	Gráfico de bosque para el término GO:0043374	45
5.13.	Gráfico de bosque para el término GO:0045410	45
5.14.	Gráfico de bosque para el término GO:0006183	46
5.15.	Gráfico de bosque para el término GO:0006228	46
5.16.	Gráfico de bosque para el término GO:1902857	47
5.17.	Gráficos de embudo para el término GO:0002438	47
5.18.	Gráficos de embudo para el término GO:0043374	48
5.19.	Resumen de los términos GO BP significativos en el metaanálisis	48
5.20.	Gráfico de bosque para el término GO:0001614	51
5.21.	Gráfico de bosque para el término GO:0004126	51
5.22.	Gráfico de bosque para el término GO:0045028	52
5.23.	Gráfico de bosque para el término GO:0048407	52
5.24.	Gráfico de bosque para la ruta KEGG hsa:04120	55

5.25. Gráfico de bosque para la ruta KEGG hsa:04330	55
5.26. Gráfico de bosque para la ruta KEGG hsa:04622	56
5.27. Gráfico de bosque para la ruta KEGG hsa:04660	56
A.1. Clustering exploratorio de las muestras del estudio GSE10072 mediante distancia de correlación	73
A.2. Análisis de componentes principales de las muestras del estudio GSE10072 .	73
A.3. Clustering exploratorio de las muestras del estudio GSE32863 mediante distancia de correlación	74
A.4. Análisis de componentes principales de las muestras del estudio GSE32863 .	74
A.5. Clustering exploratorio de las muestras del estudio GSE63459 mediante distancia de correlación	75
A.6. Análisis de componentes principales de las muestras del estudio GSE63459 .	75
A.7. Clustering exploratorio de las muestras del estudio GSE75037 mediante distancia de correlación	76
A.8. Análisis de componentes principales de las muestras del estudio GSE75037 .	76
A.9. Clustering exploratorio de las muestras del estudio GSE81089 mediante distancia de correlación	77
A.10. Análisis de componentes principales de las muestras del estudio GSE81089 .	77
A.11. Clustering exploratorio de las muestras del estudio GSE87340 mediante distancia de correlación	78
A.12. Análisis de componentes principales de las muestras del estudio GSE87340 .	78
A.13. Clustering exploratorio de las muestras de TCGA-LUAD mediante distancia de correlación	79
A.14. Análisis de componentes principales de las muestras de TCGA-LUAD	79
A.15. Gráficos de embudo para el término GO:0002863	80
A.16. Gráficos de embudo para el término GO:0045410	80
A.17. Gráficos de embudo para el término GO:0006183	81
A.18. Gráficos de embudo para el término GO:0006228	81
A.19. Gráficos de embudo para el término GO:1902857	82
A.20. Gráficos de embudo para el término GO:0001614	82
A.21. Gráficos de embudo para el término GO:0004126	83
A.22. Gráficos de embudo para el término GO:0045028	83
A.23. Gráficos de embudo para el término GO:0048407	84
A.24. Gráficos de embudo para la ruta hsa:04210	84
A.25. Gráficos de embudo para la ruta hsa:04330	85
A.26. Gráficos de embudo para la ruta hsa:04622	85
A.27. Gráficos de embudo para la ruta hsa:04660	86

Índice de tablas

5.1.	Estudios seleccionados tras la revisión sistemática	27
5.2.	Individuos incluidos en el estudio GSE10072	29
5.3.	Individuos incluidos en el estudio GSE19188	30
5.4.	Individuos incluidos en el estudio GSE31210	30
5.5.	Individuos incluidos en el estudio GSE32863	31
5.6.	Individuos incluidos en el estudio GSE63459	32
5.7.	Individuos incluidos en el estudio GSE75037	33
5.8.	Individuos incluidos en el estudio GSE81089	34
5.9.	Individuos incluidos en el estudio GSE87340	35
5.10.	Individuos incluidos en TCGA-LUAD	35
5.11.	Resumen del resultado del análisis de expresión diferencial	36
5.12.	Resumen del resultado del análisis de enriquecimiento funcional	38
5.13.	Resultados del metaanálisis para los términos GO BP	42
5.14.	Selección de términos GO BP significativos en el metaanálisis	43
5.15.	Resultados del metaanálisis para los términos GO MF	49
5.16.	Términos GO MF significativos en el metaanálisis	50
5.17.	Resultados del metaanálisis para las rutas KEGG	53
5.18.	Selección de rutas KEGG significativas en el metaanálisis	54
B.1.	Versiones del software empleado en este trabajo	87
B.2.	Términos GO BP significativos en el metaanálisis	88
B.3.	Rutas KEGG significativas en el metaanálisis	93

Acrónimos

ADN ácido desoxirribonucleico.

ARN ácido ribonucleico.

ATP adenosín trifosfato.

BH Benjamini y Hochberg.

BP procesos biológicos.

CC componentes celulares.

CCAA *Canonical Circuit Activity Analysis*.

DL DerSimonian y Laird.

FAIR *Findable, Accesible, Interoperable, Reusable*.

FE efectos fijos.

GDC *Genomic Data Commons*.

GEO *Gene Expression Omnibus*.

GO *Gene Ontology*.

GSEA *Gene Set Enrichment Analysis*.

GTP guanosín trifosfato.

HE Hedges.

HS Hunter y Schmidt.

IFN- γ interferón gamma.

IL interleucina.

KEGG *Kyoto Encyclopedia of Genes and Genomes.*

logFC logaritmo de la magnitud de cambio.

LOR logaritmos de los *odds ratio*.

MF funciones moleculares.

NK *natural killer.*

NSCLC *non-small cell lung cancer.*

PCA análisis de componentes principales.

RNA-seq secuenciación de ARN.

SCLC *small cell lung cancer.*

TCGA *The Cancer Genome Atlas.*

TME microambiente tumoral.

TMM *Trimmed Mean of M-values.*

TNF- α factor de necrosis tumoral alfa.

UTP uridina trifosfato.

1. Introducción

El cáncer es la segunda causa de muerte a nivel global, habiendo provocado 9.6 millones de muertes en 2018. Dentro de este marco, el cáncer más común es el de pulmón, con 2.09 millones de casos en 2018, y también el más mortal, con 1.76 millones de muertes [1]. A nivel epidemiológico se han descrito diferencias en el cáncer de pulmón entre hombres y mujeres, que parecen acentuarse en el subtipo adenocarcinoma [2]. Dada la importancia de la enfermedad y la disparidad epidemiológica entre sexos, en este trabajo nos proponemos analizar mediante el metaanálisis de estudios transcriptómicos el contexto molecular de las diferencias debidas al sexo en adenocarcinoma de pulmón.

1.1. Cáncer de pulmón

El cáncer es un conjunto de enfermedades en las que células anormales (que han dejado de funcionar como deberían) proliferan sin control y pueden invadir los tejidos cercanos, formando tumores malignos [3]. La mayoría de los cánceres humanos surgen a partir de los tejidos epiteliales, clasificándose como carcinomas. Estos tejidos están formados por capas de células que delimitan las paredes de las cavidades y canales corporales, además de servir como recubrimiento del cuerpo. Los carcinomas que surgen de las células que forman las capas protectoras del epitelio reciben el nombre de carcinomas de células escamosas, mientras que aquellos que surgen de las células secretoras se denominan adenocarcinomas [4].

El cáncer de pulmón es un tipo de cáncer heterogéneo que tradicionalmente se ha dividido en dos principales subtipos: cáncer de pulmón de células pequeñas, *small cell lung cancer* (SCLC), y cáncer de pulmón de células no pequeñas, *non-small cell lung cancer* (NSCLC). El NSCLC es el tipo de cáncer de pulmón más frecuente, y presenta dos subtipos predominantes: adenocarcinoma y carcinoma de células escamosas [5].

1.1.1. Epidemiología y diferencias entre sexos

El cáncer está entre las dos primeras causas de muerte antes de los 70 años de edad en 91 de 172 países analizados, y se espera que durante este siglo se convierta en la principal causa de muerte en todos los países del mundo. Dentro de este escenario, el cáncer de pulmón es el cáncer

que presenta una mayor incidencia (número de nuevos casos en un área geográfica durante un periodo de tiempo específico), y es a su vez la principal causa de muerte relacionada con el cáncer a nivel mundial (Figura 1.1) [1].

Hay una variación en las tasas de incidencia y mortalidad (número de muertes en un área geográfica y período de tiempo determinados) muy marcada entre regiones, que refleja los patrones epidemiológicos del tabaco a nivel histórico. En todas las regiones geográficas analizadas el cáncer de pulmón presenta una mayor incidencia en hombres, tal y como se muestra en la Figura 1.2. Sin embargo, en Estados Unidos se está detectando una mayor tasa de incidencia en mujeres jóvenes que en hombres jóvenes, de forma independiente al hábito tabáquico [1].

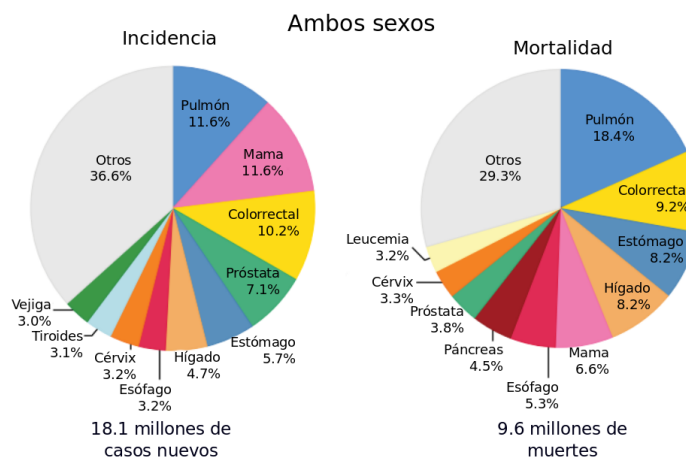


Figura 1.1: Incidencia y mortalidad del cáncer en ambos sexos. Gráficos de sectores con la distribución de casos y muertes para los diez cánceres más frecuentes en 2018 para ambos sexos. Fuente: Bray et al. [1].

La mayor parte de los casos de cáncer de pulmón, aproximadamente el 85%, se clasifican como NSCLC [6, 5], y dentro de este subtipo el adenocarcinoma supone más del 50% de los casos, que además van en aumento [7, 8]. Aunque el adenocarcinoma es el subtipo histológico más común en ambos sexos, hay una mayor predominancia de este subtipo en mujeres [9, 2]. Las mujeres además presentan una mayor tasa de supervivencia, sin importar el estadio, tratamiento, hábito tabáquico e incluso ajustando la esperanza de vida [2].

Factores de riesgo

El hábito tabáquico es el principal factor de riesgo para el cáncer de pulmón [9]. En el tabaco se han identificado hasta 73 compuestos considerados como carcinógenos en animales de laboratorio o humanos por el Centro Internacional de Investigaciones sobre el Cáncer [10].

El riesgo relativo de padecer cáncer de pulmón entre los fumadores puede variar entre 10 y 30 veces dependiendo del nivel de consumo diario de cigarrillos y los años de exposición al tabaco [9]. Además, se han encontrado firmas mutacionales relacionadas con el tabaco que persisten años después de dejar de fumar y se relacionan con el desarrollo de cáncer de pulmón y su supervivencia [11].

A pesar de que en los tumores de fumadores se observa una mayor acumulación de alteraciones genómicas en comparación con los de no fumadores, se estima que a nivel global un 15% de los cánceres de pulmón en hombres y un 53% en mujeres no están relacionados con el tabaco. Estas cifras suponen un 25% de los casos totales de cáncer de pulmón de forma que, si el cáncer de pulmón en no fumadores se tomase como una enfermedad aparte, sería la séptima causa de muerte por cáncer a nivel mundial [7]. Una serie de estudios de asociación han relacionado además el cáncer de pulmón en no fumadores al sexo femenino, a los pacientes procedentes de Asia Oriental, y al subtipo adenocarcinoma [12, 13, 2].

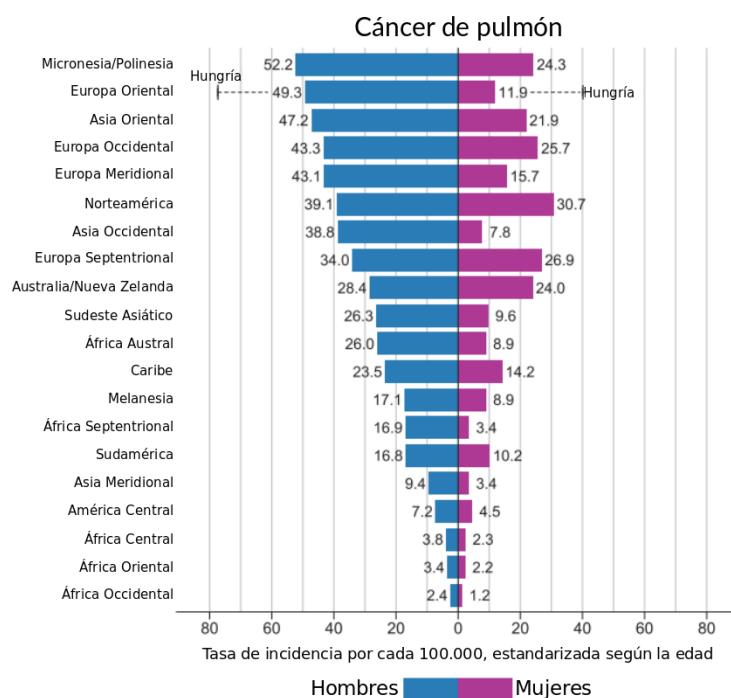


Figura 1.2: Incidencia del cáncer de pulmón por sexo y área geográfica. Gráfico de barras mostrando la tasa de incidencia por sexo y región en cáncer de pulmón en 2018. La tasa de incidencia ha sido estandarizada según la edad poblacional. Fuente: Bray et al. [1].

Las diferencias entre hombres y mujeres en las proporciones de pacientes de cáncer de pulmón no fumadores no reflejan las tendencias de consumo de tabaco, sugiriendo la existencia

de factores biológicos que estén influyendo. Así, en Europa aproximadamente el 20% de las mujeres con cáncer de pulmón y el 2-6% de los hombres con cáncer de pulmón nunca han fumado [2]. Sin embargo, en esta región geográfica el 80.6% de mujeres y el 75.4% de hombres no fuma [14].

Otros factores de riesgo son la exposición a carcinógenos de origen ocupacional (como el asbestos en el trabajo en minas), exposición a combustibles de biomasa domésticos, altos niveles de contaminación ambiental, presencia de otras enfermedades pulmonares (como la enfermedad pulmonar obstructiva crónica), la dieta y factores genéticos [9].

1.1.2. Bases moleculares del adenocarcinoma de pulmón

A nivel celular, el cáncer es un desorden genético en el que se suceden una serie de mutaciones y cambios moleculares, de forma que su combinación lleva al estado de malignidad celular [15]. Estas mutaciones causan una activación inapropiada de las rutas de transducción de señal, promoviendo el crecimiento celular, la angiogénesis y la adopción de un fenotipo metastásico, entre otros procesos [15]. En las células cancerosas las mutaciones se van acumulando de forma progresiva, de forma que el perfil genético de los pacientes en estadio temprano es diferente de aquellos en los que la enfermedad está más avanzada. El cáncer de pulmón es una enfermedad heterogénea a nivel molecular [16], por lo que el conocimiento de su biología es crucial para el desarrollo de nuevas terapias dirigidas contra las alteraciones que presenta cada tumor.

Los genes que participan en el desarrollo de cualquier tipo de cáncer pueden clasificarse en oncogenes o supresores de tumores. Los oncogenes son aquellos genes que pueden transformar células e inducir cáncer, por lo que sus mutaciones suelen ser activadoras. Por el contrario, los genes supresores de tumores son aquellos cuya desactivación parcial o total puede favorecer el desarrollo de cáncer, generalmente porque son limitadores de la proliferación, por lo que suelen sufrir mutaciones inactivadoras [4].

En el adenocarcinoma de pulmón, los oncogenes que se encuentran mutados con más frecuencia son *KRAS* y *EGFR*, además de los supresores de tumores *TP53*, *KEAP1*, *STK11*, *RIT1* y *NFI* [8, 16]. La alteración de estos genes afecta a rutas de señalización celular claves para el desarrollo tumoral, implicadas en la respuesta a estrés oxidativo, la proliferación, y la progresión del ciclo celular entre otros procesos (Figura 1.3) [8].

En no fumadores hay una mayor prevalencia de mutaciones accionables (potencialmente respondedoras a una terapia dirigida) como las que se producen en *EGFR*, *ROS1* o *ALK*, y

además presentan predominantemente transiciones de citosina a timina (C > T). Al contrario, los fumadores presentan más mutaciones no accionables, como las de *KRAS* o *TP53*, y una proporción de 5 a 6 veces mayor de mutaciones, predominantemente de citosina a adenina (C > A) [5, 12, 16].

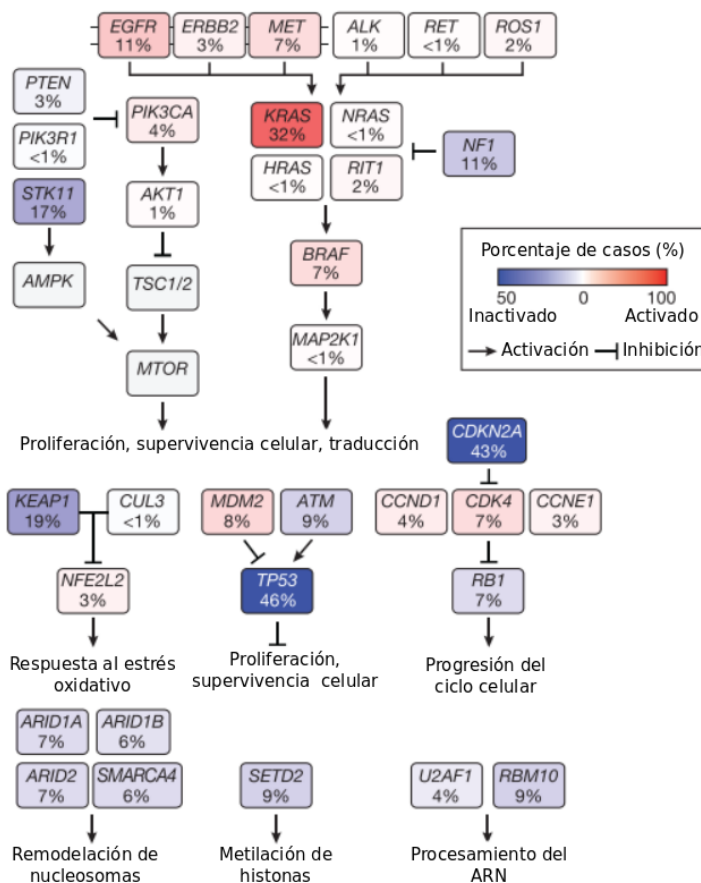


Figura 1.3: Alteraciones en las rutas de señalización en adenocarcinoma de pulmón. Las alteraciones afectan a componentes clave de las rutas de respuesta a estrés oxidativo, proliferación y progresión del ciclo celular, remodelación de nucleosomas, metilación de histonas y procesamiento del RNA. Fuente: The Cancer Genome Atlas Research Network [8].

También parecen jugar un papel en la tumorigénesis las hormonas, ya que se ha descrito la expresión de receptores de estrógenos ($ER\alpha$ y $ER\beta$, mayormente $ER\beta$) en tumores de pulmón tanto en hombres como en mujeres. Aunque el estrógeno promueve el crecimiento de tejido pulmonar, sano y maligno, no se han encontrado asociaciones entre la expresión de *ESR2* (gen de $ER\beta$) y el riesgo o mortalidad del cáncer de pulmón [2].

La caracterización genómica del NSCLC ha permitido identificar los oncogenes que

promueven la transformación tumoral en los diferentes subtipos moleculares de este cáncer, lo cual nos permite realizar terapias dirigidas específicas de subtipo. Sin embargo, esta aproximación actualmente sólo puede aplicarse al 15-20% de los pacientes, y en la mayoría suelen surgir mecanismos de resistencia a los fármacos [10].

1.1.3. Microambiente tumoral y respuesta inmunitaria

El microambiente tumoral (TME) comprende la vasculatura y moléculas que rodean al cáncer, la matriz extracelular y células sanas como fibroblastos o células inmunitarias infiltrantes (que pasan de la sangre al tumor), y juega un papel central en la iniciación y progresión del cáncer de pulmón [17, 18, 19]. No es algo ajeno al tumor, sino que los eventos genéticos que ocurren en las células cancerosas influyen en la composición del TME [16]. La composición de células del sistema inmunitario en el tumor depende del estadio de la enfermedad, y puede contribuir tanto al desarrollo del tumor [18] como a su supresión [19].

El sistema inmunitario tiene dos facetas: una adaptativa y otra innata. En su vertiente adaptativa el sistema “aprende” de un primer contacto con el agente extraño (reconocido como no propio), mientras que en la innata las células tienen desde su formación la habilidad de reconocer agentes infecciosos o células anormales [4]. Las células de la respuesta adaptativa son los linfocitos T y los linfocitos B. La respuesta innata consiste en células dendríticas, linfocitos *natural killer* (NK), granulocitos y mastocitos, entre otras células [20].

Las células T se dividen en linfocitos T CD4+ (o células T cooperadoras) y T CD8+ (o células T citotóxicas). Son capaces de secretar interferón gamma (IFN- γ) y factor de necrosis tumoral alfa (TNF- α), entre otras citocinas, que ejercen efectos antitumorales por medio de la citotoxicidad y la estimulación de la respuesta inmune [4, 17, 19, 20]. Las células T CD8+ son capaces de reconocer los neoantígenos que crean algunas mutaciones tumorales, de forma que se activan y atacan a las células malignas [16]. Los linfocitos B infiltrantes en el tumor también juegan un papel clave antitumoral en el TME, activando las células T CD4+ y evitando la inmunosupresión [17]. Estos y otros componentes del sistema inmunitario, como los linfocitos T reguladores, interaccionan en una compleja red regulando la respuesta inmunitaria final (Figura 1.4), con efectos tanto antitumorales como protumorales [17, 19].

Uno de los principales mecanismos de inmunosupresión en el TME es la inactivación de las células T CD8+. Las células dendríticas producen una serie de factores capaces de ampliar la población de linfocitos T reguladores, que inhiben a los T CD8+, además de inhibirlos directamente por medio del ligando 1 de la proteína de muerte celular programada (PDL1). Aparte

de los T CD8+, otros efectores antitumorales son los linfocitos NK, pero pueden desarrollar fenotipos disfuncionales con pérdida de su potencial de desgranulación y una reducción de la expresión de IFN- γ . Lo mismo ocurre con las células B, que pueden generar anticuerpos específicos del tumor, capaces de activar la respuesta inmunitaria, o adoptar un estado disfuncional capaz de inducir a las células T reguladoras [17].

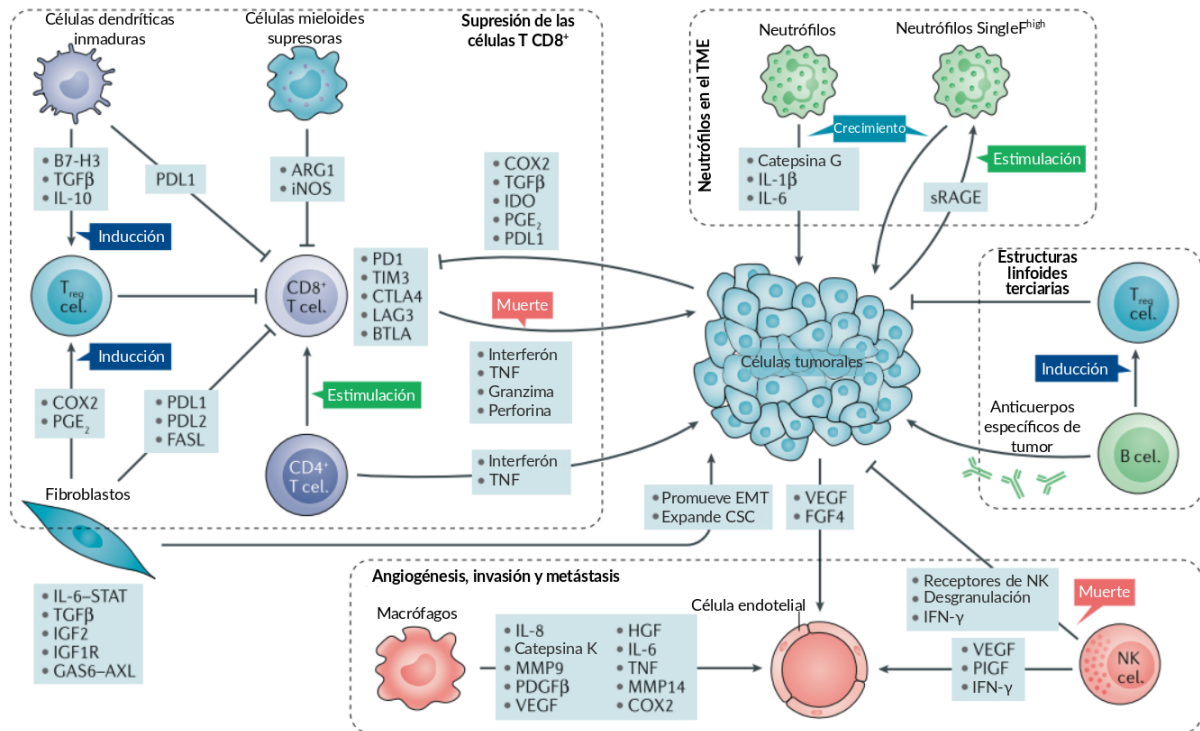


Figura 1.4: El microambiente en cáncer de pulmón. Los principales mecanismos antitumorales del sistema inmunitario es la citotoxicidad mediada por linfocitos T CD8+ y linfocitos NK. Sin embargo, el sistema inmunitario actúa como un arma de doble filo, capaz de promover el crecimiento tumoral y la inmunosupresión en colaboración con otras células del TME y con las propias células cancerosas. Fuente: Altorki et al. [17].

1.2. Datos ómicos y tecnologías de alto rendimiento

El ácido desoxirribonucleico (ADN) es la molécula que contiene la información genética de las células, y está compuesto por dos cadenas antiparalelas polinucleotídicas. Los nucleótidos que conforman los ácidos nucleicos son la adenina (A), citosina (C), guanina (G), timina (T) y uracilo (U). El ADN, que se compone exclusivamente de A, C, G y T, puede transcribirse

a ácido ribonucleico (ARN), en el que la timina se sustituye por el uracilo [21].

El ARN, además de realizar toda una gama de funciones reguladoras, sirve como molde en el proceso de traducción para la síntesis proteica. La molécula que actúa como puente entre la información genética codificada en el ADN y las proteínas recibe el nombre de ARN mensajero o ARNm [21]. Este proceso desde el gen contenido en la secuencia del ADN hasta la proteína efectora final conforma el dogma central de la biología molecular. Desde su formulación en 1970 (y su posterior revisión y ampliación [22], Figura 1.5), el interés por estudiar los procesos descritos en él no ha dejado de aumentar.

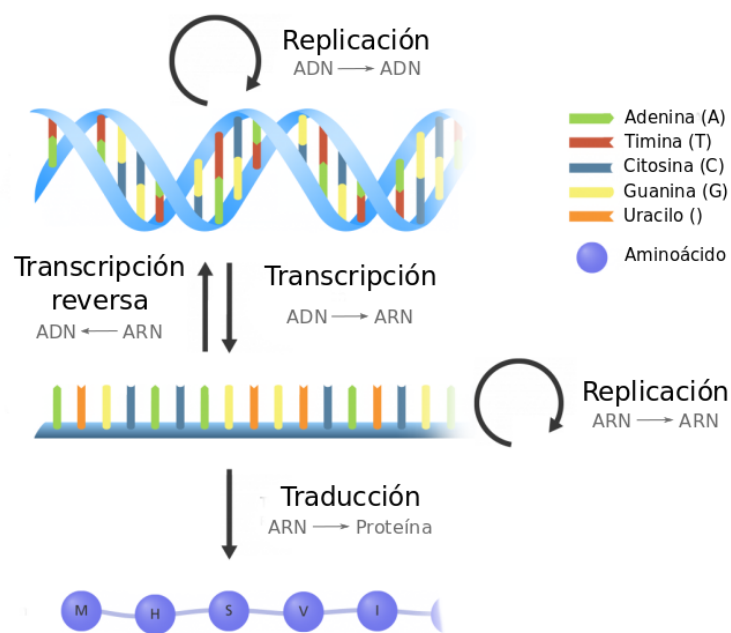


Figura 1.5: Dogma central de la biología molecular ampliado.

La palabra genómica fue acuñada en 1986 por Thomas H. Roderick [23] para referirse al estudio del genoma o conjunto completo de genes del ser humano. Más adelante Marc Wilkins propuso en 1995 [23] la palabra proteómica para el estudio del conjunto de proteínas de un organismo. Desde entonces las ómicas son las disciplinas que se encargan del estudio de grandes cantidades de información tanto dentro como fuera del dogma central de la biología molecular, como la transcriptómica o la metabolómica [23, 24].

La gran cantidad de datos generados por las técnicas de alto rendimiento, en parte debido a su continuo abaratamiento, ha generado la necesidad de disponer de toda una batería de bases de datos y repositorios en los que almacenarlos, clasificarlos y ordenarlos. Estas bases de datos están en continuo cambio, renovación y actualización, de forma que sus cifras varían de

año en año. A fecha de enero de 2019 hay disponibles 1613 bases de datos biológicas [25], cifra que refleja la magnitud de los datos continuamente generados.

1.2.1. Transcriptómica

El transcriptoma está conformado por el total de transcritos de ARN en la célula, tanto codificantes de proteínas como no codificantes, y la transcriptómica es la ciencia que se encarga de estudiarlos [26]. El estudio de los transcritos permite conocer las características de la expresión génica específica de la región estudiada, aportando información crucial para entender la dinámica del metabolismo celular y tisular, y la influencia del transcriptoma en las enfermedades [24]. Esta disciplina ha podido desarrollarse gracias al avance técnico que ha llevado hasta las tecnologías de alto rendimiento, primero los microarrays y más adelante la secuenciación de ARN (RNA-seq).

Los microarrays están formados por un soporte sólido al que se unen una serie de sondas de ADN en una disposición concreta. Para su funcionamiento, requieren el marcaje del ADN o ARN a estudiar con una molécula fluorescente o radiactiva, y su posterior hibridación con las sondas del microarray. Fue la primera técnica que permitía detectar la actividad de miles de transcritos en un único procesamiento [27, 28].

La secuenciación de ARN es una técnica más compleja, en la que el ARN se convierte en ADN por transcripción reversa y, mediante diferentes tecnologías según la plataforma que se use, se obtiene su secuencia de nucleótidos. Esta técnica nos permite identificar los transcritos de la célula y hacer una cuantificación directa de los mismos en un único ensayo. Esto implica que no partimos de un conjunto preestablecido de sondas, como en el caso de los microarrays, si no que podemos identificar nuevos transcritos, detectar isoformas de los mismos, y analizar otros tipos de ARN, como los ARN pequeños de interferencia o los micro ARN [26, 29].

1.2.2. Análisis transcriptómicos

La estrategia de análisis de los datos ómicos se debe adaptar al tipo de datos a analizar y al objetivo del estudio. En el caso de los datos transcriptómicos se requiere un preprocesamiento de los datos, que dependerá de la tecnología utilizada. El resultado del preprocesamiento es una matriz de expresión, que contiene la cuantificación de cada unidad biológica de interés para cada una de las muestras analizadas [30]. Esta matriz nos servirá como punto de partida para las diferentes estrategias de análisis de los datos. Algunas de estas estrategias son [31]:

- **Análisis de expresión diferencial:** El objetivo de esta aproximación es identificar qué genes (u otro tipo de transcritos no codificantes) varían su nivel de expresión entre varias condiciones experimentales.
- **Predicción de clases:** a partir de un set de entrenamiento, compuesto por perfiles transcriptómicos pertenecientes a las clases de interés, se realiza una clasificación de nuevos perfiles de forma supervisada.
- **Análisis cluster, o descubrimiento de clases:** Al contrario que en la predicción de clases, se trata de una clasificación no supervisada. Las posibles clases no se conocen de antemano, si no que surgen de la clasificación de las muestras en función de su similitud a nivel transcriptómico.

1.2.3. Caracterización funcional

A partir del análisis de expresión diferencial se puede obtener una lista de elementos biológicos de interés: elementos que varían entre las condiciones que se quieren comparar. Este listado de elementos puede combinarse con la información biológica contenida en las bases de datos disponibles, facilitando la interpretación de los resultados. En concreto, son de especial interés las anotaciones funcionales, que informan de las funciones asociadas a las unidades biológicas y presentan un vocabulario estructurado [30]. Algunas de las bases de datos que incluyen esta información son:

- *Gene Ontology* (GO), que contiene un modelo de los sistemas biológicos desde el nivel molecular hasta el celular [32]. Incorpora la información clasificada en tres grandes ontologías: procesos biológicos, funciones moleculares y componentes celulares.
- *Kyoto Encyclopedia of Genes and Genomes* (KEGG), que contiene información sistemática sobre la función de los genes y su organización en rutas [33].

Los resultados obtenidos a partir del análisis de los datos transcriptómicos y la información procedente de las bases de datos se combinan mediante metodologías de caracterización funcional. Una estrategia para ello es el análisis de enriquecimiento o análisis de sobrerrepresentación [30], que permite caracterizar las funciones de un grupo de genes (o unidad biológica) de interés. A partir de las funciones anotadas de los genes de interés se realiza una comparación con las funciones de otro grupo de genes de referencia, de forma que se evalúa estadísticamente la proporción de genes con cada función, pudiendo detectar una sobrerrepresentación o enriquecimiento de dicha función en uno de los grupos de genes.

Otra aproximación son los análisis sobre grupos de genes [30], como el *Gene Set Enrichment Analysis* (GSEA) [34]. Esta metodología emplea la lista completa de genes y los clasifica según su nivel de expresión diferencial, de forma que se analiza si los genes pertenecientes a un mismo grupo funcional tienden a estar sobreexpresados o infraexpresados.

A partir de las estrategias anteriores, surgen también las estrategias de análisis de actividad de rutas, como el *Canonical Circuit Activity Analysis* (CCAA) [35]. Este método tiene en cuenta la circuitería de las rutas y las interacciones proteicas que se dan en ellas para computar la señal final de salida sobre cada función considerando la información disponible sobre la expresión génica.

1.2.4. Metaanálisis funcional

Las revisiones sistemáticas y el metaanálisis son herramientas esenciales para la síntesis de la evidencia científica. El metaanálisis puede definirse como una revisión y síntesis cuantitativa de los resultados de estudios independientes sobre una misma temática [36, 37]. Combinar la información de varios estudios mejora el poder estadístico de los resultados, obteniendo una medida combinada del efecto de interés. Este incremento del poder estadístico permite al metaanálisis enfrentarse a estudios con conclusiones conflictivas, estimando un efecto medio ponderado en todo el grupo de estudios analizados [30, 36].

Aunque el metaanálisis surgió inicialmente en el campo de las ciencias sociales [38], se ha aplicado con éxito a los análisis de expresión génica, permitiendo descubrir nuevas dianas terapéuticas [39], subtipos de enfermedad [40] o predecir la supervivencia de los pacientes [41]. La mayoría de métodos de metaanálisis de estudios transcriptómicos se centran a nivel génico, pero aplicarlo a nivel de función puede mejorar la detección de las funciones que están realmente enriquecidas en los grupos de interés, ya que los resultados individuales suelen ser inconsistentes debido a problemas como el ruido de la tecnología empleada, el diseño experimental o el reducido tamaño muestral [30].

2. Hipótesis

En la población general, los hombres presentan una mayor incidencia de cáncer de pulmón, mientras que en no fumadores la incidencia es mayor en mujeres [2, 12, 13]. Además, si nos fijamos en el subtipo más común (adenocarcinoma), éste predomina en mujeres, que también tiene una mayor tasa de supervivencia [1, 9, 2]. Todo ello sugiere que las diferencias observadas puedan ser debidas a las características genéticas y hormonales de ambos sexos.

Teniendo en cuenta las diferencias epidemiológicas entre sexos, es probable que existan diferencias entre hombres y mujeres a nivel de activación y represión de las funciones moleculares que subyacen a la enfermedad. Estudiar el contexto molecular del NSCLC, subtipo adenocarcinoma, puede llevarnos a la identificación de nuevos biomarcadores capaces de ampliar el repertorio de terapias actual.

3. Objetivos

El objetivo de este trabajo es identificar mediante un abordaje in silico las bases moleculares que afectan de forma diferencial a hombres y mujeres en cáncer de pulmón de subtipo adenocarcinoma. Para lograrlo, se han propuesto tres objetivos específicos a cumplir:

- Revisión sistemática y selección de estudios de cáncer de pulmón con datos de microarrays o RNA-seq en los repositorios *Gene Expression Omnibus* (GEO), ArrayExpress y *Genomic Data Commons* (GDC).
- Análisis individual de cada estudio seleccionado, incluyendo un análisis de expresión diferencial, un análisis de enriquecimiento funcional y un análisis de rutas.
- Metaanálisis de los resultados funcionales obtenidos a nivel de enriquecimiento funcional y rutas.

4. Material y métodos

Todo el análisis bioinformático descrito a continuación ha sido realizado en el lenguaje de programación R [42]. El código empleado y las funciones creadas están disponibles en gitlab (<https://gitlab.com/metafundev/metafunr>). Todas las librerías o paquetes indicados en cada uno de los pasos se encuentran detalladas en el Anexo B.

4.1. Revisión sistemática y selección de estudios

La revisión sistemática es una metodología que permite cotejar toda la evidencia empírica que cumple ciertos criterios de elegibilidad para responder a preguntas concretas. Este proceso incluye el establecimiento de objetivos claros, la búsqueda sistemática de estudios que cumplan con los criterios de elegibilidad establecidos, una evaluación de la validez de los estudios y una síntesis de los resultados obtenidos a partir de dichos estudios [37]. Para el desarrollo de este trabajo se han seguido las directrices y el flujo de trabajo propuestos en la declaración PRISMA [37] (Figura 4.1), fruto de una colaboración internacional y multidisciplinaria para estandarizar los procesos de metaanálisis y revisión sistemática.

En primer lugar, se llevó a cabo una búsqueda exhaustiva de los estudios disponibles en bases de datos públicas sobre adenocarcinoma de pulmón, que contuviesen información sobre la expresión génica. En concreto, se escanearon las bases de datos GEO [43] y ArrayExpress [44] con las palabras clave *expression profiling by array*, *expression profiling by high throughput sequencing*, *adenocarcinoma of lung*, *lung adenocarcinoma*, *non-small-cell lung carcinoma* y *homo sapiens*. Se incluyeron como palabras excluyentes *cell line* y *cell lines*. También se utilizó la base de datos GDC [45], en concreto los datos disponibles del proyecto *The Cancer Genome Atlas* (TCGA) sobre adenocarcinoma de pulmón (TCGA-LUAD).

De los estudios obtenidos a partir de la búsqueda, se seleccionaron aquellos que contenían información sobre el sexo, el estadio de la enfermedad y el hábito tabáquico de los pacientes. Se excluyeron aquellos estudios realizados sobre líneas celulares humanas o sin muestras control y aquellos con pacientes tratados o con un tamaño muestral por grupo experimental menor a cuatro sujetos.

Los estudios que finalmente fueron incluidos en el análisis pertenecen a las bases de datos GEO y GDC, por lo que los datos se obtuvieron mediante los paquetes GEOquery [46]

y TCGAbiolinks [47].

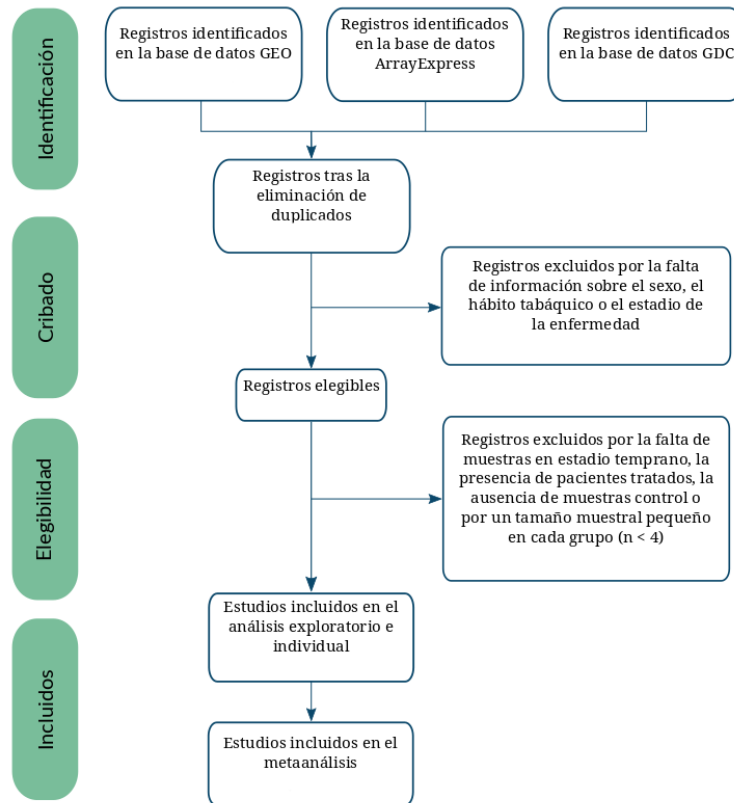


Figura 4.1: Flujo de información a través de las diferentes fases de una revisión sistemática. Esquema adaptado de Liberati, A et al. [37].

4.2. Análisis individual de los estudios

Para poder aplicar la metodología del metaanálisis es necesario en primer lugar analizar cada uno de los resultados por separado, combinando finalmente los resultados.

4.2.1. Procesamiento de los datos

Normalización de los datos clínicos

Debido a la variabilidad en la naturaleza de los estudios y en la terminología empleada

en cada uno de ellos, se llevó a cabo una normalización de la nomenclatura de los datos clínicos disponibles para evitar problemas al integrar los datos en el metaanálisis. Para la normalización se siguieron las siguientes pautas:

- Grupo "Control": Pacientes sanos, divididos en dos grupos según el sexo.
- Grupo "Adenocarcinoma": Pacientes con NSCLC adenocarcinoma, divididos en dos grupos según el sexo.
- Variable factorial "estadio": Se ha dividido en estadio temprano (I-IIb) o tardío (III-IV), y han sido excluidas las muestras de estadio tardío.
- Variable factorial "fumadores": Se ha dividido en fumadores (tanto actuales como ex-fumadores) y no fumadores.

Anotación

La anotación de las sondas es un paso esencial en el procesamiento de microarrays, y varía en función del modelo y la plataforma comercial. Los códigos de las sondas de cada uno de los microarrays se transformaron en un identificador común, el código Entrez, que identifica cada uno de los genes en la base de datos del *National Center for Biotechnology Information* (NCBI, Estados Unidos) [48]. Para facilitar la automatización del flujo de trabajo, los identificadores de gen de los estudios de RNA-seq también se cambiaron por identificadores Entrez.

En aquellos casos en que varias sondas de un chip de ADN hibridan con secuencias distintas de un mismo gen, el hecho de convertir el identificador de la sonda a identificador Entrez provoca que aparezcan genes duplicados en la matriz de expresión. Para el análisis es necesario obtener un valor único de expresión para cada gen por muestra, por lo que se creó la función `medianReps`, disponible en gitlab. Esta función calcula la mediana de los niveles de expresión de todas las sondas que apuntan a un mismo gen, en cada uno de los pacientes.

Análisis exploratorio

Se realizó un análisis exploratorio de cada estudio seleccionado y anotado con el objetivo de evaluar la distribución de las muestras en grupos de interés y detectar comportamientos anómalos en las muestras.

Para evitar la introducción de nuevos sesgos y más variabilidad, se partió de los datos normalizados de los estudios de microarrays seleccionados. En primer lugar se representaron los niveles de expresión de los genes de cada una de las muestras en un diagrama de cajas para

comprobar que no hubiese anomalías. En el caso de los estudios de RNA-seq, se descargaron las matrices de conteos y se normalizaron utilizando el paquete `edgeR` [49] por el método *Trimmed Mean of M-values* (TMM) que, junto con el método DESeq, es el que mejor rendimiento ofrece en la normalización [29].

A partir de los datos normalizados se procedió al análisis de agrupamiento o clustering, tanto con distancia euclídea como de correlación, con el objetivo de observar posibles agrupaciones de los niveles de expresión de las muestras según el sexo o la enfermedad de los pacientes. Además, se realizó un análisis de componentes principales (PCA) con el mismo objetivo.

4.2.2. Análisis de expresión diferencial

Para el análisis de expresión diferencial se compararon las diferencias en mujeres debidas a la enfermedad con las diferencias en hombres debidas a la enfermedad, utilizando el siguiente contraste de interés:

$$\begin{aligned} & (\textit{Adenocarcinoma Mujer} - \textit{Control Mujer}) - \\ & (\textit{Adenocarcinoma Hombre} - \textit{Control Hombre}) \quad (4.1) \end{aligned}$$

Para ello, se combinaron las variables factoriales “sexo” (Hombre o Mujer) y “adenocarcinoma” (Control o Adenocarcinoma), generando los cuatro posibles grupos experimentales. La información sobre el hábito tabáquico (variable “fumador”: Fumador o No fumador) y el emparejamiento de muestras en algunos de los estudios evaluados, se incorporó en un modelo lineal mediante variables control, utilizando las funciones del paquete `limma` [50], con el objetivo de detectar las diferencias de expresión entre sexos. Se calcularon los coeficientes estimados y los errores estándares a partir de la matriz de contrastes y el modelo lineal, y los estadísticos de la expresión diferencial. Finalmente se ajustó el p-valor mediante el método de Benjamini y Hochberg (BH) [51].

Las funciones utilizadas y el flujo de trabajo empleado, tanto para los datos provenientes de microarrays como de RNA-seq, pueden encontrarse en gitlab en la función `diffExp`.

4.2.3. Análisis de enriquecimiento funcional

Una vez obtenidos los datos de expresión diferencial de los genes, se realizó el análisis de enriquecimiento funcional mediante GSEA [52] con la librería `mdgsa` [53]. Las funciones y el flujo de trabajo pueden encontrarse en la función `doGsea` en `gitlab`.

Para el enriquecimiento funcional con los términos GO, se realizó previamente una correspondencia entre los identificadores Entrez de los genes y sus términos GO mediante con la librería `biomaRt` [54]. Todo el proceso se realizó para cada una de las categorías GO de interés (procesos biológicos, funciones moleculares o componentes celulares) propagando la anotación, de forma que los genes asociados a un término GO heredan la anotación de los términos ancestros de su término GO asociado. Finalmente, se filtraron las anotaciones muy específicas o demasiado genéricas (anotaciones con un número de genes asociados menor a 10 o superior a 500).

Utilizando los estadísticos obtenidos a partir de la expresión diferencial, los genes se ordenaron en función de sus p-valores y su estadístico de contraste. Una vez ordenados, la clasificación se transformó para adecuar su distribución a una variable independiente de un modelo de regresión logística univariante y se realizó un GSEA. Se consideraron significativos los términos con un p-valor ajustado inferior a 0.05.

En el caso de las rutas KEGG, se realizó la anotación con el paquete `org.Hs.eg.db` [55]. Todo el proceso se llevó a cabo utilizando `doGsea`, como en caso de los GO, excepto la propagación de la anotación, que no puede aplicarse a estructuras no jerárquicas como es el caso de las rutas KEGG.

4.2.4. Análisis de rutas de señalización

Hipathia [35] integra el método CCAA de análisis de actividad de rutas, permitiendo la modelización de la señalización celular a partir de los conocimientos disponibles sobre transducción de señal y datos transcriptómicos. Este método se basa en un algoritmo iterativo capaz de calcular la intensidad de la señal que pasa a través de los nodos de una red teniendo en cuenta el nivel de expresión de cada gen (o nodo) y la intensidad de la señal que llega a él. Esta aproximación nos permite estimar la señal final que llega a cada uno de los nodos efectores de la ruta, con las funciones anotadas que correspondan.

Las rutas incluidas en *Hipathia* provienen de la base de datos KEGG. Dado que una

única ruta puede estar asociada a varios nodos efectores con funciones diferentes, *Hipathia* divide cada ruta en subrutas que son analizadas de forma independiente. Cada subruta está compuesta por un nodo efector final y todos los nodos que llevan hasta él.

El flujo de trabajo de *Hipathia* comienza con la traducción de los genes a identificador Entrez. Dado que el identificador empleado en este trabajo es Entrez, el primer paso realizado fue la normalización de los valores de expresión a valores comprendidos entre 0 y 1. A continuación se calculó el nivel de activación de cada subruta para cada una de las muestras del estudio. Una vez obtenidos los valores de activación, se realizó la comparación de interés (Ecuación 4.1) a través del `limma` con la función `diffExp`.

4.3. Metaanálisis

A partir de los resultados del enriquecimiento funcional, tanto en términos GO como en rutas KEGG, disponemos de las medidas de los logaritmos de los *odds ratios* (o razón de ventajas) y su varianza para cada una de las funciones evaluadas. Estas dos medidas para cada función en cada estudio se condensaron en dos matrices: la de logaritmos de los *odds ratio* (LOR) y la de las varianzas de los LOR.

Para mejorar el conocimiento de los datos se realizó una exploración de las matrices, incluyendo la descripción de la magnitud de efecto y su variabilidad, la detección de valores atípicos, y la cuantificación de valores perdidos [30].

4.3.1. Determinación de la medida combinada del efecto

El metaanálisis se realizó para cada una de las funciones detectadas en el análisis de enriquecimiento funcional. Generalmente, en el metaanálisis se evalúa primero la heterogeneidad de los estudios y, en función de sus resultados, se decide el modelo de estimación de la variabilidad del efecto. Sin embargo, trabajar a nivel de función supuso realizar miles de metaanálisis simultáneos (uno por función), por lo que la valoración de la heterogeneidad de los estudios para confirmar la idoneidad del método seleccionado se realizó a posteriori [30]. La heterogeneidad hace referencia a la variabilidad en los efectos evaluados, debida a la diversidad propia de los datos y a la de las metodologías empleadas en los estudios [56].

Todos los metaanálisis se realizaron con la función `metaGO`, disponible en gitlab, que

incorpora las funciones necesarias de las librerías *metafor* [57] y *mdgsa*. Se utilizó el modelo de efectos fijos (FE), y los modelos de efectos aleatorios DerSimonian y Laird (DL) [58], Hedges (HE) [59] y Hunter y Schmidt (HS) [60], para su evaluación. Finalmente, el método escogido para el metaanálisis fue DL, por ajustarse mejor a las características de los estudios seleccionados (diferentes tecnologías y datos ómicos).

El resultado del metaanálisis es un conjunto de estimadores de la medida del efecto combinado para cada una de las funciones o rutas estudiadas, además de los indicadores de heterogeneidad del metaanálisis [30, 61]:

- QE y QEp son el estadístico de contraste y p-valor, respectivamente, del método DL, utilizados para detectar la heterogeneidad entre estudios. La hipótesis nula hace referencia a la no heterogeneidad entre los estudios.
 - LOR es la estimación del efecto combinado de todos los estudios. El signo indica la sobrerrepresentación de una determinada función en uno de los dos grupos de la comparación realizada (Ecuación 4.1), siendo mayor en la segunda clase si el signo es negativo. La magnitud del valor es un indicador cuantitativo de dicha sobrerrepresentación.
 - Intervalo de confianza del LOR (al 95%), construido a partir de la variabilidad estimada en el modelo seleccionado. Si el intervalo de confianza no incluye el 0, el LOR se considera significativo.
 - El p-valor informa sobre el nivel de significación del efecto combinado. Dado que el p-valor no contempla el escenario de multiplicidad del análisis (múltiples funciones) es necesario realizar una corrección del mismo. Esta corrección fue realizada mediante el método BH.
 - τ^2 describe la heterogeneidad entre los estudios.
 - El estadístico I^2 indica el porcentaje de variabilidad debida a la heterogeneidad en los estudios. Cuanto menor sea este valor, menor relevancia tiene la heterogeneidad detectada.
 - H^2 es el cociente entre la variabilidad total y la variabilidad en el muestreo.
-

4.3.2. Evaluación de la heterogeneidad

El test Q propuesto por DL presenta una baja potencia a la hora de trabajar con un número pequeño de estudios primarios, por lo que el análisis de heterogeneidad se complementa con varios métodos que se resumen gráficamente en gráficos de embudo [30], además de las representaciones gráficas de los estimadores previamente descritos.

- Gráficos de embudo: evalúan la variabilidad de los distintos estudios y la presencia de sesgos. Se muestra la magnitud del efecto medido (eje X) frente a una medida de precisión (eje Y), como la desviación estándar o el inverso de la varianza.
- Gráficos radiales: permiten evaluar la consistencia de los efectos según su nivel de precisión. Se muestran la inversa de los errores estándares (precisión, eje X) y el tamaño de los efectos observados estandarizados por su error estándar (eje Y).

4.3.3. Representación de resultados

Además de los resultados crudos del metaanálisis, también se generaron gráficos de bosque para cada una de las funciones. Estos gráficos nos permiten evaluar los resultados globales y el peso en el resultado de cada función de cada estudio individual [30].

En este tipo de gráficos, a la izquierda se enumeran los estudios incluidos en el metaanálisis y a la derecha la estimación de la medida resumen individual de cada uno, con su intervalo de confianza al 95%. En el centro del gráfico se visualiza la medida de efecto (cuadrado negro) y su intervalo de confianza, siendo proporcional el tamaño del cuadrado a la precisión de las estimaciones. En la parte inferior central se muestra el resultado del metaanálisis, de forma que su posición respecto a la línea de efecto nulo nos informa sobre la significación estadística del resultado, y su anchura nos informa de su intervalo de confianza. Finalmente, en la parte inferior izquierda encontramos el modelo de análisis empleado (DL) y a la derecha el intervalo de confianza al 95% (Figura 4.2).

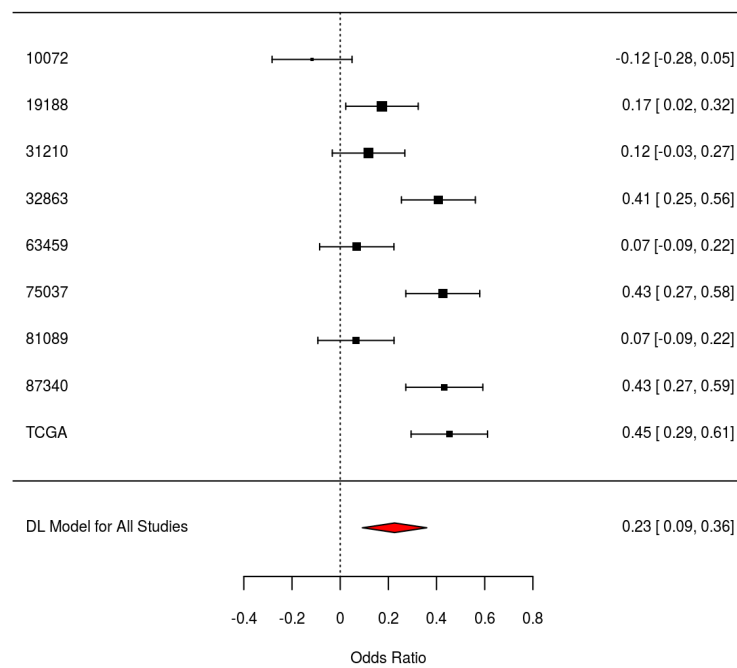
GO:0002285 - Activación linfocitaria implicada en la respuesta inmune

Figura 4.2: Ejemplo de un gráfico de bosque Los odds ratio del eje X se encuentran en escala logarítmica

5. Resultados

En este apartado se muestran los resultados obtenidos, siguiendo el orden establecido en el apartado anterior. Las figuras y tablas complementarias pueden encontrarse en los Anexos A y B, respectivamente.

5.1. Revisión sistemática y selección de estudios

En la revisión sistemática encontramos, tras la eliminación de duplicados, un total de 207 estudios candidatos. Hasta 157 estudios fueron excluidos por carecer de información sobre el sexo de los pacientes, su hábito tabáquico o el estadio de la enfermedad. De los 50 estudios restantes, se descartaron 41 por no disponer de muestras en estadio temprano, no tener muestras control, haber sometido a tratamiento a los pacientes, o presentar un tamaño muestral muy pequeño en cada grupo ($n < 4$). Finalmente, 9 estudios fueron seleccionados para el metaanálisis (Tabla 5.1, Figura 5.1).

Tabla 5.1: Estudios seleccionados tras la revisión sistemática. Se indica la plataforma utilizada en cada estudio y las publicaciones asociadas.

Estudio	Plataforma	Artículo
GSE10072	Affymetrix Human Genome U133A Array	[11]
GSE19188	Affymetrix Human Genome U133 Plus 2.0 Array	[62]
GSE31210	Affymetrix Human Genome U133 Plus 2.0 Array	[63, 64]
GSE32863	Illumina HumanWG-6 v3.0 expression beadchip	[13]
GSE63459	Illumina HumanRef-8 v3.0 expression beadchip	[65]
GSE75037	Illumina HumanWG-6 v3.0 expression beadchip	[66]
GSE81089	Illumina HiSeq 2500	[67]
GSE87340	Illumina HiSeq 2000	[68]
TCGA-LUAD	Illumina HiSeq 2000	[8]

En el estudio GSE19188 no estaba disponible para muchos individuos la información sobre el sexo, el estadio de la enfermedad y el hábito tabáquico, por lo que se contactó con el responsable para que nos proporcionara estos datos e incorporar este estudio en el metaanálisis.

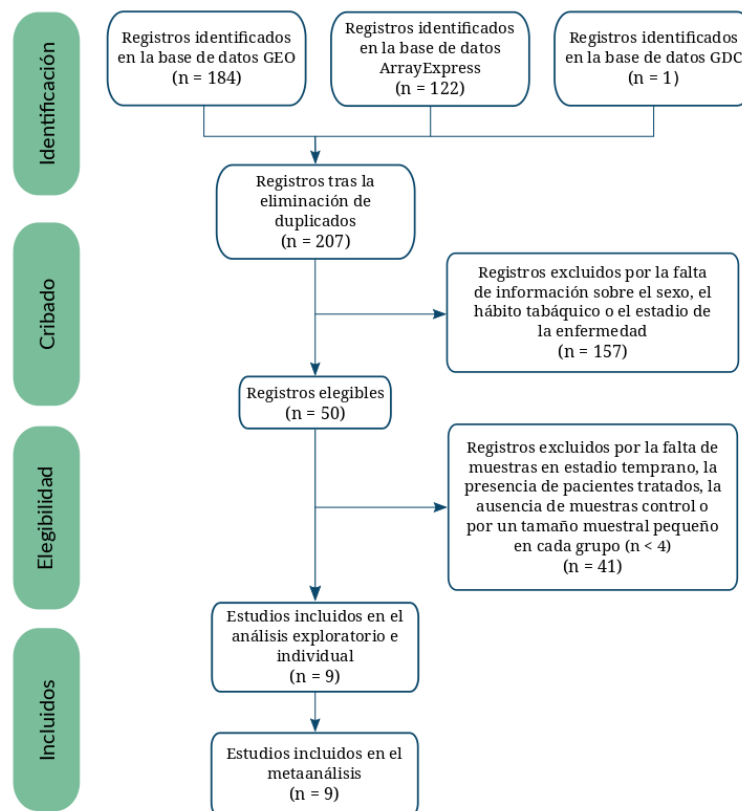


Figura 5.1: Diagrama de flujo de la revisión sistemática, siguiendo la declaración PRISMA.

El mayor factor limitante a la hora de seleccionar estudios fue la inclusión de información sobre el sexo del paciente. Otros factores limitantes durante la selección de estudios fueron el tamaño muestral, la ausencia de controles o la presencia de pacientes tratados.

5.2. Análisis individual de los estudios

5.2.1. Procesamiento de los datos

Tras la obtención de los datos de cada estudio y la estandarización de la nomenclatura de los datos clínicos, se evaluó la normalización de las matrices de expresión de los datos de microarrays mediante diagramas de cajas. De los estudios de RNA-seq se descargaron las matrices de conteos y se normalizaron por el método TMM del paquete edgeR. Se analizó entonces también la normalización realizada mediante diagramas de cajas.

A partir de los datos normalizados, se describió la cantidad de individuos por sexo y por enfermedad (control o adenocarcinoma) en cada estudio (Figura 5.2).

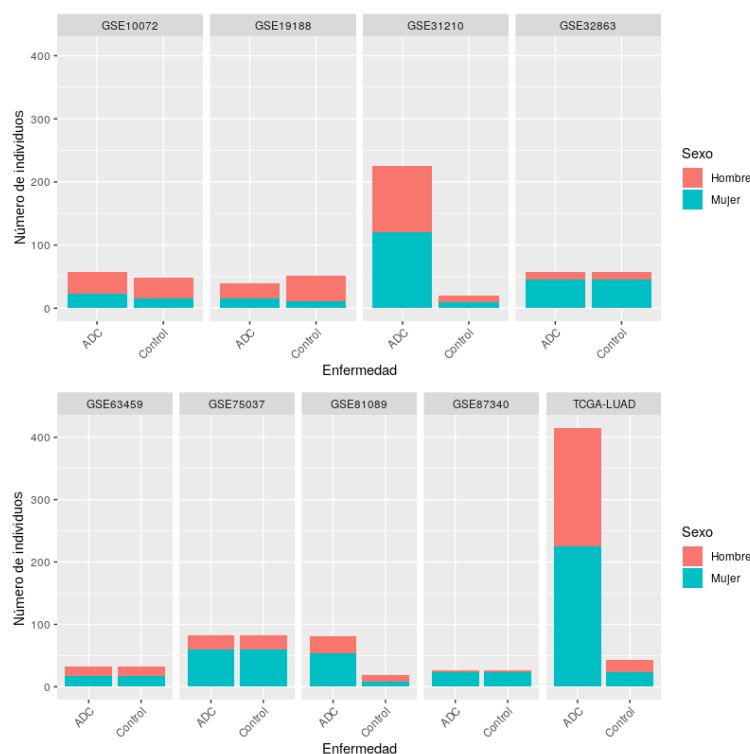


Figura 5.2: Número de pacientes en cada estudio. Se presentan distribuidos por sexo y enfermedad.

GSE10072

En este estudio [11] se analizó la firma de expresión génica que provoca el tabaco y su papel en el desarrollo de adenocarcinoma y la supervivencia de los pacientes, a partir de tejido pulmonar. La Tabla 5.2 muestra la distribución de individuos por sexo y enfermedad en este estudio, tras eliminar las muestras que pertenecen a pacientes que no son de estadio temprano.

Tabla 5.2: Individuos incluidos en el estudio GSE10072. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	8	29	37
Adenocarcinoma	13	30	43
Total	21	59	80

Tanto el análisis de agrupamiento (Anexo I, Figura A.1) como el PCA (Anexo I, Figura A.2) mostraron un comportamiento normal de las muestras, agrupando en función de la presencia o ausencia de enfermedad.

GSE19188

En este estudio [62] se analizó la expresión génica en diferentes tipos de NSCLC para identificar una firma de 75 genes capaz de clasificar las muestras, y también encontraron una firma pronóstica de 17 genes, a partir de tejido pulmonar. La Tabla 5.3 muestra la distribución de individuos por sexo y enfermedad en este estudio, tras eliminar las muestras que pertenecen a pacientes que no son de estadio temprano.

Tabla 5.3: Individuos incluidos en el estudio GSE19188. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	11	41	52
Adenocarcinoma	11	21	32
Total	22	62	84

Tanto el análisis de agrupamiento (Figura 5.3) como el PCA (Figura 5.4) mostraron un agrupamiento de las muestras GSM47570 y GSM47578 con las muestras control, siendo adenocarcinomas. Por ello, se omitieron de los siguientes análisis.

GSE31210

En este estudio [63, 64] se analizó la expresión génica de adenocarcinomas de estadio I o II. La Tabla 5.4 muestra la distribución de individuos por sexo y enfermedad en este estudio. Todos los pacientes de este estudio se encontraban en estadio I o II en el momento de obtención del tejido pulmonar, por lo que no se tuvo que eliminar ninguna muestra para el análisis.

Tabla 5.4: Individuos incluidos en el estudio GSE31210. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	9	11	20
Adenocarcinoma	121	105	226
Total	130	116	246

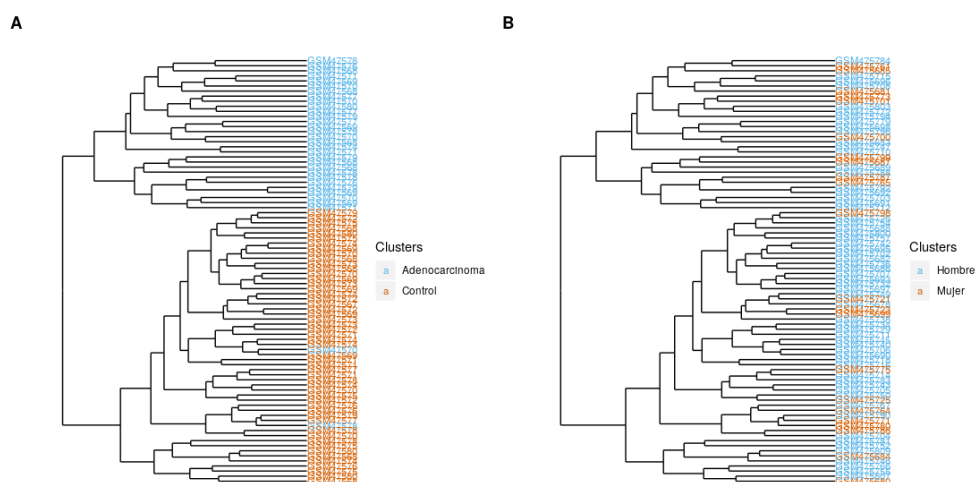


Figura 5.3: Clustering exploratorio de las muestras del estudio GSE19188 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

Tanto el análisis de agrupamiento (Figura 5.5) como el PCA (Figura 5.6) mostraron un comportamiento normal de las muestras, a excepción de la muestra GSM773784, que se agrupaba en ambos casos con los adenocarcinomas siendo control. Por ello, quedó excluida en los siguientes análisis.

GSE32863

En este estudio [13] se analizó la expresión génica de adenocarcinomas y tejido pulmonar sano adyacente, para integrarlo con información sobre metilación. La Tabla 5.5 muestra la distribución de individuos por sexo y enfermedad en este estudio, después de excluir aquellos individuos en estadio tardío.

Tabla 5.5: Individuos incluidos en el estudio GSE32863. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	35	10	45
Adenocarcinoma	35	10	45
Total	70	20	90

Tanto el análisis de agrupamiento (Anexo I, Figura A.3) como el PCA (Anexo I, Figura A.4) mostraron un comportamiento normal de las muestras, agrupando las muestras en función

de la presencia o ausencia de enfermedad.

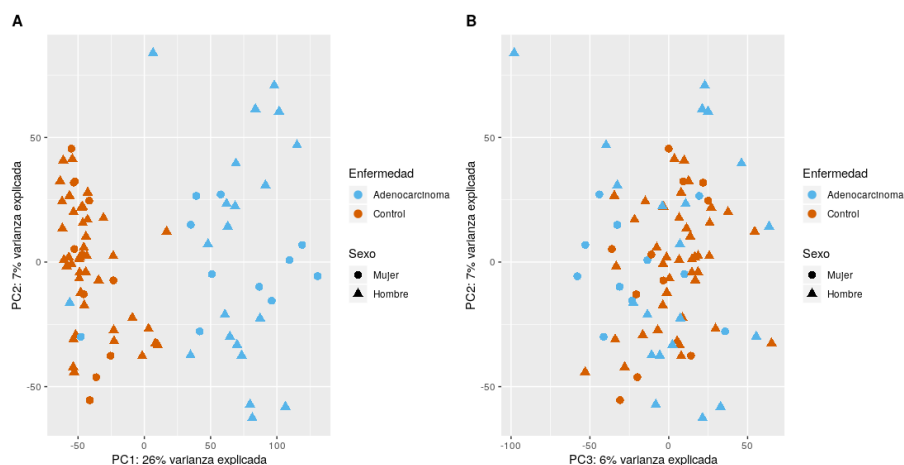


Figura 5.4: Análisis de componentes principales de las muestras del estudio GSE19188. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

GSE63459

En este estudio [65] se analizó la expresión génica con el objetivo de caracterizar la expresión del adenocarcinoma de pulmón en estadio I. La Tabla 5.6 muestra la distribución de individuos por sexo y enfermedad en este estudio, después de excluir dos muestras (GSM1550225 y GSM1550226) de las cuales no se dispone de información sobre el hábito tabáquico.

Tabla 5.6: Individuos incluidos en el estudio GSE63459. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	17	14	31
Adenocarcinoma	17	15	32
Total	34	29	63

Tanto el análisis de agrupamiento (Anexo I, Figura A.5) como el PCA (Anexo I, Figura A.6) mostraron un comportamiento normal de las muestras, agrupando las muestras en función de la presencia o ausencia de enfermedad.

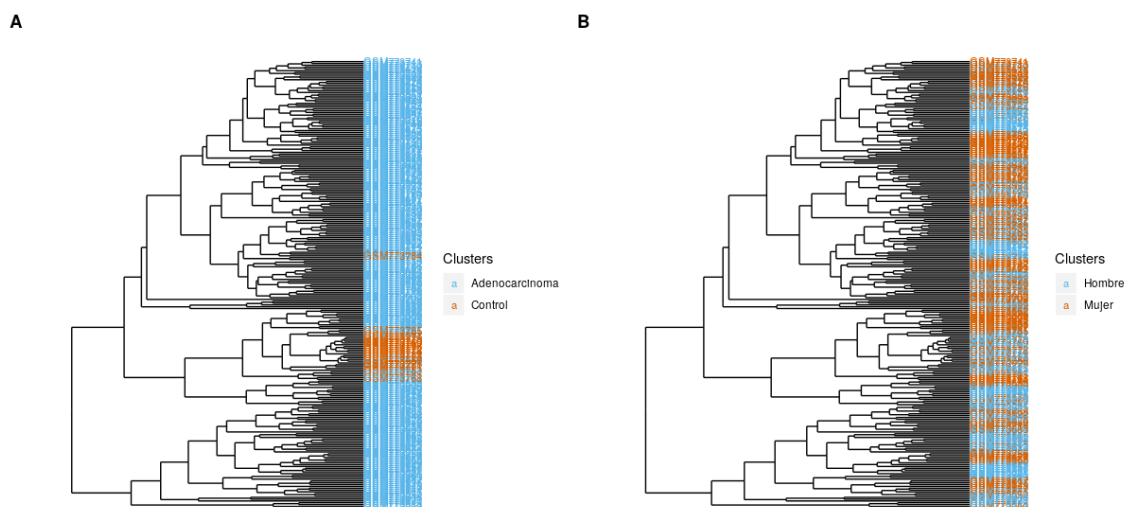


Figura 5.5: Clustering exploratorio de las muestras del estudio GSE31210 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

GSE75037

En este estudio [66] se analizó la expresión génica en muestras emparejadas de adenocarcinoma y tejido sano con el objetivo de desarrollar una firma de expresión como apoyo a la clasificación histológica rutinaria de los tumores. La Tabla 5.7 muestra la distribución de individuos por sexo y enfermedad en este estudio. Ninguna muestra fue excluida, ya que todas pertenecían a pacientes en estadio temprano.

Tabla 5.7: Individuos incluidos en el estudio GSE75037. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	59	24	83
Adenocarcinoma	59	24	86
Total	118	48	166

Tanto el análisis de agrupamiento (Anexo I, Figura A.7) como el PCA (Anexo I, Figura A.8) mostraron un comportamiento normal de las muestras, agrupando las muestras en función de la presencia o ausencia de enfermedad.

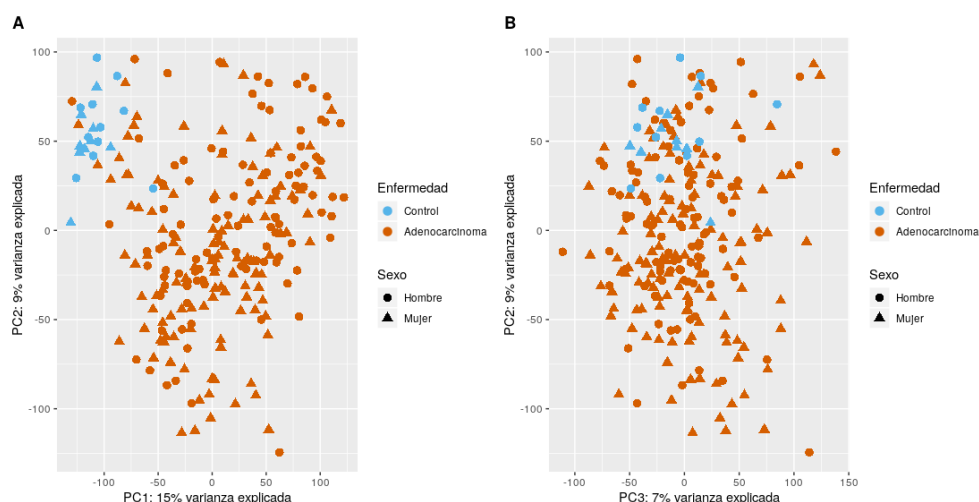


Figura 5.6: Análisis de componentes principales de las muestras del estudio GSE31210. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

GSE81089

En este estudio [67] se analizó la expresión génica en NSCLC con el objetivo de caracterizar antígenos biomarcadores. La Tabla 5.8 muestra la distribución de individuos por sexo y enfermedad en este estudio. Ninguna muestra fue excluida, ya que todas pertenecían a pacientes en estadio temprano.

Tabla 5.8: Individuos incluidos en el estudio GSE81089. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	8	11	19
Adenocarcinoma	53	24	81
Total	61	39	100

Tanto el análisis de agrupamiento (Anexo I, Figura A.9) como el PCA (Anexo I, Figura A.10) mostraron un comportamiento normal de las muestras, agrupando las muestras en función de la presencia o ausencia de enfermedad.

GSE87340

En este estudio [68] se analizó la expresión génica para combinarla con secuenciación

de exomas y encontrar mutaciones que afecten a la desregulación de las rutas de señalización. La Tabla 5.9 muestra la distribución de individuos por sexo y enfermedad en este estudio. Ninguna muestra fue excluida, ya que todas pertenecían a pacientes en estadio temprano.

Tabla 5.9: Individuos incluidos en el estudio GSE87340. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	23	4	27
Adenocarcinoma	23	4	27
Total	46	8	54

Tanto el análisis de agrupamiento (Anexo I, Figura A.11) como el PCA (Anexo I, Figura A.12) mostraron un comportamiento normal de las muestras, agrupando las muestras en función de la presencia o ausencia de enfermedad.

TCGA-LUAD

Este estudio [8] forma parte del proyecto *The Cancer Genome Atlas*, un programa de genómica del cáncer cuyo objetivo era caracterizar molecularmente hasta 33 tipos de cáncer distintos. La Tabla 5.10 muestra la distribución de individuos por sexo y enfermedad en este estudio, tras eliminar las muestras que no pertenecen al grupo de estadio temprano.

Tabla 5.10: Individuos incluidos en TCGA-LUAD. Se muestran distribuidos por sexo y enfermedad.

	Mujeres	Hombres	Total
Control	24	19	43
Adenocarcinoma	225	190	415
Total	249	209	458

Tanto el análisis de agrupamiento (Anexo I, Figura A.13) como el PCA (Anexo I, Figura A.14) mostraron un comportamiento normal de las muestras, agrupando las muestras en función de la presencia o ausencia de enfermedad.

5.2.2. Análisis de expresión diferencial

La tabla 5.11 muestra el número de genes expresados de forma diferencial en cada estudio (p-valor ajustado por BH < 0.05) para la comparación de interés (Ecuación 4.1). También se muestran, con fin exploratorio, los resultados de las comparaciones entre mujeres control y mujeres con adenocarcinoma, y entre hombres control y hombres con adenocarcinoma. Los resultados se muestran agrupados en *Up* si el logaritmo de la magnitud de cambio (logFC) es mayor a 0, o en *Down* si el logFC es menor a 0.

Tabla 5.11: Resumen del resultado del análisis de expresión diferencial. Se realizó el análisis para dos comparaciones exploratorias (Mujer Adenocarcinoma - Mujer Control, MA - MC; Hombre Adenocarcinoma - Hombre Control, HA - HC) y la comparación de interés, (MA - MC) - (HA - HC).

		(MA - MC) - (HA - HC)	MA - MC	HA - HC
GSE10072	Up	0	1199	3182
	Down	0	1296	2688
GSE19188	Up	0	2348	5828
	Down	0	2111	3830
GSE31210	Up	0	2117	2598
	Down	1	2143	2279
GSE32863	Up	1	5393	2581
	Down	1	4584	2284
GSE63459	Up	0	2409	1561
	Down	0	2064	1611
GSE75037	Up	1	5912	4268
	Down	1	5240	3698
GSE81089	Up	3	2654	3564
	Down	0	3383	4262
GSE87340	Up	1	4887	1875
	Down	3	4958	1841
TCGA-LUAD	Up	1	5861	5397
	Down	0	5569	5269

Los genes significativos con logFC positivo son:

- *EIF1AY*, factor eucariótico de iniciación de la traducción 1A ligado a Y, en GSE75037 y GSE32863.
- *TTY16*, transcrito específico de testículo ligado a Y 16, en GSE81089 y TCGA-LUAD.
- *ZFY-AS1*, ARN antisentido 1 de *ZFY*, en GSE81089.
- *SRY*, región determinadora del sexo del cromosoma Y, en GSE81089.
- *BTG1*, factor BTG de antiproliferación 1, en GSE87340.

Los genes significativos con logFC negativo son:

- *FCN3*, ficolina 3, en GSE31210.
- *TCEA1*, factor de elongación transcripcional A2, en GSE32863.
- *FBXL16*, proteína F box con repeticiones de leucina 16, en GSE75037.
- *GPR135*, receptor acoplado a proteínas G 135, en GSE87340.
- *EFR3B*, homólogo B de *EFR3*, en GSE87340.
- *ZNF853*, proteína dedo de zinc 863, en GSE87340.

5.2.3. Análisis de enriquecimiento funcional

Los resultados del análisis de GSEA para términos GO y rutas KEGG en la comparación de interés (Ecuación 4.1) se resumen en la Tabla 5.12. En ellas se muestran las funciones significativas (p-valor ajustado por BH < 0.05) agrupadas según su LOR, de forma que quedan etiquetadas como *Up* si están sobrerrepresentadas en mujeres, y como *Down* si lo están en hombres. Los términos GO además se encuentran clasificados en procesos biológicos (BP), funciones moleculares (MF) y componentes celulares (CC).

Se realizó un diagrama UpSet [69] para cada uno de los grupos funcionales analizados (BP, MF, CC y rutas KEGG) para cuantificar las funciones compartidas por varios estudios (Figuras 5.7 y 5.8).

Este tipo de diagramas muestran la intersección entre conjuntos. Así, cada fila de la parte inferior del diagrama corresponde a un estudio, y se muestra al lado de su nombre un diagrama de barras con el número de rutas o términos GO significativos encontrados. El gráfico de barras de la parte superior del diagrama indica el tamaño de la intersección de los grupos indicados por puntos en la parte inferior del diagrama. Así, cada punto o conjunto de puntos conectados se corresponderían con un área diferente en un diagrama de Venn.

Tabla 5.12: Resumen del resultado del análisis de enriquecimiento funcional. Los términos *Up* y *Down* hacen referencia al grupo de la comparación en el que las funciones están sobrerrepresentadas, *Up* en mujeres y *Down* en hombres.

	Términos GO BP significativos			Términos GO MF significativos			Términos GO CC significativos			Rutas KEGG significativas		
	Up	Down	Total	Up	Down	Total	Up	Down	Total	Up	Down	Total
GSE10072	26	153	179	29	8	37	16	40	56	1	5	6
GSE19188	7	12	19	0	3	3	11	20	31	1	4	5
GSE31210	17	0	17	0	0	0	7	0	7	10	1	11
GSE32863	341	57	398	15	5	20	23	45	68	17	6	23
GSE63459	0	26	26	0	3	3	8	27	35	1	4	5
GSE75037	210	45	255	8	2	10	18	32	50	21	6	27
GSE81089	2	1	3	7	1	8	7	4	11	1	1	2
GSE87340	178	62	240	3	0	3	48	13	61	26	1	27
TCGA	294	228	522	28	30	58	21	70	91	30	17	47

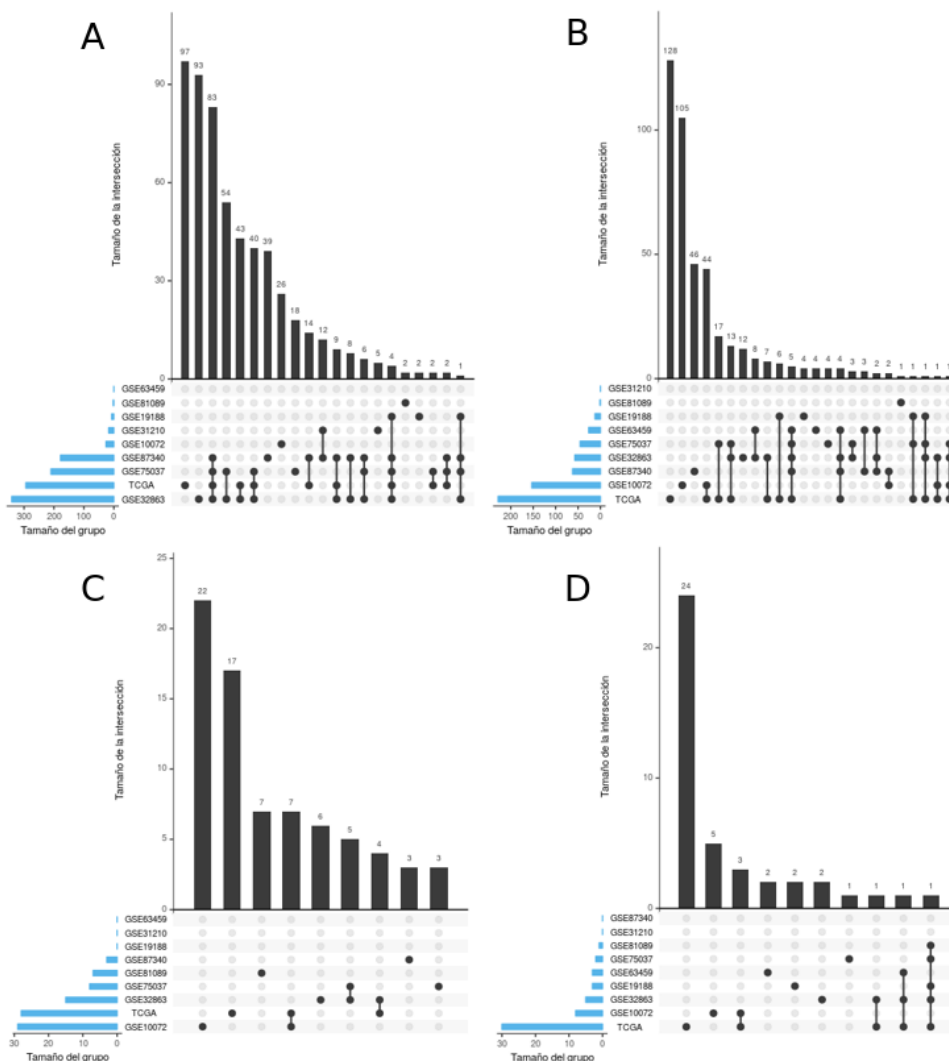


Figura 5.7: Diagrama UpSet de los términos GO BP y MF. (A) BP catalogados como *Up*, (B) BP catalogados como *Down*, (C) MF catalogados como *Up* y (D) MF catalogados como *Down*.

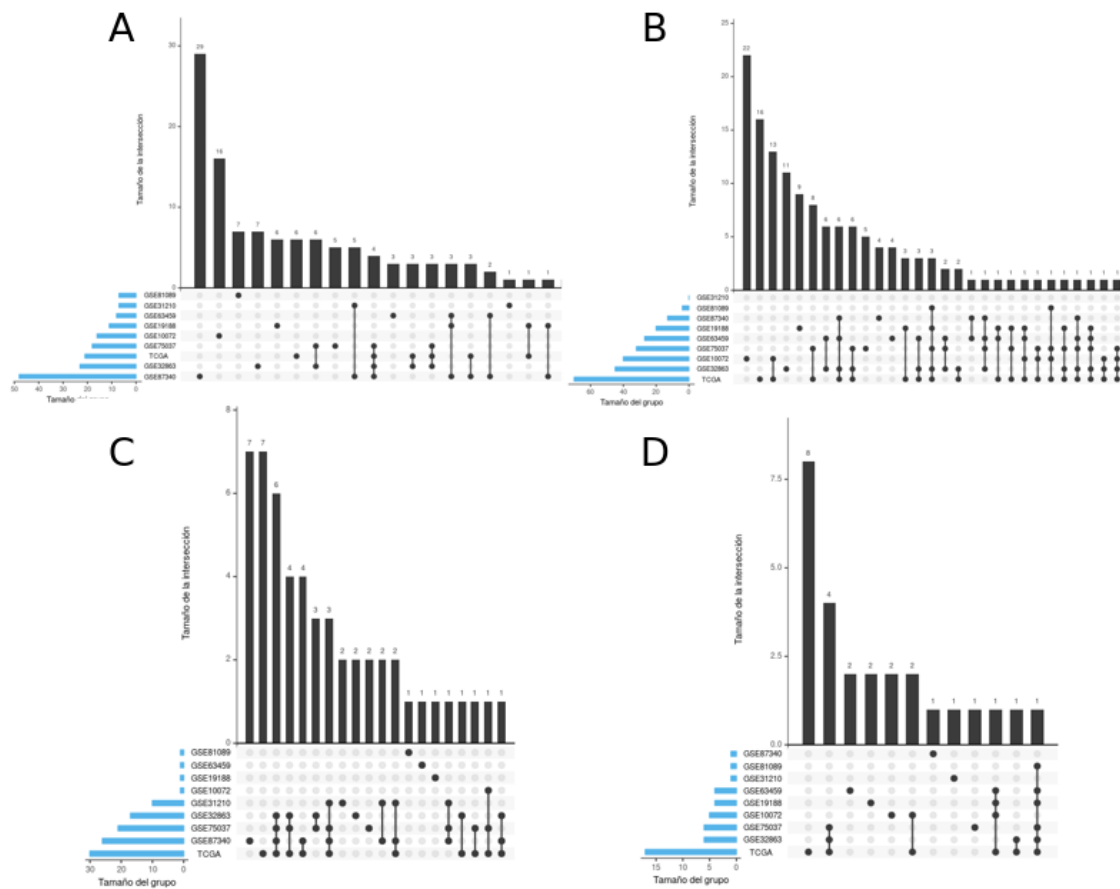


Figura 5.8: Diagrama UpSet de los términos GO CC y las rutas KEGG. (A) CC catalogados como *Up*, (B) CC catalogados como *Down*, (C) rutas KEGG catalogados como *Up* y (D) rutas KEGG catalogados como *Down*.

5.2.4. Análisis de rutas de señalización

En el análisis de rutas de señalización, realizado con Hipathia para la comparación de interés (Ecuación 4.1), no se obtuvo ninguna ruta que tuviera un p-valor ajustado por BH inferior a 0.05.

5.3. Metaanálisis

El metaanálisis funcional se realizó de forma independiente para los términos GO de las diferentes categorías (BP, MF y CC) y las rutas KEGG. En general, se encontró una mayor sobrerepresentación de términos GO en mujeres, y no hubo resultados significativos (p -valor ajustado por BH menor a 0.05) para componentes celulares (Figura 5.9).

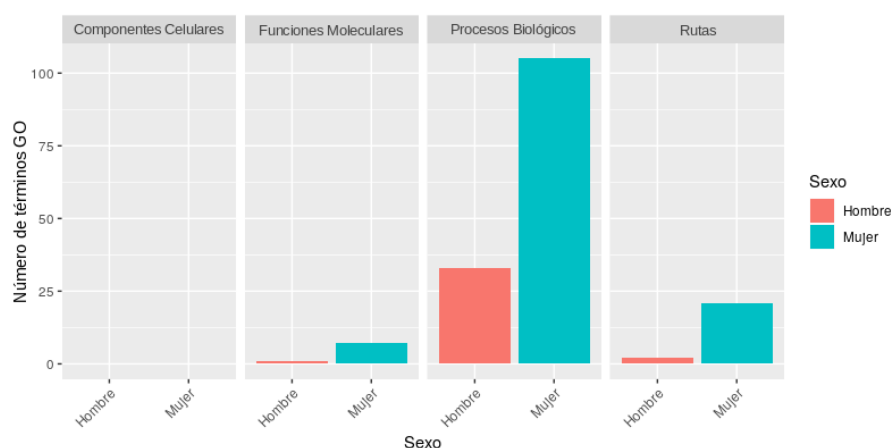


Figura 5.9: Número de funciones y rutas significativas. (Se muestran distribuidas en función del sexo en el que están sobrerepresentadas (valor del LOR) y por categoría analizada (BP, MF, CC y rutas KEGG)).

5.3.1. Términos GO para procesos biológicos

La Tabla 5.13 muestra el resumen de los resultados obtenidos para el metaanálisis de los términos GO BP en cada uno de los cuatro métodos de metaanálisis utilizados.

El metaanálisis por el modelo DL mostró un total de 105 términos GO asociados a procesos biológicos significativamente sobrerepresentados en mujeres y 33 en hombres. La Tabla 5.14 muestra un extracto de estos términos significativos junto a sus estimadores del metaanálisis. La totalidad de los BP significativos junto con sus estimadores de la medida de efecto puede encontrarse en el Anexo II, Tabla B.2.

El LOR y el intervalo de confianza para cada estudio individual, y el global del metaanálisis, para cada uno de los términos GO de la Tabla 5.14 pueden observarse en las Figuras 5.10 a 5.16.

Tabla 5.13: Resultados del metaanálisis para los términos GO BP. *Sig. Up* y *Sig. Down* hacen referencia a las funciones que presentan un p-valor ajustado por BH menor a 0.05, mientras que *Sig. LOR-Up* y *Sig. LOR-Down* hacen referencia a aquellas que además presentan una magnitud de LOR mayor a 0.5.

Método	Up	Down	Sig. Up	Sig. Down	Sig. LOR-Up	Sig. LOR-Down
DL	3414	2916	105	33	4	4
HE	3414	2915	100	33	4	3
HS	3414	3914	217	60	5	8
FE	3414	2906	909	420	10	11

Las Figuras 5.14, 5.15 y 5.16 muestran la información funcional solo de aquellos estudios en los que se ha podido medir el efecto. En el caso de la Figura 5.16, por ejemplo, faltan todos los estudios cuyos datos provienen de microarrays. Esto ocurre cuando ninguno de los genes analizados en esos estudios está asociado al término GO analizado, por lo que el estudio no puede aportar ningún valor al metaanálisis de ese término concreto.

Los diagramas de embudo de los términos mostrados en la Tabla 5.14 (Figuras 5.17 y 5.18, y Anexo I Figuras A.15 a A.19) muestran cuántos estudios de los analizados se encuentran en la región de confianza, delimitada por la zona blanca. En el caso del término GO:0002438 (Figura 5.17) se observa que un estudio aparece fuera de la región de confianza, en concreto TCGA-LUAD. Esto indica que en este estudio, esta función es más variable en su LOR que el resto de estudios. Esto ocurre de forma puntual en algunos términos analizados, para diferentes estudios, y en la mayoría de ocasiones todos los estudios se encuentran dentro de la región de confianza, como en la Figura 5.18.

Dada la cantidad de términos GO de procesos biológicos significativos encontrados, se resumieron con REVIGO [70]. En la Figura 5.19 se puede observar que se ha detectado una gran cantidad de procesos biológicos relacionados con la respuesta inmunitaria sobrerrepresentados en mujeres ($LOR > 0$). También se ha encontrado una mayor representación de términos relacionados con la señalización purinérgica en mujeres, y de términos relacionados con la regulación de la reparación del ADN y la transición de las fases del ciclo celular en hombres.

Tabla 5.14: Selección de términos GO BP significativos en el metaanálisis. Se muestran los estimadores de la medida de efecto y los indicadores de la heterogeneidad.

ID	Nombre	LI 95%	LOR	LS 95%
GO:0002863	<i>positive regulation of inflammatory response to antigenic stimulus</i>	0.317	0.6	0.883
GO:0002438	<i>acute inflammatory response to antigenic stimulus</i>	0.119	0.294	0.469
GO:0045410	<i>positive regulation of interleukin-6 biosynthetic process</i>	0.235	0.430	0.626
GO:0043374	<i>CD8-positive, alpha-beta T cell differentiation</i>	0.241	0.479	0.716
GO:1902857	<i>positive regulation of non-motile cilium assembly</i>	-0.942	-0.591	-0.239
GO:0006228	<i>UTP biosynthetic process</i>	-0.818	-0.516	-0.215
GO:0006183	<i>GTP biosynthetic process</i>	-0.685	-0.543	-0.222

ID	FDR	QE	QEp	SE	τ^2	I ²	H ²
GO:0002863	0.01	17.011	0.03	0.144	0.099	52.972	2.126
GO:0002438	0.047	12.496	0.13	0.089	0.026	35.98	1.562
GO:0045410	0.039	8.014	0.432	0.096	0.000	0.179	1.002
GO:0043374	0.018	14.325	0.074	0.121	0.058	44.155	1.791
GO:1902857	0.046	0.075	0.963	0.179	0.000	0.000	1.000
GO:0006228	0.043	2.034	0.565	0.154	0.000	0.000	1.000
GO:0006183	0.023	6.627	0.357	0.118	0.009	9.459	1.104

GO:0002438 (*acute inflammatory response to antigenic stimulus*)

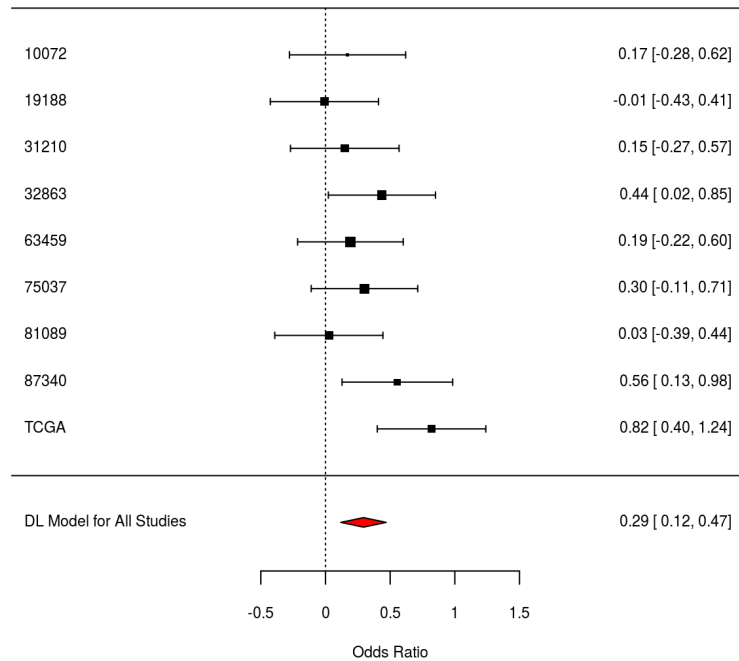


Figura 5.10: Gráfico de bosque para el término GO:0002438.

GO:0002863 (*positive regulation of inflammatory response to antigenic stimulus*)

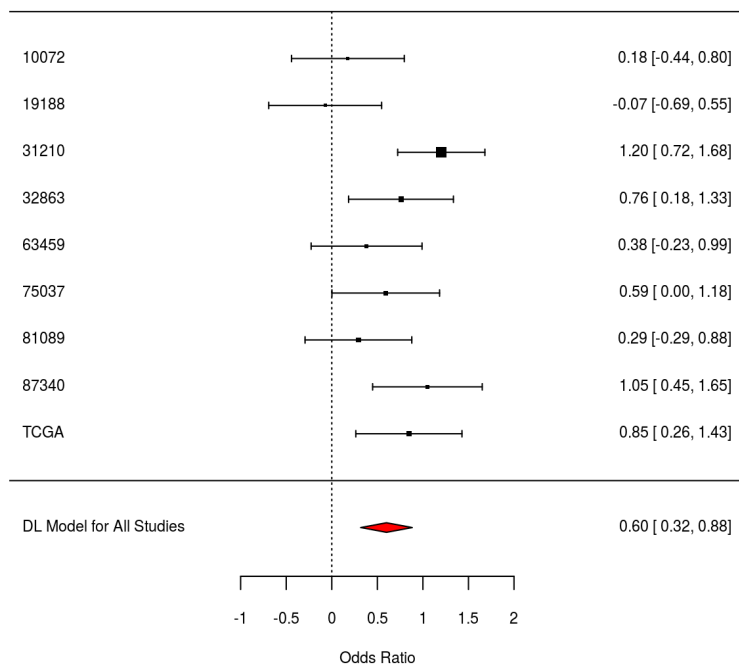
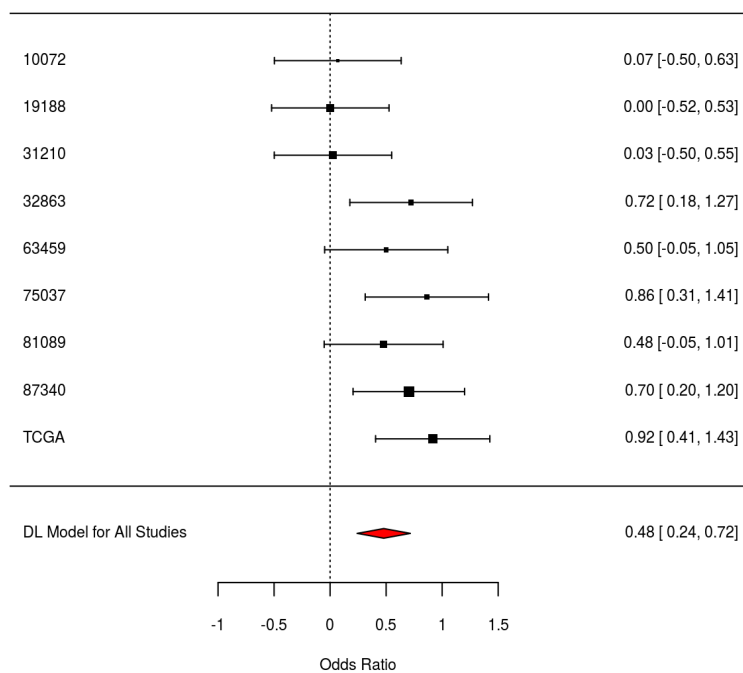
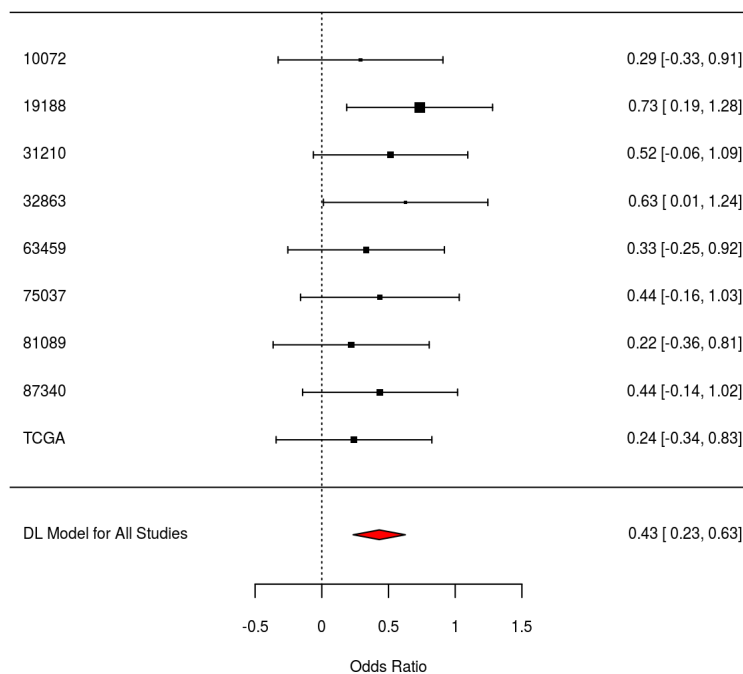


Figura 5.11: Gráfico de bosque para el término GO:0002863.

GO:0043374 (*CD8-positive, alpha-beta T cell differentiation*)**Figura 5.12: Gráfico de bosque para el término GO:0043374.****GO:0045410** (*positive regulation of interleukin-6 biosynthetic process*)**Figura 5.13: Gráfico de bosque para el término GO:0045410.**

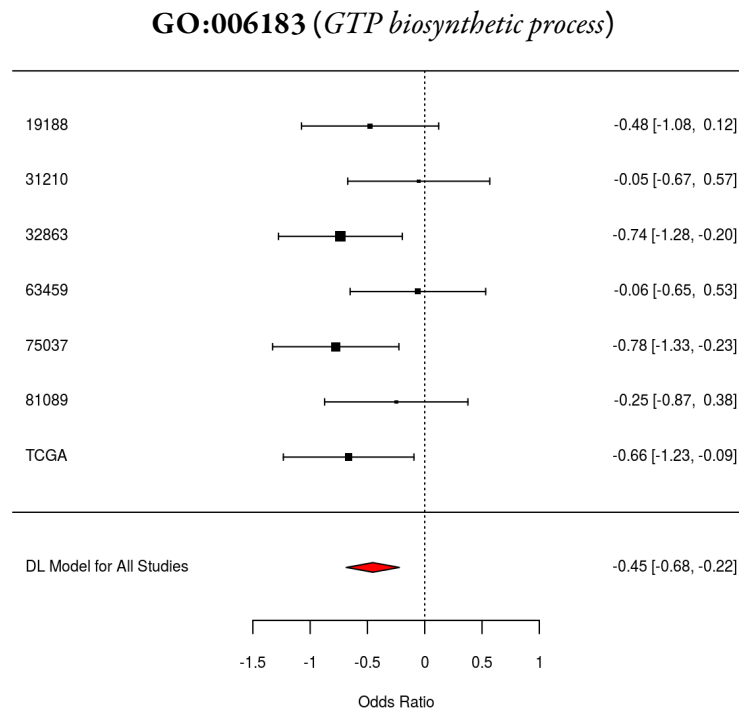


Figura 5.14: Gráfico de bosque para el término GO:0006183.

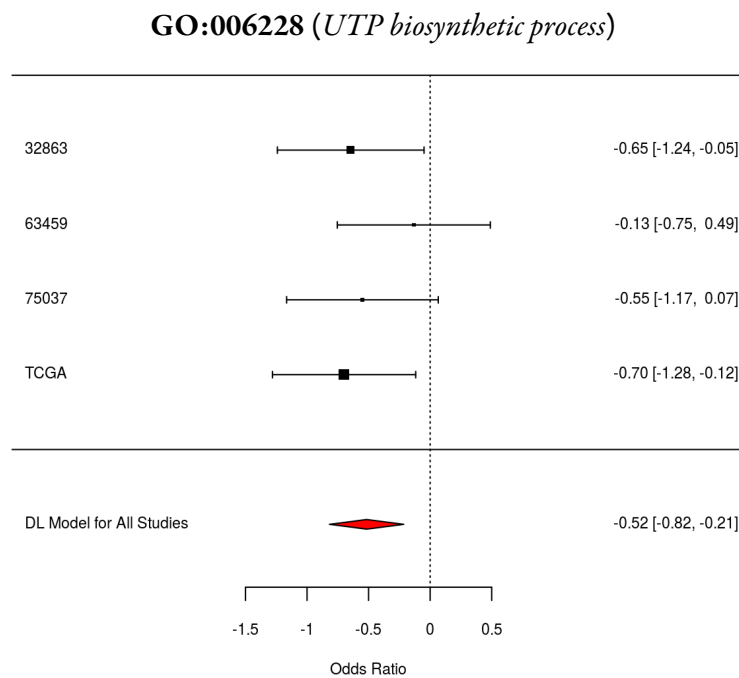


Figura 5.15: Gráfico de bosque para el término GO:0006228.

GO:1902857 (*positive regulation of non-motile cilium assembly*)

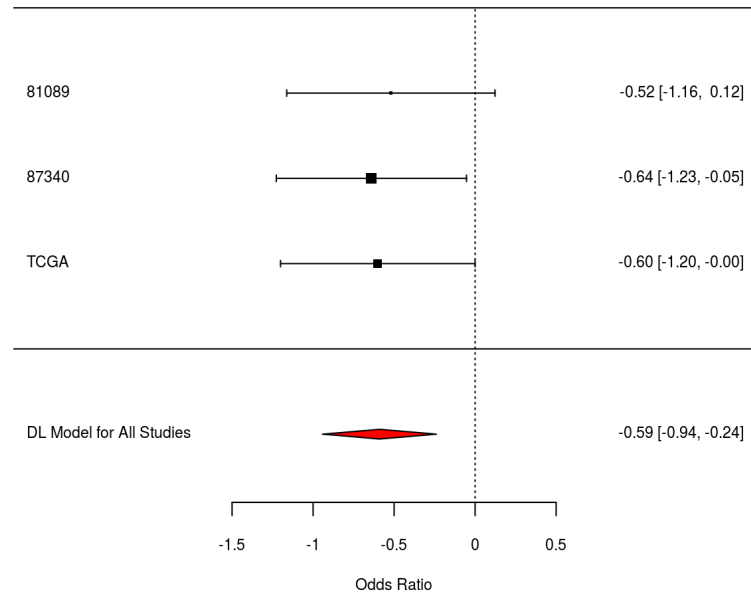


Figura 5.16: Gráfico de bosque para el término GO:1902857.

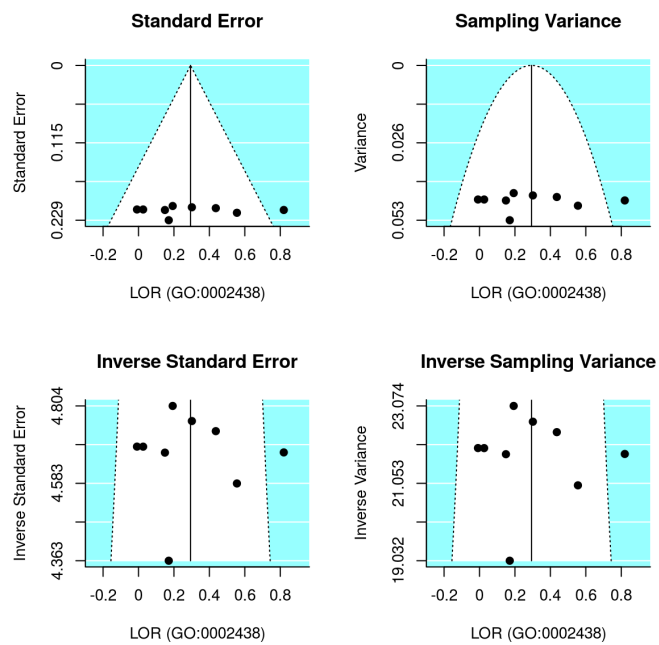


Figura 5.17: Gráficos de embudo para el término GO:0002438.

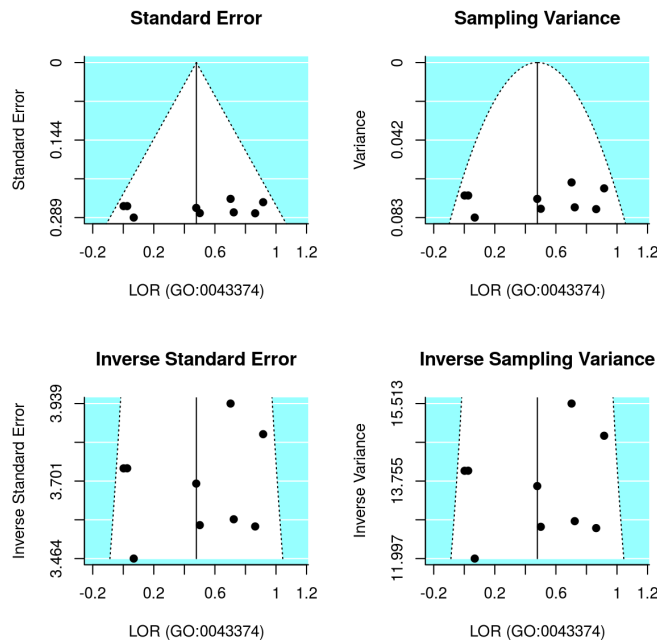


Figura 5.18: Gráficos de embudo para el término GO:0043374.

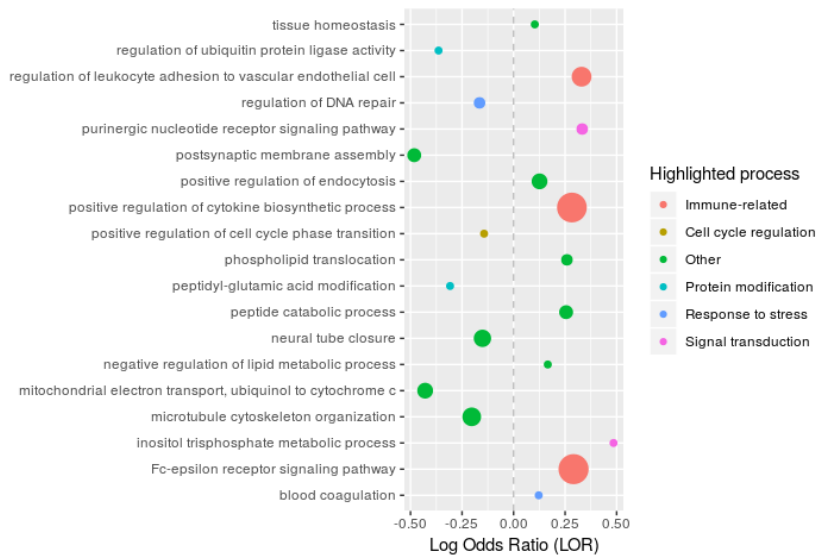


Figura 5.19: Resumen de los términos GO BP significativos en el metaanálisis. El tamaño de los puntos representa la cantidad de términos GO resumidos en ese término en concreto. El LOR indica la magnitud de la diferencia entre hombres y mujeres, siendo positivo si está sobrerrepresentado en mujeres y negativo si está sobrerrepresentado en hombres.

5.3.2. Términos GO para funciones moleculares

La Tabla 5.15 muestra el resumen de los resultados obtenidos para el metaanálisis de los términos GO MF en cada uno de los cuatro métodos de metaanálisis utilizados.

Tabla 5.15: Resultados del metaanálisis para los términos GO BP. *Sig. Up* y *Sig. Down* hacen referencia a las funciones que presentan un p-valor ajustado por BH menor a 0.05, mientras que *Sig. LOR-Up* y *Sig. LOR-Down* hacen referencia a aquellas que además presentan una magnitud de LOR mayor a 0.5.

Método	Up	Down	Sig. Up	Sig. Down	Sig. LOR-Up	Sig. LOR-Down
DL	654	539	7	1	0	0
HE	654	538	7	1	0	0
HS	653	540	9	1	1	0
FE	651	539	83	35	1	1

El metaanálisis por el modelo DL mostró un total de 7 términos GO asociados a funciones moleculares significativamente sobrerrepresentados en mujeres y 1 en hombres. Los términos significativos pueden consultarse en la Tabla 5.16, junto a sus estimadores de la medida de efecto e indicadores de la heterogeneidad.

El LOR y el intervalo de confianza para cada estudio individual, y el global del metaanálisis, para algunos de los términos GO de la Tabla 5.16 pueden observarse en las Figuras 5.20 a 5.23, así como sus diagramas de embudo en el Anexo I, Figuras A.20 a A.23.

Tabla 5.16: Términos GO MF significativos en el metaanálisis. Se muestran los estimadores de la medida de efecto y los indicadores de la heterogeneidad.

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0001608	<i>G protein-coupled nucleotide receptor activity</i>	0,158	0,332	0,505	0,031
GO:0001614	<i>purinergic nucleotide receptor activity</i>	0,193	0,355	0,518	0,011
GO:0004126	<i>cytidine deaminase activity</i>	0,196	0,427	0,658	0,045
GO:0005126	<i>cytokine receptor binding</i>	0,057	0,116	0,176	0,031
GO:0016502	<i>nucleotide receptor activity</i>	0,193	0,355	0,518	0,011
GO:0035586	<i>purinergic receptor activity</i>	0,182	0,351	0,521	0,016
GO:0045028	<i>G protein-coupled purinergic nucleotide receptor activity</i>	0,158	0,332	0,505	0,031
GO:0048407	<i>platelet-derived growth factor binding</i>	-0,734	-0,494	-0,254	0,016

ID	QE	QEp	SE	τ^2	I ²	H ²
GO:0001608	7,347	0,5	0,089	0	0	1
GO:0001614	10,63	0,224	0,083	0,015	24,741	1,329
GO:0004126	12,002	0,151	0,118	0,042	33,347	1,5
GO:0005126	16,832	0,032	0,03	0,004	52,471	2,104
GO:0016502	10,63	0,224	0,083	0,015	24,741	1,329
GO:0035586	13,652	0,091	0,086	0,028	41,401	1,707
GO:0045028	7,347	0,5	0,089	0	0	1
GO:0048407	13,109	0,108	0,122	0,052	38,973	1,639

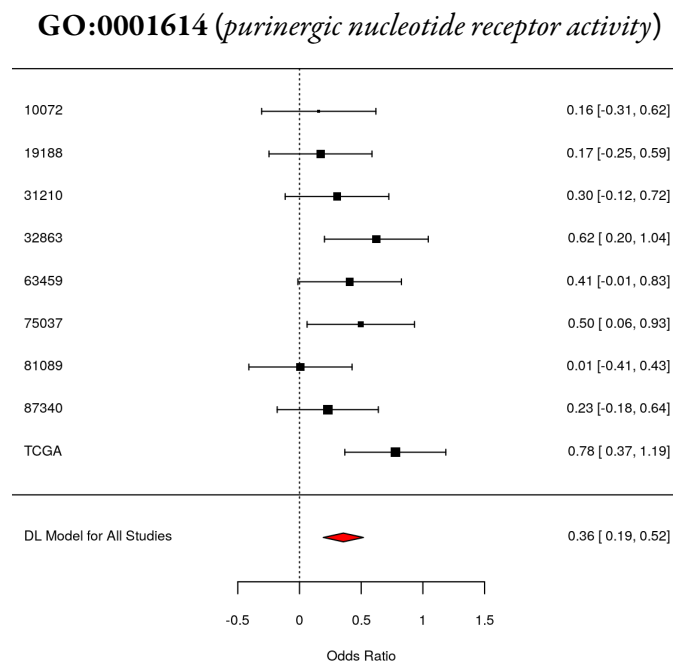


Figura 5.20: Gráfico de bosque para el término GO:0001614.

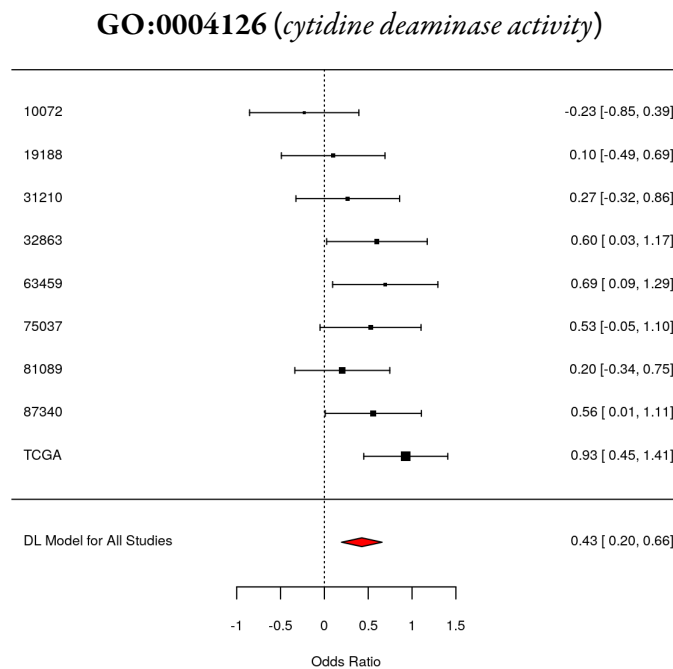


Figura 5.21: Gráfico de bosque para el término GO:0004126.

GO:0045028 (*G protein-coupled purinergic nucleotide receptor activity*)

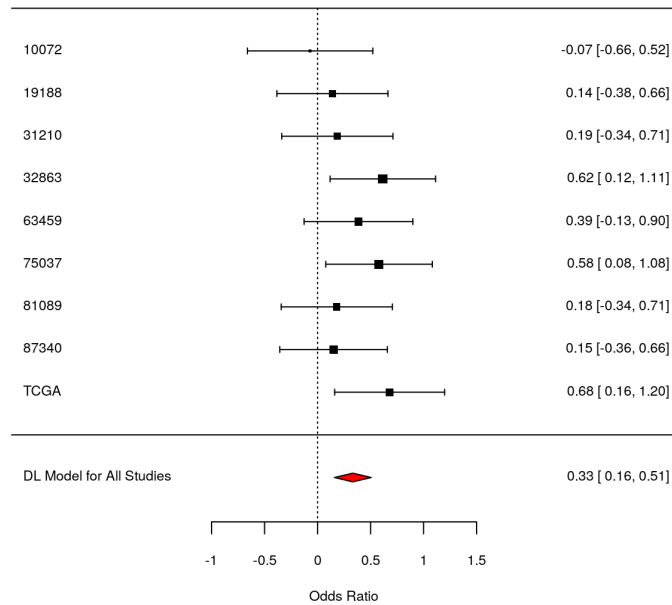


Figura 5.22: Gráfico de bosque para el término GO:0045028.

GO:0048407 (*platelet-derived growth factor binding*)

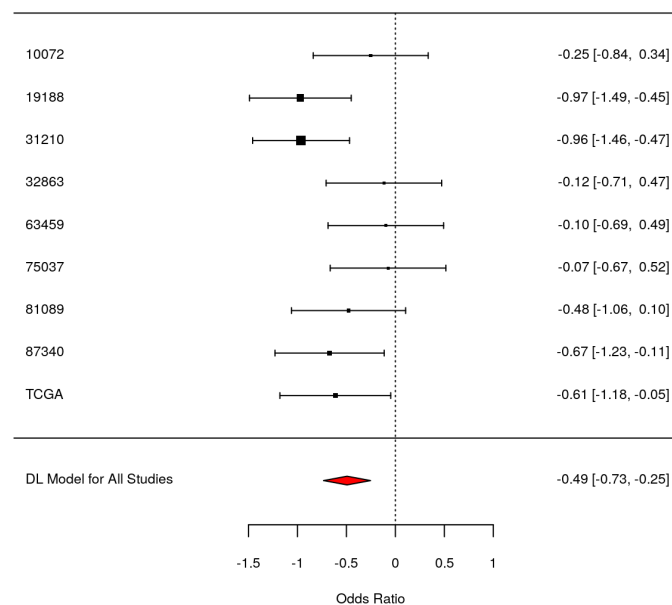


Figura 5.23: Gráfico de bosque para el término GO:0048407.

5.3.3. Rutas KEGG

La Tabla 5.17 muestra el resumen de los resultados obtenidos para el metaanálisis de las rutas KEGG en cada uno de los cuatro métodos de metaanálisis utilizados.

Tabla 5.17: Resultados del metaanálisis para los términos GO BP. *Sig. Up* y *Sig. Down* hacen referencia a las funciones que presentan un p-valor ajustado por BH menor a 0.05, mientras que *Sig. LOR-Up* y *Sig. LOR-Down* hacen referencia a aquellas que además presentan una magnitud de LOR mayor a 0.5.

Método	Up	Down	Sig. Up	Sig. Down	Sig. LOR-Up	Sig. LOR-Down
DL	136	76	21	2	0	0
HE	136	76	21	2	0	0
HS	136	76	32	2	0	0
FE	135	77	55	13	0	0

El metaanálisis por el modelo DL mostró un total de 21 rutas KEGG significativamente sobrerrepresentados en mujeres y 2 en hombres. Un extracto de los términos significativos pueden consultarse en la Tabla 5.18, sus estimadores de la medida de efecto e indicadores de la heterogeneidad. El resto de términos pueden consultarse en el Anexo II, Tabla B.3.

El LOR y el intervalo de confianza para cada estudio individual, y el global del metaanálisis, para algunas de las rutas KEGG de la Tabla 5.18 pueden observarse en las Figuras 5.24 a 5.27, así como sus diagramas de embudo en el Anexo I, Figuras A.24 a A.27.

Tabla 5.18: Selección de rutas KEGG significativas en el metaanálisis. Se muestran los estimadores de la medida de efecto y los indicadores de la heterogeneidad.

ID	Nombre	LI 95%	LOR	LS 95%	FDR
hsa:04210	<i>Apoptosis</i>	0,089	0,16	0,231	0,002
hsa:04330	<i>Notch signaling pathway</i>	-0,253	-0,157	-0,06	0,025
hsa:04621	<i>NOD-like receptor signaling pathway</i>	0,094	0,191	0,288	0,006
hsa:04622	<i>RIG-I-like receptor signaling pathway</i>	0,088	0,181	0,275	0,006
hsa:04660	<i>T cell receptor signaling pathway</i>	0,064	0,2	0,337	0,039
hsa:04670	<i>Leukocyte transendothelial migration</i>	0,037	0,116	0,196	0,039
hsa:05160	<i>Hepatitis C</i>	0,034	0,092	0,151	0,03
hsa:05320	<i>Autoimmune thyroid disease</i>	0,126	0,382	0,638	0,039

ID	QE	QEp	SE	τ^2	I ²	H ²
hsa:04210	3,565	0,894	0,036	0	0	1
hsa:04330	7,277	0,507	0,049	0	0	1
hsa:04621	9,893	0,273	0,05	0,004	19,133	1,237
hsa:04622	10,406	0,238	0,048	0,005	23,118	1,301
hsa:04660	36,742	0	0,07	0,034	78,226	4,593
hsa:04670	13,048	0,11	0,04	0,006	38,687	1,631
hsa:05160	4,501	0,809	0,03	0	0	1
hsa:05320	54,808	0	0,13	0,13	85,404	6,851

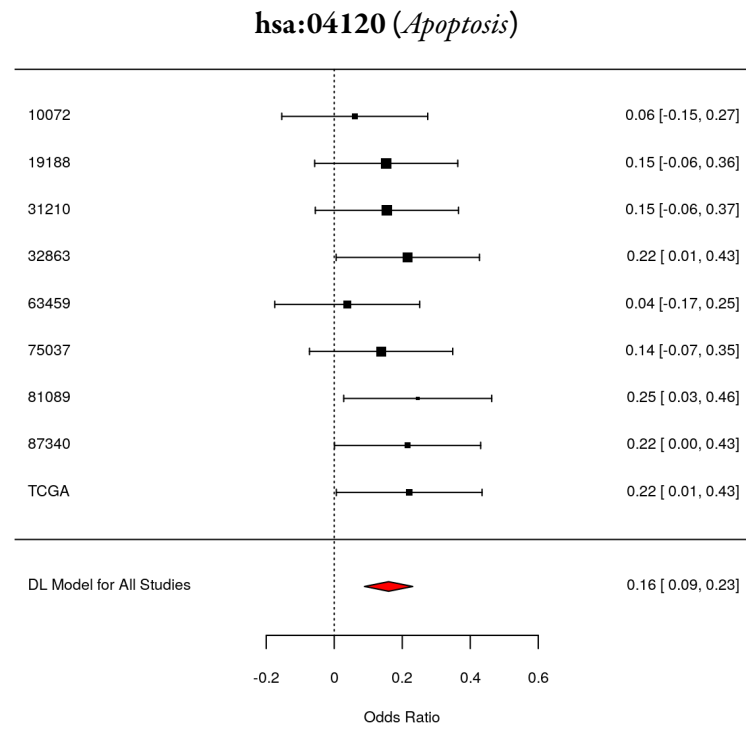


Figura 5.24: Gráfico de bosque para la ruta KEGG hsa:04120.

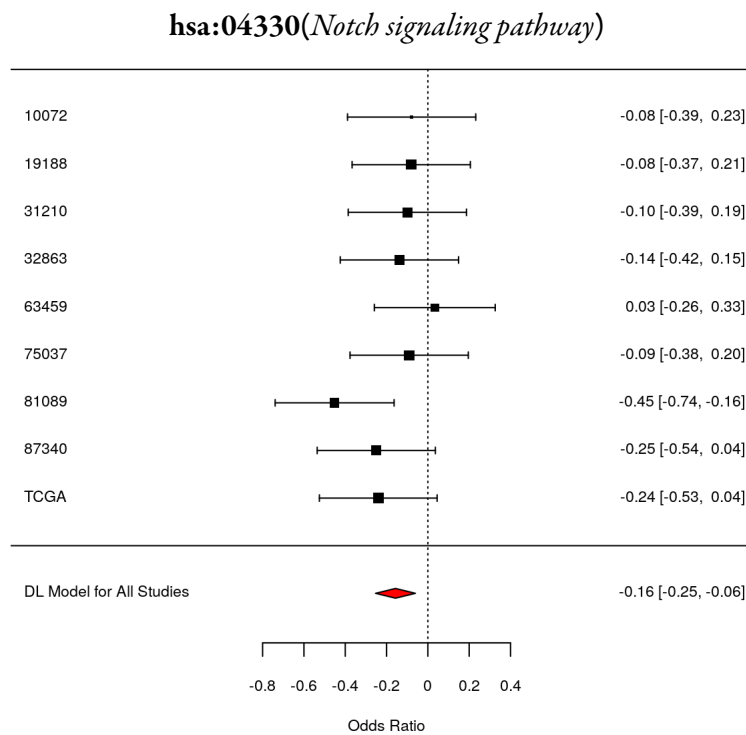


Figura 5.25: Gráfico de bosque para la ruta KEGG hsa:04330.

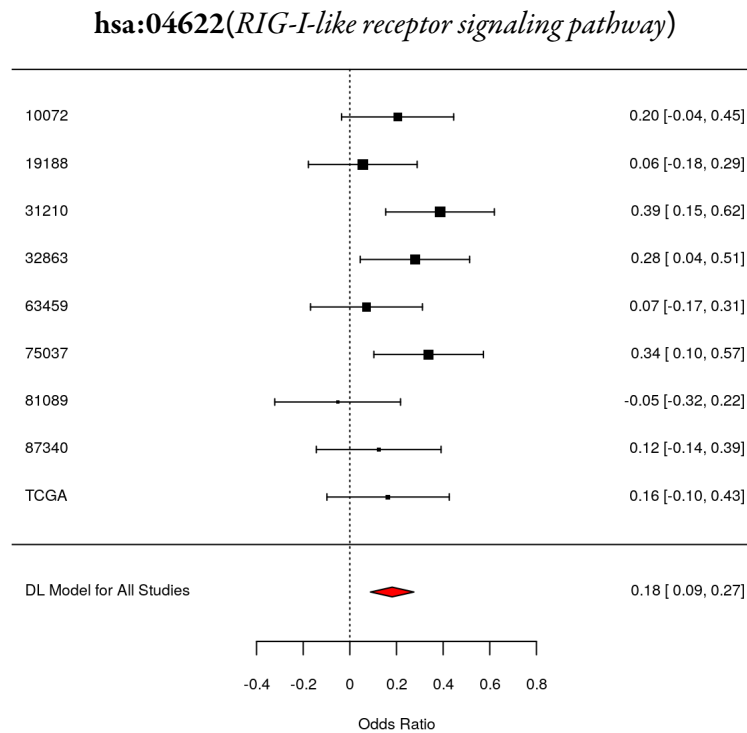


Figura 5.26: Gráfico de bosque para la ruta KEGG hsa:04622.

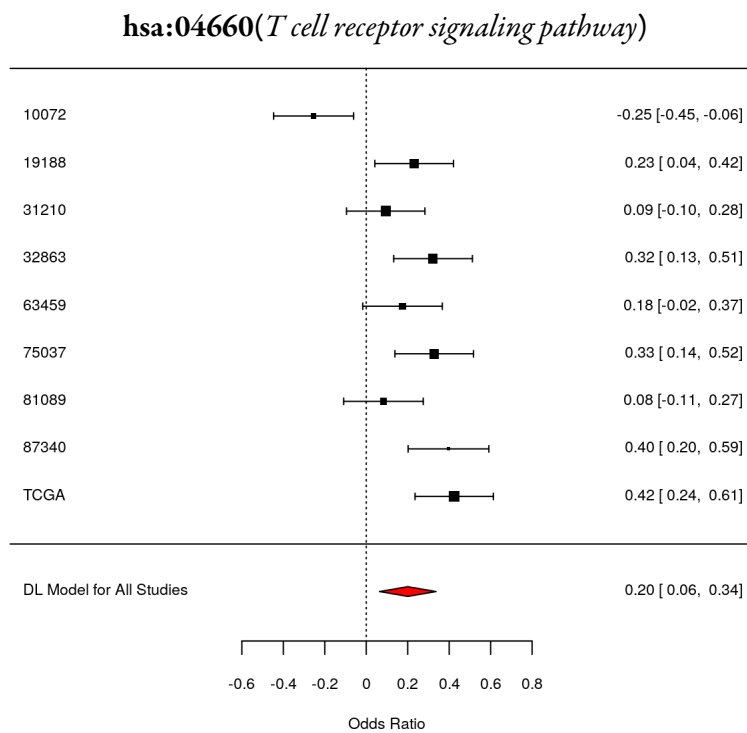


Figura 5.27: Gráfico de bosque para la ruta KEGG hsa:04660.

6. Discusión

El cáncer de pulmón de células no pequeñas, incluido el subtipo adenocarcinoma, presenta diferencias a nivel de incidencia, prevalencia y severidad específicas en hombres y mujeres [1, 7, 9, 2, 71], al igual que muchas otras patologías, como las enfermedades autoinmunes [72] o los desórdenes psicológicos [73]. A pesar de estas diferencias, las mujeres en general están infrarrepresentadas en la investigación biomédica [74], de forma que muchas investigaciones no incorporan la variable sexo en el estudio. Dotar a la investigación biomédica de esta perspectiva es esencial, así como el análisis de los resultados de forma específica al sexo, ya que estas diferencias entre hombres y mujeres afectan a los biomarcadores que presenta cada sexo (a la hora de diagnosticar una enfermedad o su severidad, por ejemplo), o a la respuesta a medicamentos y terapias [74].

Debido a la falta de concienciación de la comunidad científica sobre la importancia de incluir una muestra representativa de ambos sexos [74], hay una escasez de estudios con datos de expresión génica disponibles en las bases de datos que contengan esta información. Además, el objetivo inicial de estos estudios no suele ser estudiar las diferencias causadas por el sexo, sino que el sexo queda incorporado como una característica clínica más del paciente. Esta falta de perspectiva sobre el sexo en el diseño de los estudios, sumado a la falta de un estándar común en los datos disponibles, dificulta su reutilización y reanálisis. En este sentido, y teniendo en cuenta que la falta de estandarización es un problema que afecta en general a la comunidad científica, expertos de diferentes áreas han propuesto los principios *Findable, Accesible, Interoperable, Reusable* (FAIR) [75]. Estos principios son un requerimiento para aquellas investigaciones en el marco del programa europeo H2020, y son la clave para una mejora de la generación de conocimiento e innovación a partir de la reutilización de datos generados [75, 76]. Por todo ello, para realizar este trabajo hemos tenido que integrar diversos estudios con distintos objetivos, realizados en diferentes plataformas, y sin un estándar de nomenclatura para sus metadatos.

Para abordar las diferencias de sexo en adenocarcinoma de pulmón hemos optado por el metaanálisis, un método robusto para el análisis conjunto de datos que nos permite incrementar el poder estadístico de los hallazgos, mejorar la precisión, y responder a preguntas que no habían sido planteadas en los estudios individuales [56, 36]. Esta técnica nos ha permitido combinar los estudios obtenidos a partir de la búsqueda en bases de datos y generar información conjunta que de otra forma no hubiésemos podido obtener, ni por la intersección ni por la suma de los resultados individuales.

En el metaanálisis funcional a nivel de los términos GO asociados a procesos biológi-

cos hemos detectado un enriquecimiento de términos relacionados con la respuesta inmune en mujeres (Figura 5.19). Los términos GO:0043374, GO:003637, GO:0046635 y GO:0046638 hacen referencia a la activación y diferenciación de linfocitos T CD8+, y la regulación positiva de estos procesos. Los linfocitos CD8+ están asociados al efecto antitumoral del TME [15, 17, 20]. Los niveles de linfocitos CD8+ se han propuesto como biomarcador pronóstico en cáncer de pulmón [77] y en el subtipo adenocarcinoma [78], correlacionado con mayor supervivencia y menor recurrencia.

También se han encontrado sobrerrepresentados en mujeres el término GO:0030890 y la ruta hsa04662, que hacen referencia a la proliferación de linfocitos B y su señalización. En NSCLC cuando estas células están activadas actúan como presentadoras de antígenos ante las CD4+, convirtiéndolas en cooperadoras con acción antitumoral que acaban por estimular a las CD8+ [17, 79].

Aparte de los términos que hacen referencia directamente a los linfocitos T y B, los términos GO:0045086, GO:0032743 y GO:0032623 están relacionados con la producción de interleucina (IL) 2, y con la regulación positiva de su producción y biosíntesis. La IL-2 favorece el crecimiento y la actividad de linfocitos T y B, amplificando la respuesta linfocitaria [80], por lo que una mayor producción de IL-2 podría estar influyendo en una mayor actividad de los linfocitos T CD8+. Además, la IL-2 se ha utilizado como terapia en NSCLC con una mejora significativa de la supervivencia [77]. También se ha detectado sobrerrepresentada en mujeres la ruta KEGG hsa04660 (*T cell receptor signaling pathway*), relacionada con los linfocitos T CD8+ y la producción de IL-2.

La IL-2 forma parte de la familia de las citocinas, proteínas que juegan un papel clave en la comunicación y señalización celular. En mujeres parece haber una mayor regulación de la biosíntesis de citocinas (GO:0042035, hsa04060). Sin embargo, la producción de IL-2 no es la única que aparece sobrerrepresentada. La biosíntesis de citocinas como la interleucina 6 (IL-6, GO:0045410) o la interleucina 17 (IL-17, GO:0032740), y la secreción de interleucina 1-beta (IL-1 β , GO:0050718) también han sido detectadas. La IL-6 se ha correlacionado en varias ocasiones con una peor prognosis en pacientes de NSCLC de estado avanzado [81, 82]. Un análisis de redes en mujeres no fumadoras la ha detectado como un nodo central en la patología [82], lo cual es consistente con nuestros resultados, apuntando a un papel importante de la IL-6 en la progresión tumoral en mujeres pacientes de NSCLC.

La IL-1 se asocia con la progresión tumoral y la metástasis en estado avanzado [83, 84] y la IL-17 además de favorecer la metástasis es capaz de inhibir la apoptosis, evitando la muerte celular [17, 85]. Otra citocina cuya producción ha sido significativa es la interleucina 10 (IL-10), también de efecto positivo para el tumor. Se trata de una citocina inmunosupresora, capaz

de dotar al tumor de la capacidad de escapar al sistema inmunitario [86].

La producción de IFN- γ (GO:0032729) y TNF- α (GO:0032760), también de la familia de las citocinas, tienen efectos antitumorales como la inducción de la apoptosis, la paralización del ciclo celular o la activación de células efectoras del sistema inmune [17, 20, 87, 88]. Sin embargo, su papel es dual, pues interferón- γ puede favorecer la angiogénesis y la tumorigénesis [87], y el TNF la metástasis [89].

Todos los términos GO descritos, sean positivos o negativos para el tumor, señalan una mayor respuesta inflamatoria debida al tumor en mujeres, que influye de manera compleja en el cáncer. En concreto, parece ser una respuesta inflamatoria aguda ante estímulos antigénicos (GO:0002863, GO:0002864), relacionada con la inmunidad celular. Aunque la inflamación crónica favorece al desarrollo tumoral en NSCLC [87], los biomarcadores de inflamación en NSCLC se correlacionan con una mayor supervivencia [90] e incluso en cáncer escamoso de vejiga se usa la inflamación aguda inducida como terapia [91]. En otros estudios en adenocarcinoma de pulmón, el microambiente tumoral con características de inflamación se asocia con una alta carga de neoantígenos clonales, detectándose un enriquecimiento en células T efectoras activadas y una mayor expresión de proteínas asociadas con la presentación de antígenos, la migración (GO:1904996 y hsa04670) y la acción efectora de las células T, y algunos reguladores negativos de su actividad [16], lo cual se corresponde con nuestros resultados en mujeres.

Ya sea con una tendencia hacia la actividad antitumoral o protumoral, nuestros resultados apuntan hacia una mayor activación del sistema inmunitario en las mujeres. Esto podría sugerir que las terapias inmuno-oncológicas dirigidas a activar la respuesta antitumoral serían más efectivas en hombres, mientras que las terapias cuyo objetivo es bloquear la acción protumoral tendrían una mayor repercusión en mujeres.

Parte de los procesos biológicos y funciones moleculares sobrerrepresentadas en mujeres hacen referencia a la señalización purinérgica (GO:0001614, GO:0045028, GO:0035590, GO:0035589, GO:0035588 y GO:0035587). Esta señalización implica a la adenosina y al adenosín trifosfato (ATP), entre otros nucleótidos, que juegan un papel fundamental como mensajeros extracelulares tanto en situaciones de inflamación como en el microambiente tumoral [92]. La señalización purinérgica se transduce mediante toda una variedad de receptores, que en función de la célula en la que se expresan o las concentraciones de ATP pueden favorecer el crecimiento del tumor y la inmunosupresión, o la muerte tumoral y la citotoxicidad [93]. Así, en NSCLC se ha descrito el efecto citotóxico del receptor P2X₄ (mediante la autofagia y la apoptosis inmunogénica), la acción dual del receptor P2X₇ (autofagia, apoptosis inmunogénica y muerte celular por un lado, proliferación y migración celular por otro), y la inducción de la proliferación por parte del receptor A_{2A}R [93, 94].

Por otro lado, otra serie de términos GO apuntan a un mayor crecimiento celular en hombres, con una mayor actividad mitocondrial (GO:0006122) y producción de uridina trifosfato (UTP) y guanosín trifosfato (GTP) (GO:0006228, GO:0006186). El GTP es un metabolito importante en la biosíntesis de macromoléculas, actuando como ladrillo del ARN y el ADN, y funciona como fuente energética en muchos procesos celulares, ambos procesos claves para el cáncer [95]. El término GO:0048407, unión del factor de crecimiento derivado de plaquetas, también apunta a un mayor crecimiento tumoral [96] en hombres. Este factor de crecimiento también está implicado en la curación de heridas, que según Mostertz et al. [71] en NSCLC se encuentra sobreactivada en pacientes hombres de alto riesgo. Además, en mujeres hemos detectado una regulación negativa del metabolismo de lípidos (GO:0045833), que está relacionado con un menor crecimiento celular [97], y un una sobrerrepresentación de rutas relacionadas con la apoptosis y la muerte celular (hsa:004210, hsa:004622, hsa:004650).

Con nuestro estudio, hemos sido capaces de detectar diferencias entre hombres y mujeres a nivel funcional en adenocarcinoma de pulmón en estadio temprano. Los estudios de Shi et al. [82] y Araujo et al. [98] apoyan nuestros resultados, mostrando repetidamente un enriquecimiento de términos relacionados con el sistema inmunitario en mujeres. El hecho de que otras funciones celulares no coincidan entre los dos estudios, ni con el nuestro, puede deberse a varios factores:

- En el caso de Shi et al. [82], el estudio se hizo solo en mujeres con NSCLC y mujeres control, sin comparar con los hombres. Nuestra aproximación aporta una mejor visión de cómo afecta el cáncer a hombres y mujeres, y evita positivos comunes a ambos, atribuibles a diferencias biológicas independientes del sexo.
- En ambos estudios mezclan subtipos de NSCLC y estadios de la enfermedad, factores que pueden introducir mucha variabilidad en los datos.
- En nuestro caso, hemos incorporado la información sobre el hábito tabáquico para su corrección en el modelo lineal, pero sería interesante poder realizar el análisis sólo en muestras de estadio temprano de no fumadores. Para ello sería necesario la incorporación a las bases de datos de nuevos estudios con este diseño experimental.

Nuestro estudio abre la puerta a futuras investigaciones acerca del efecto del sistema inmunitario en mujeres, poniendo un punto de atención en el papel dual de la mayor parte de los participantes del microambiente tumoral. Sería especialmente interesante combinar nuevas técnicas, como la secuenciación de ARN de célula única, para observar qué genes y funciones se están sobreexpresando en las células tumorales, y cuales en las del sistema inmunitario en mujeres y así diseñar terapias dirigidas a evitar el crecimiento tumoral y la inmunosupresión.

Por otro lado, queda patente la existencia de diferencias biológicas entre hombres y mujeres en el adenocarcinoma de pulmón, que pueden condicionar la supervivencia del paciente, la agresividad del tumor, o la resistencia a terapias. Incorporar este enfoque al diseño experimental en biomedicina junto con la aplicación de los principios FAIR a los datos generados es necesario para poder generar unos resultados en sintonía con la medicina personalizada y de precisión.

7. Conclusiones

1. El metaanálisis es un método robusto para la evaluación e integración de datos provenientes de una revisión sistemática de estudios ómicos.
2. Hemos detectado una sobrerrepresentación de procesos biológicos y rutas metabólicas relacionadas con la respuesta inmunitaria y la inflamación en mujeres, previamente asociados con la enfermedad y, en algunos casos concretos, con las pacientes.
3. Algunos procesos biológicos y funciones moleculares relacionados con la señalización purinérgica, vinculados con la patología y la respuesta inmunitaria, están significativamente sobreexpresados en mujeres.
4. Hemos descrito la sobrerrepresentación de términos relacionados con el crecimiento y el avance del ciclo celular en hombres, mientras que otros términos parecen apuntar a un menor crecimiento celular en mujeres.
5. La estrategia propuesta en este trabajo permite detectar y caracterizar diferencias debidas al sexo en la salud mediante abordajes ómicos con un enfoque estadístico robusto, incorporando esta perspectiva necesaria para la medicina personalizada y de precisión en los estudios biomédicos.

Bibliografía

- [1] Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**, 394–424 (2018).
- [2] North, C. M. & Christiani, D. C. Women and Lung Cancer: What is New? *Seminars in Thoracic and Cardiovascular Surgery* **25**, 87–94 (2013).
- [3] National Cancer Institute, N. What is Cancer? (2019). URL <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer>.
- [4] Weinberg, R. A. *The Biology of Cancer* (Garland Science, New York, NY, 2014), 2 edn.
- [5] Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K. K. Non-small-cell lung cancers: A heterogeneous set of diseases. *Nature Reviews Cancer* **14**, 535–546 (2014).
- [6] Wood, S. L., Pernemalm, M., Crosbie, P. A. & Whetton, A. D. Molecular histology of lung cancer: From targets to treatments. *Cancer Treatment Reviews* **41**, 361–375 (2015).
- [7] Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers - A different disease. *Nature Reviews Cancer* **7**, 778–790 (2007).
- [8] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- [9] Barta, J. A., Powell, C. A. & Wisnivesky, J. P. Global Epidemiology of Lung Cancer. *Annals of Global Health* **85** (2019).
- [10] Hecht, S. S. Lung carcinogenesis by tobacco smoke. *International Journal of Cancer* **131**, 2724–2732 (2012).
- [11] Landi, M. T. *et al.* Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLOS ONE* **3**, 1–8 (2008).
- [12] Karlsson, A. *et al.* Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. *Clinical Cancer Research* **20**, 4912–4924 (2014).
- [13] Selamat, S. A. *et al.* Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Research* **22**, 1197–1211 (2012).

-
- [14] World Health Organization. Tobacco control (2016). URL <http://apps.who.int/gho/data/view.main.SDG3aWHOREGv?lang=en>.
- [15] Jenkins, R. W., Thummalapalli, R., Carter, J., Cañadas, I. & Barbie, D. A. Molecular and Genomic Determinants of Response to Immune Checkpoint Inhibition in Cancer. *Annual Review of Medicine* **69**, 333–347 (2017).
- [16] Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).
- [17] Altorki, N. K. *et al.* The lung microenvironment: an important regulator of tumour growth and metastasis. *Nature Reviews Cancer* **19**, 9–31 (2019).
- [18] Hanahan, D. & Coussens, L. M. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* **21**, 309–322 (2012).
- [19] Palucka, A. K. & Coussens, L. M. The Basis of Oncoimmunology. *Cell* **164**, 1233–1247 (2016).
- [20] Wang, M. *et al.* Role of tumor microenvironment in tumorigenesis. *Journal of Cancer* **8**, 761–773 (2017).
- [21] Alberts, B. *et al.* *Molecular Biology of the Cell* (Garland Science, New York, 2015), 6 edn.
- [22] Nelson, D. L. & Cox, M. M. *Lehninger. Principles of Biochemistry* (W. H. Freeman, New York, 2017), 7 edn.
- [23] Yadav, S. P. The wholeness in suffix *-omics*, *-omes*, and the word *om*. *Journal of Biomolecular Techniques* **18**, 277 (2007).
- [24] Manzoni, C. *et al.* Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics* **19**, 286–302 (2018).
- [25] Rigden, D. J. & Fernández, X. The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. *Nucleic Acids Research* **47**, D1–D7 (2018).
- [26] Zhong Wang, M. G. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics Zhong. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* **10**, 57–63 (2001).
- [27] Doménech-Sáncheza, A. & Vilab, J. Fundamento, tipos y aplicaciones de los arrays de ADN en la microbiología médica. *Enfermedades Infecciosas y Microbiología Clínica* **22**, 46–54 (2004).
-

-
- [28] Schena, M. Genome analysis with gene expression microarrays. *BioEssays* **18**, 427–431 (1996).
- [29] Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).
- [30] García-García, F. *Methods of functional enrichment analysis in genomic studies*. Ph.D. thesis (2016).
- [31] Grant, G. R., Manduchi, E. & Stoeckert Jr., C. J. Analysis and management of microarray gene expression data. *Current Protocols in Molecular Biology* **77**, 19.6.1–19.6.30 (2007).
- [32] Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25 (2000).
- [33] Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
- [34] Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
- [35] Hidalgo, M. R. *et al.* High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget* **8**, 5160–5178 (2017).
- [36] Normand, S. L. T. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 248–287 (1999).
- [37] Liberati, A. *et al.* The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLOS Medicine* **6**, 1–28 (2009).
- [38] Glass, G. V. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher* **5**, 3–8 (1976).
- [39] Chen, R. *et al.* A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer research* **74**, 2892–2902 (2014).
- [40] Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast cancer research* **10**, R65 (2008).
-

-
- [41] Sørli, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* **100**, 8418–8423 (2003).
- [42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).
- [43] Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2012).
- [44] Kolesnikov, N. *et al.* ArrayExpress update—simplifying data submissions. *Nucleic acids research* **43**, D1113–D1116 (2015).
- [45] Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* **375**, 1109–1112 (2016).
- [46] Davis, S. & Meltzer, P. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **14**, 1846–1847 (2007).
- [47] Colaprico, A. *et al.* TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2015).
- [48] Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **39**, D52–D57 (2010).
- [49] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
- [50] Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).
- [51] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
- [52] Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
- [53] Montaner, D. & Dopazo, J. Multidimensional gene set analysis of genomic data. *PLOS ONE* **5**, 1–13 (2010).
-

-
- [54] Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols* **4**, 1184–1191 (2009).
- [55] Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human* (2018). R package version 3.7.0.
- [56] Higgins, J. & Green, S. *Cochrane Handbook for Systematic Reviews of Interventions*, vol. 2011 (2011).
- [57] Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**, 1–48 (2010).
- [58] DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188 (1986).
- [59] Hedges, L. V., Gurevitch, J. & Curtis, P. S. The Meta-Analysis of Response Ratios in Experimental Ecology. *Ecology* **80**, 1150–1156 (1999).
- [60] Schmidt, F. L. & Hunter, J. E. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (SAGE Publications, London, 2015).
- [61] Schwarzer, G., Carpenter, J. R. & Rücker, G. *Meta-Analysis with R. Use R!* (Springer International Publishing, Cham, 2015).
- [62] Hou, J. *et al.* Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and Survival Prediction. *PLOS ONE* **5**, 1–12 (2010).
- [63] Okayama, H. *et al.* Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Research* **72**, 100–111 (2012).
- [64] Yamauchi, M. *et al.* Epidermal Growth Factor Receptor Tyrosine Kinase Defines Critical Prognostic Genes of Stage I Lung Adenocarcinoma. *PLOS ONE* **7**, 1–11 (2012).
- [65] Robles, A. I. *et al.* An integrated prognostic classifier for Stage I lung adenocarcinoma based on mRNA, microRNA, and DNA methylation biomarkers. *Journal of Thoracic Oncology* **10**, 1037–1048 (2015).
- [66] Girard, L. *et al.* An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clinical Cancer Research* **22**, 4880–4889 (2016).
-

-
- [67] Mezheyeuski, A. *et al.* Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *Journal of Pathology* **244**, 421–431 (2018).
- [68] Sun, Z. *et al.* Conserved recurrent gene mutations correlate with pathway deregulation and clinical outcomes of lung adenocarcinoma in never-smokers. *BMC Medical Genomics* **7**, 1–11 (2014).
- [69] Lex, A., Gehlenborg, N., Strobel, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics* **20**, 1983–1992 (2014).
- [70] Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* **6**, 1–9 (2011).
- [71] Mostert, W. *et al.* Age- and Sex-Specific Genomic Profiles in Non-Small Cell Lung Cancer. *JAMA* **303**, 535–543 (2010).
- [72] Whitacre, C. C. Sex differences in autoimmune disease. *Nature Immunology* **2**, 777–780 (2001).
- [73] Green, T., Flash, S. & Reiss, A. L. Sex differences in psychiatric disorders: what we can learn from sex chromosome aneuploidies. *Neuropsychopharmacology* **44**, 9–21 (2019).
- [74] Kim, A. M., Ting, C. M. & Woodruff, T. K. Sex bias in trials and treatment must end. *Nature* **465**, 688–689 (2010).
- [75] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
- [76] European Commission. Guidelines on FAIR Data Management in Horizon 2020. Tech. Rep. (2016).
- [77] Zeng, D.-Q. *et al.* Prognostic and predictive value of tumor-infiltrating lymphocytes for clinical therapeutic research in patients with non-small cell lung cancer. *Oncotarget* **7**, 13765–13781 (2016).
- [78] Ye, S.-L., Li, X.-Y., Zhao, K. & Feng, T. High expression of CD8 predicts favorable prognosis in patients with lung adenocarcinoma. *Medicine* **96**, e6472 (2017).
- [79] Bruno, T. C. Evaluating the antitumor role of B cells in patients with non-small cell lung cancer. *Journal of Clinical Oncology* **35**, 75 (2017).
-

-
- [80] Nelson, B. H. IL-2, Regulatory T Cells, and Tolerance. *The Journal of Immunology* **172**, 3983–3988 (2014).
- [81] Silva, E. M. *et al.* High systemic IL-6 is associated with worse prognosis in patients with non-small cell lung cancer. *PLOS ONE* **12**, 1–12 (2017).
- [82] Shi, K., Li, N., Yang, M. & Li, W. Identification of Key Genes and Pathways in Female Lung Cancer Patients Who Never Smoked by a Bioinformatics Analysis. *Journal of Cancer* **10**, 51–60 (2019).
- [83] Lewis, A. M., Varghese, S., Xu, H. & Alexander, H. R. Interleukin-1 and cancer progression: the emerging role of interleukin-1 receptor antagonist as a novel therapeutic agent in cancer treatment. *Journal of Translational Medicine* **4**, 48 (2006).
- [84] Gottschlich, A., Endres, S. & Kobold, S. Can we use interleukin-1 β blockade for lung cancer treatment? *Translational Lung Cancer Research* **7**, S160–S164 (2018).
- [85] Wu, F., Xu, J. & Huang, Q. The Role of Interleukin-17 in Lung Cancer. *Mediators of Inflammation* **2016**, 6 (2016).
- [86] Vahl, J. M. *et al.* Interleukin-10-regulated tumour tolerance in non-small cell lung cancer. *British journal of cancer* **117**, 1644–1655 (2017).
- [87] Ni, L. & Lu, J. Interferon gamma in cancer immunotherapy. *Cancer Medicine* **7**, 4509–4516 (2018).
- [88] Roberts, N. J., Zhou, S., Diaz Jr, L. A. & Holdhoff, M. Systemic use of tumor necrosis factor alpha as an anticancer agent. *Oncotarget* **2**, 739–751 (2011).
- [89] Shang, G. S., Liu, L. & Qin, Y. W. IL-6 and TNF- α promote metastasis of lung cancer by inducing epithelial-mesenchymal transition. *Oncology Letters* **13**, 4657–4660 (2017).
- [90] Tolia, M. *et al.* Prognostic Significance of Serum Inflammatory Response Markers in Newly Diagnosed Non-Small Cell Lung Cancer before Chemoradiation. *BioMed research international* **2015**, 5 (2015).
- [91] Korniluk, A., Koper, O., Kemon, H. & Dymicka-Piekarska, V. From inflammation to cancer. *Irish Journal of Medical Science* **186**, 57–62 (2017).
- [92] Di Virgilio, F. & Adinolfi, E. Extracellular purines, purinergic receptors and tumor growth. *Oncogene* **36**, 293–303 (2017).
-

-
- [93] Vijayan, D., Smyth, M. J. & Teng, M. W. L. Purinergic Receptors: Novel Targets for Cancer Immunotherapy. In *Oncoimmunology*, 115–141 (Springer International Publishing, Cham, 2018).
- [94] Di Virgilio, F., Sarti, A. C., Falzoni, S., De Marchi, E. & Adinolfi, E. Extracellular ATP and P2 purinergic signalling in the tumour microenvironment. *Nature Reviews Cancer* **18**, 601–618 (2018).
- [95] Sasaki, A. T. Dynamic Role of the GTP Energy Metabolism in Cancers. *The Keio Journal of Medicine* **65**, 21–21 (2016).
- [96] Farooqi, A. A. & Siddik, Z. H. Platelet-derived growth factor (PDGF) signalling in cancer: rapidly emerging signalling landscape. *Cell biochemistry and function* **33**, 257–265 (2015).
- [97] Yano, K. Lipid metabolic pathways as lung cancer therapeutic targets: A computational study. *International Journal of Molecular Medicine* **29**, 519–529 (2012).
- [98] Araujo, J. M. *et al.* Repeated observation of immune gene sets enrichment in women with non-small cell lung cancer. *Oncotarget* **7**, 20282–20292 (2016).
-

A. Anexo I - Figuras

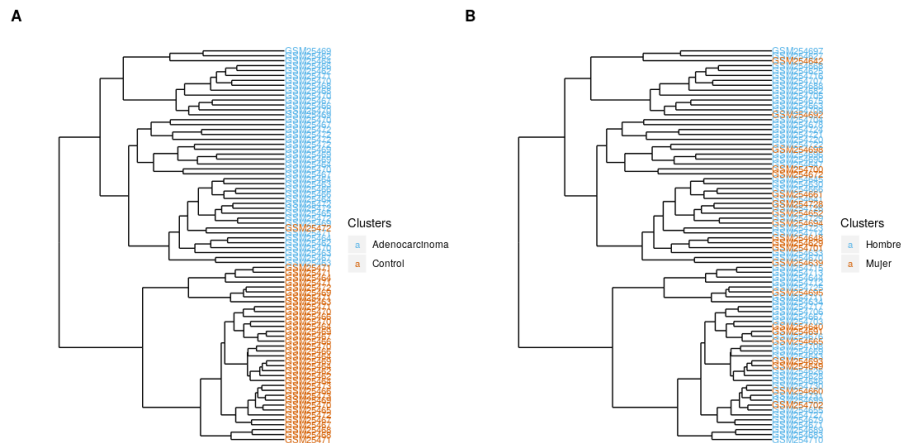


Figura A.1: Clustering exploratorio de las muestras del estudio GSE10072 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

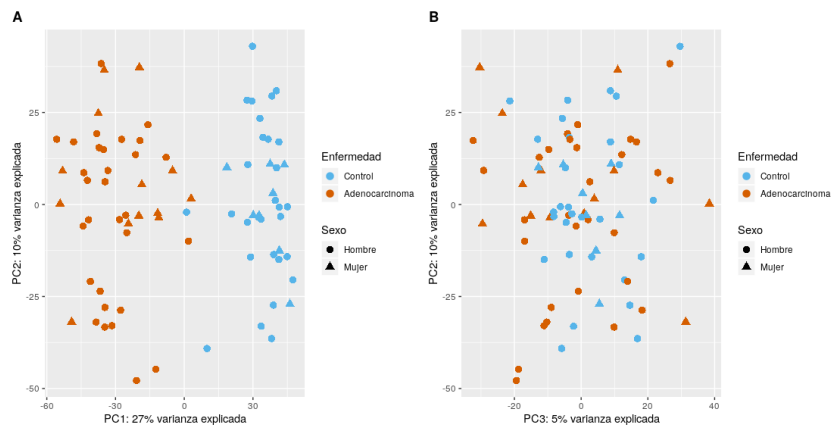


Figura A.2: Análisis de componentes principales de las muestras del estudio GSE10072. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

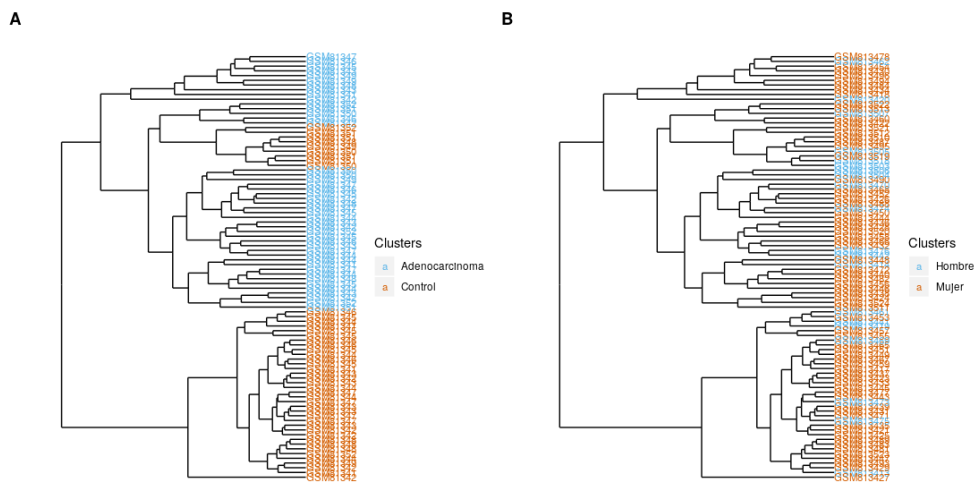


Figura A.3: Clustering exploratorio de las muestras del estudio GSE32863 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

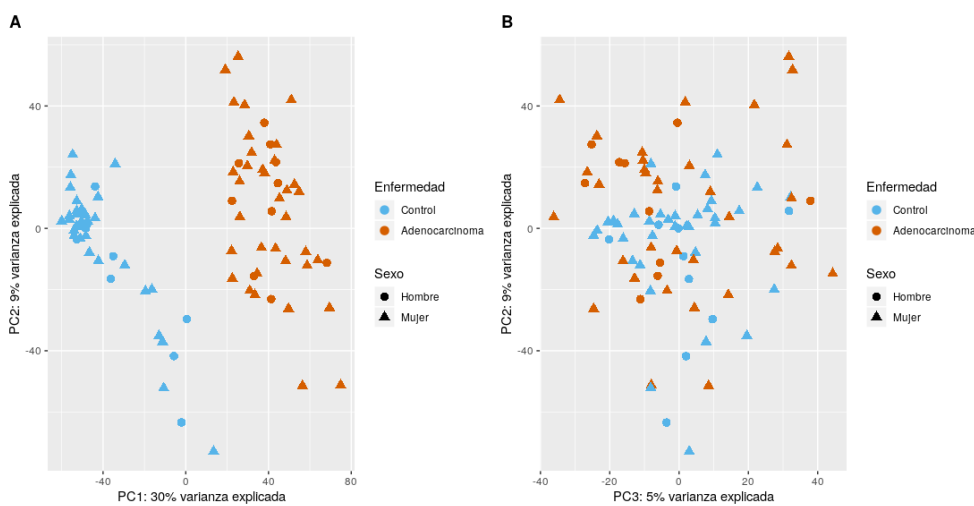


Figura A.4: Análisis de componentes principales de las muestras del estudio GSE32863. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

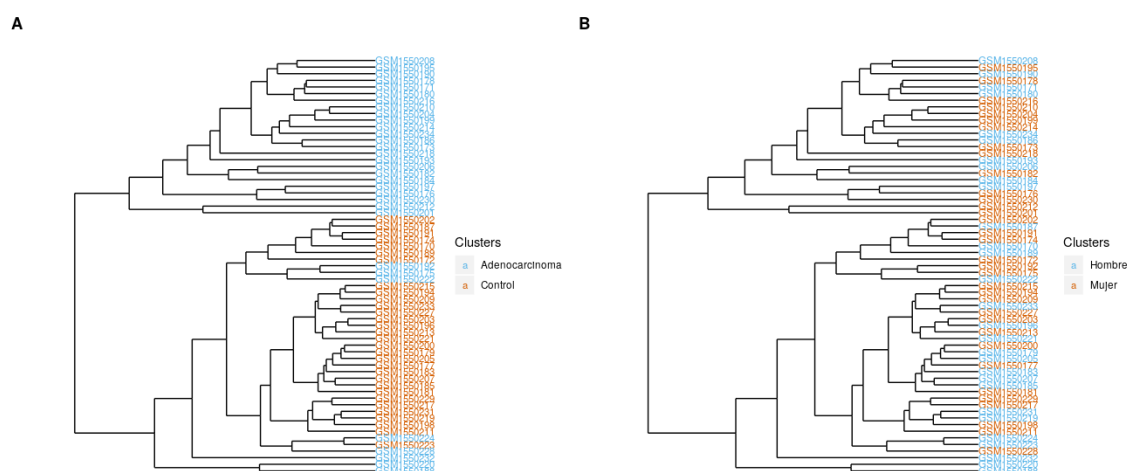


Figura A.5: Clustering exploratorio de las muestras del estudio GSE63459 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

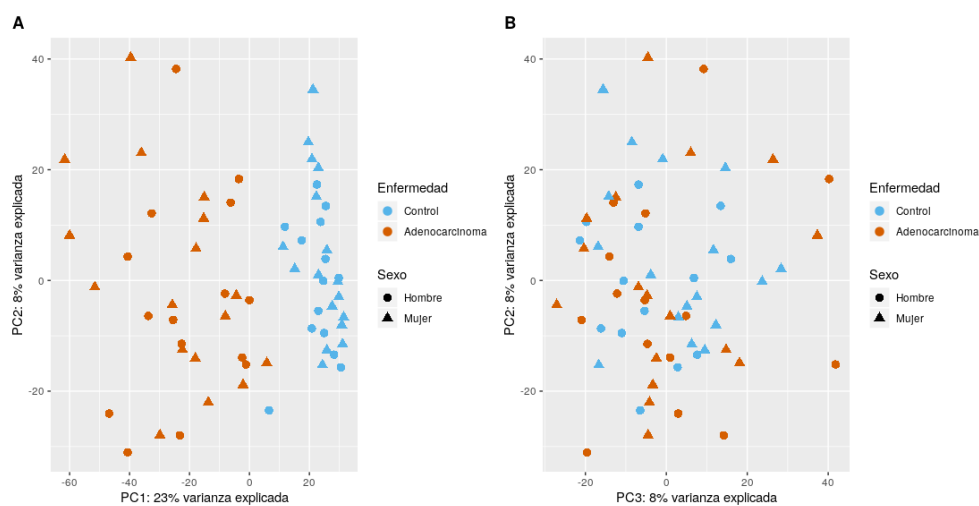


Figura A.6: Análisis de componentes principales de las muestras del estudio GSE63459. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

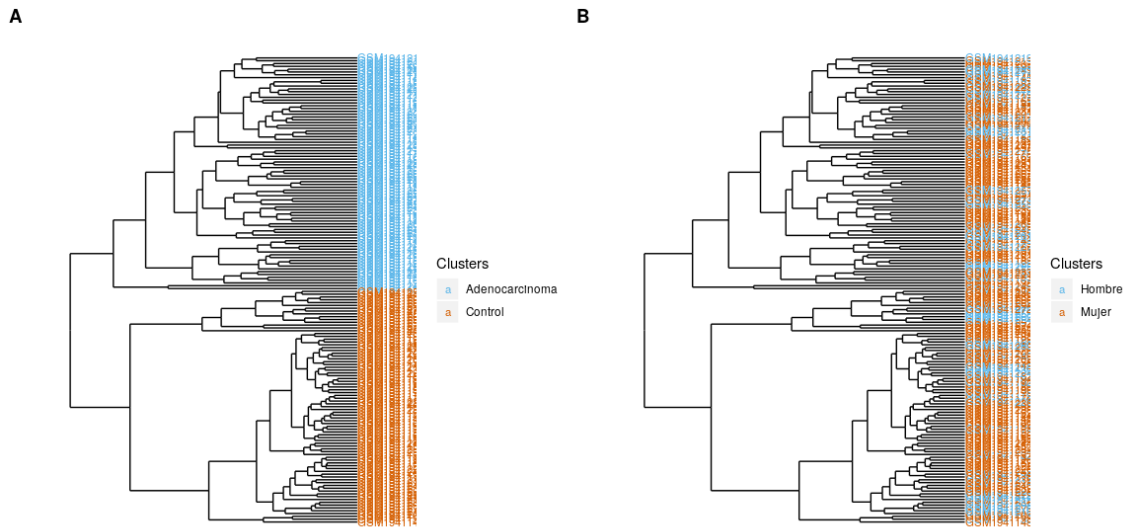


Figura A.7: Clustering exploratorio de las muestras del estudio GSE75037 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

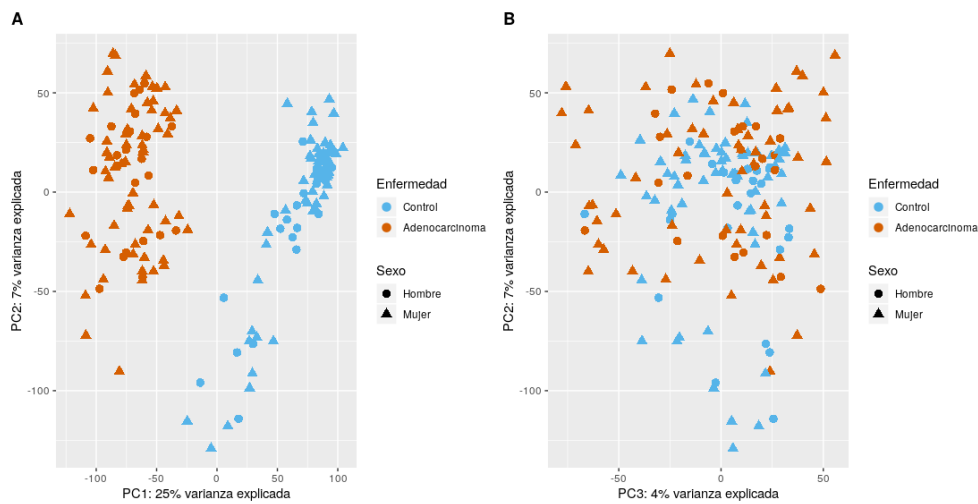


Figura A.8: Análisis de componentes principales de las muestras del estudio GSE75037. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

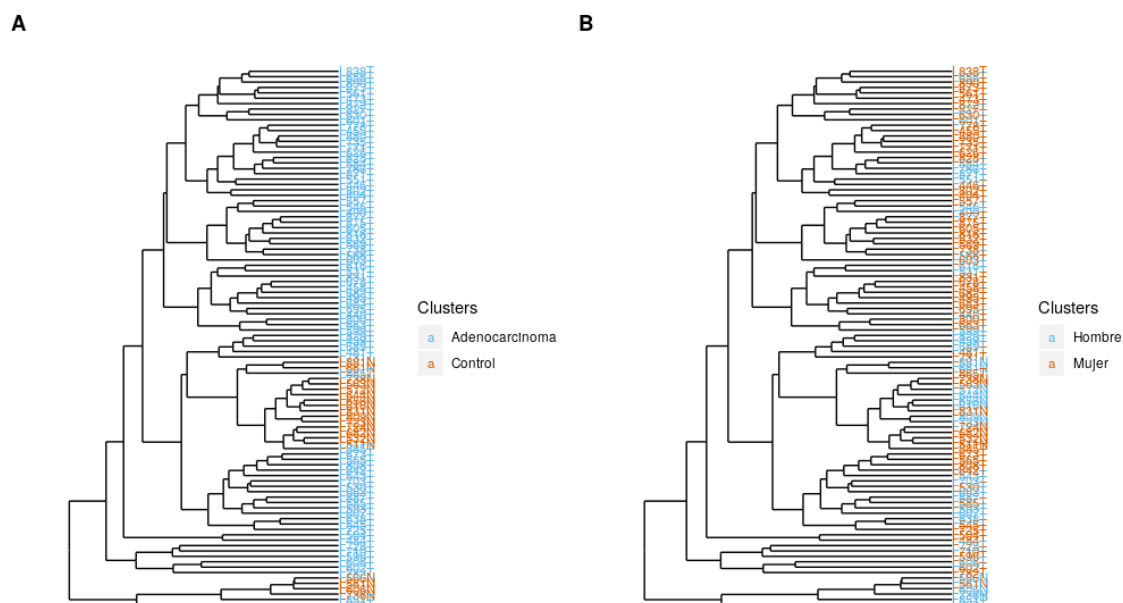


Figura A.9: Clustering exploratorio de las muestras del estudio GSE81089 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

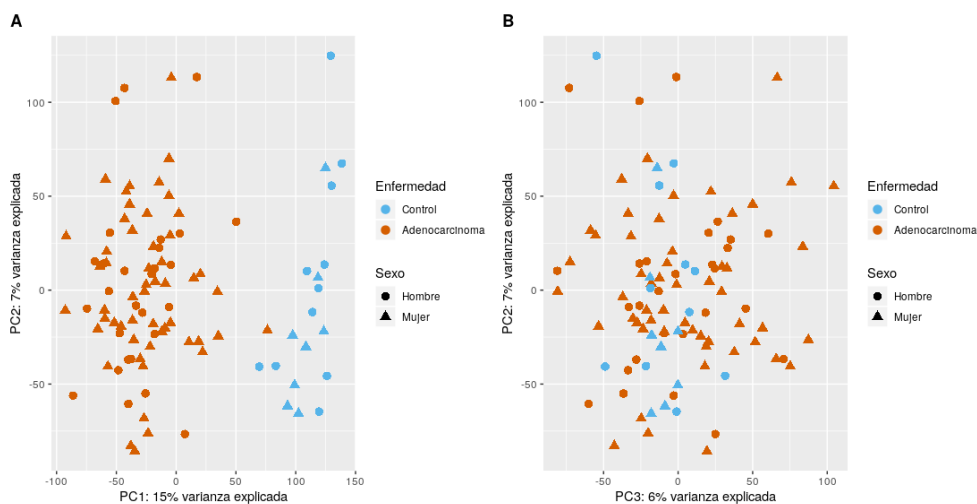


Figura A.10: Análisis de componentes principales de las muestras del estudio GSE81089. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

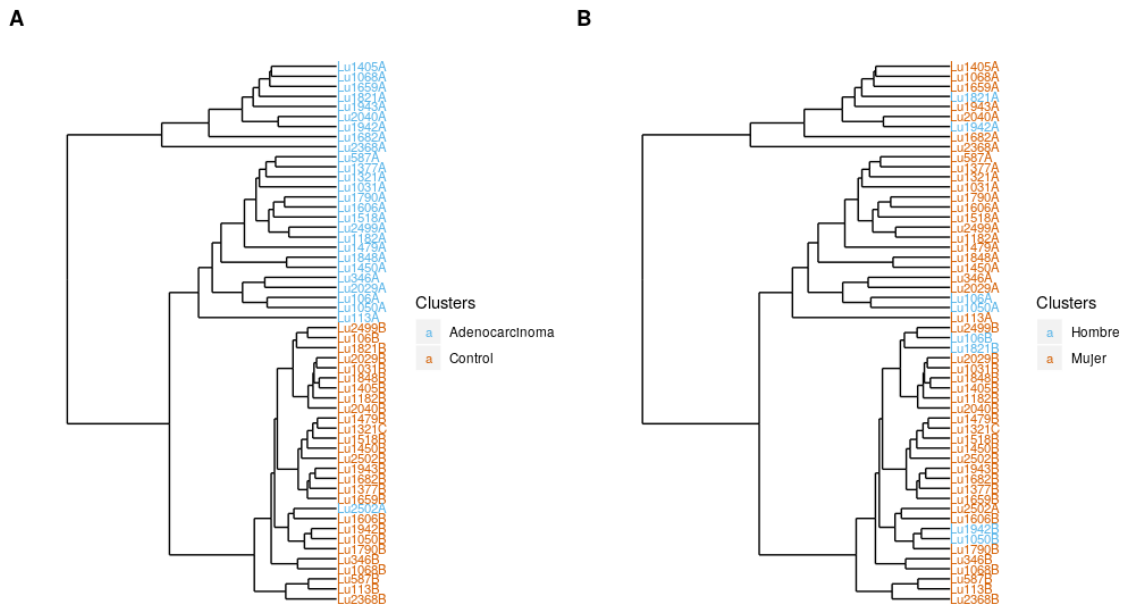


Figura A.11: Clustering exploratorio de las muestras del estudio GSE87340 mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

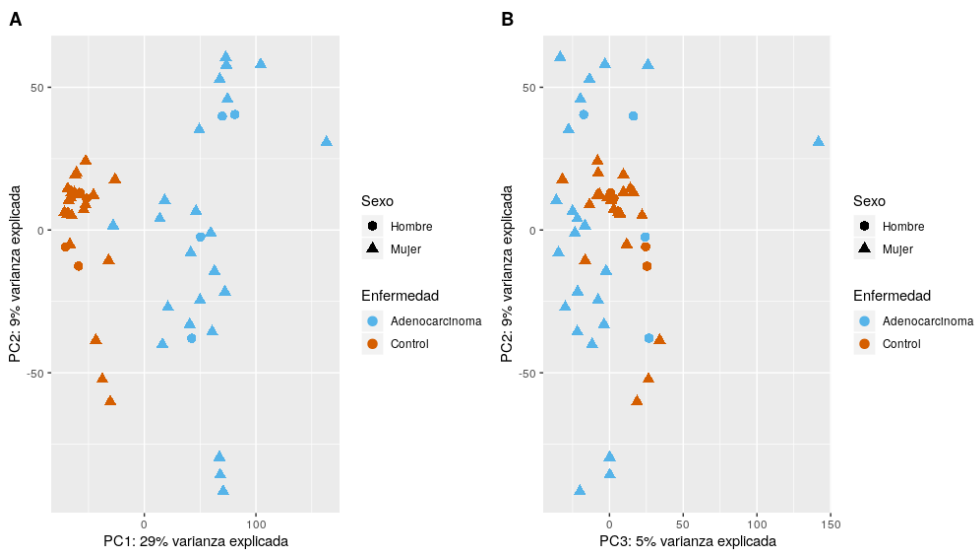


Figura A.12: Análisis de componentes principales de las muestras del estudio GSE87340. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

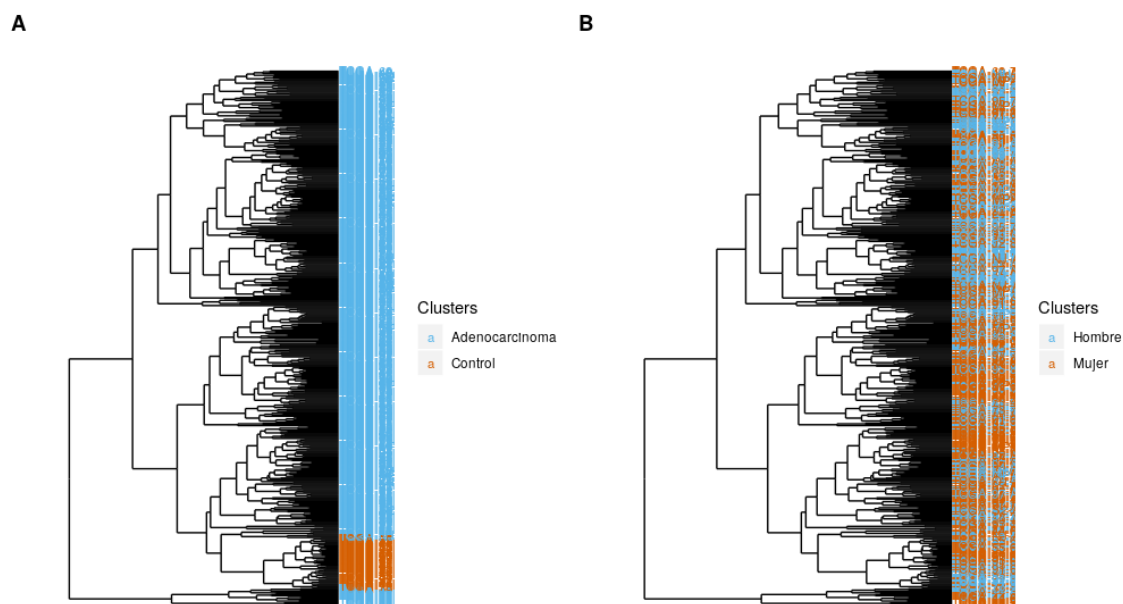


Figura A.13: Clustering exploratorio de las muestras de TCGA-LUAD mediante distancia de correlación. Las muestras se encuentran coloreadas por enfermedad (A) o por el sexo de los individuos (B).

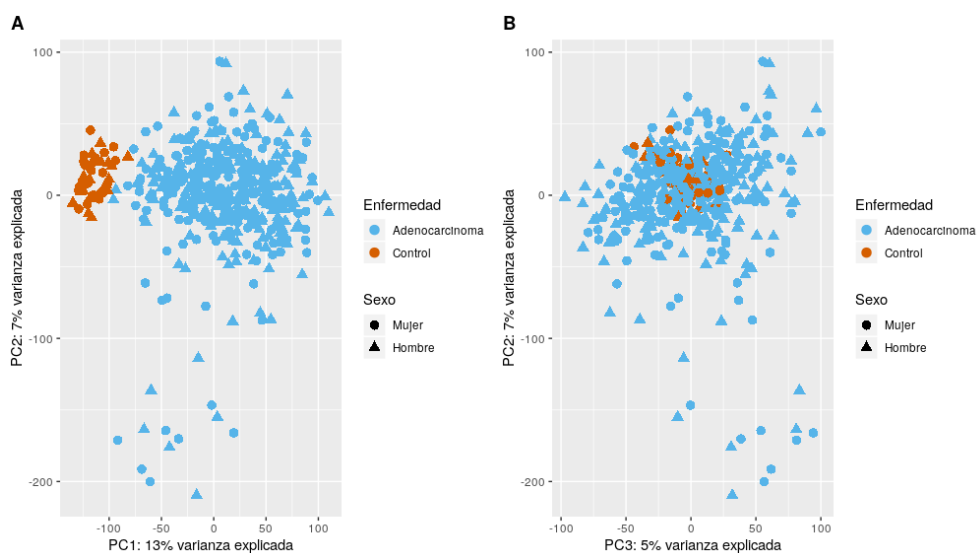


Figura A.14: Análisis de componentes principales de las muestras del TCGA-LUAD. El color indica la enfermedad y la forma geométrica el sexo. Se compara la componente principal 1 frente a la 2 (A) y la 3 frente a la 2 (B).

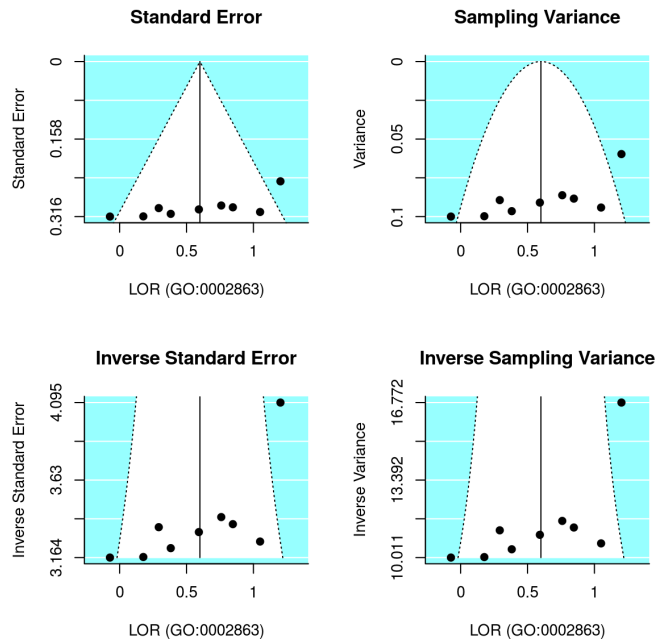


Figura A.15: Gráficos de embudo para el término GO:0002863.

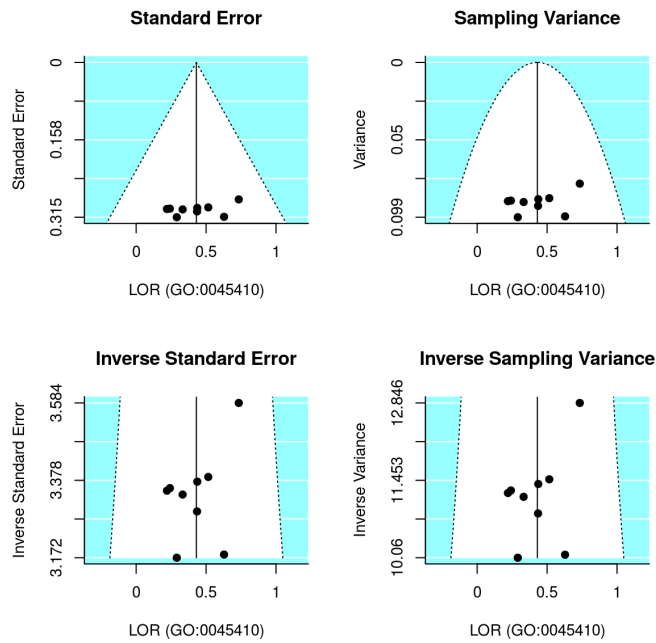


Figura A.16: Gráficos de embudo para el término GO:0045410.

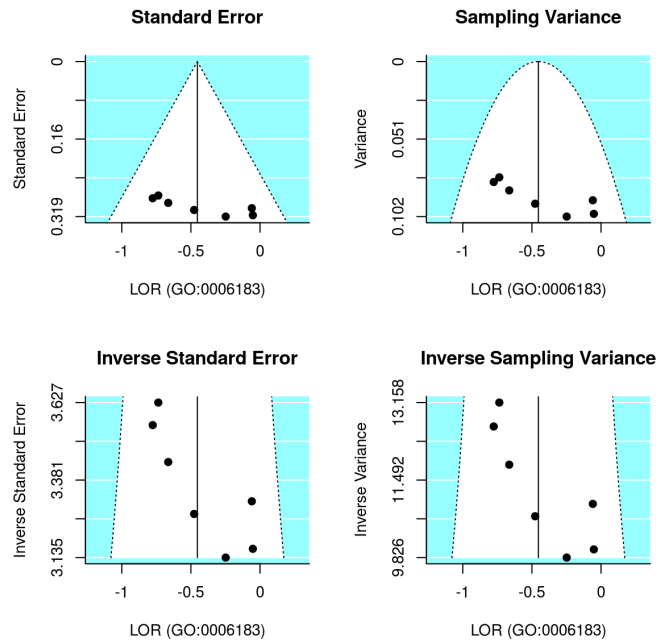


Figura A.17: Gráficos de embudo para el término GO:0006183.

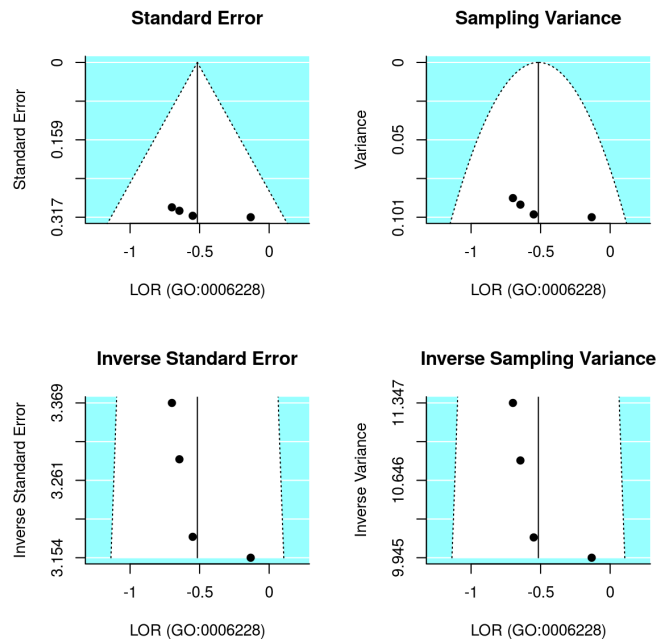


Figura A.18: Gráficos de embudo para el término GO:0006228.

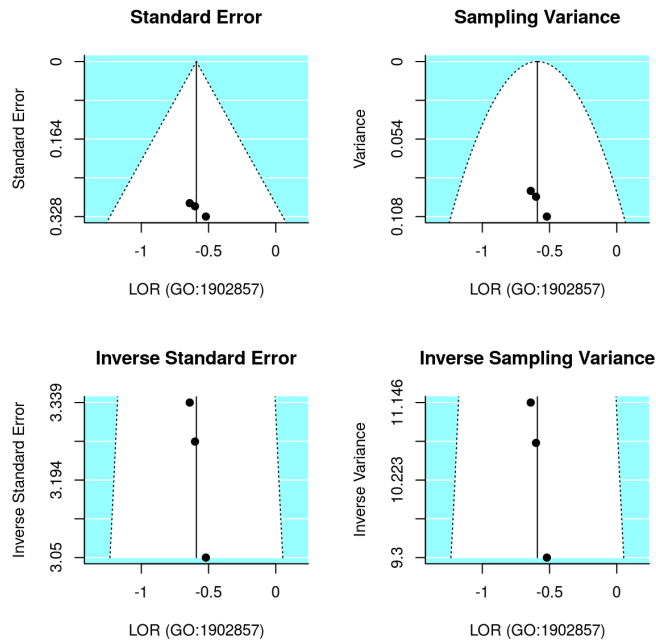


Figura A.19: Gráficos de embudo para el término GO:1902857.

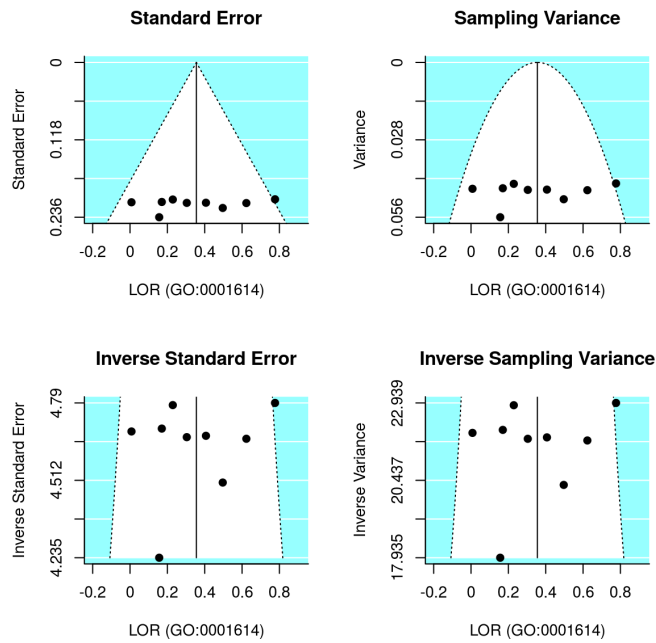


Figura A.20: Gráficos de embudo para el término GO:0001614.

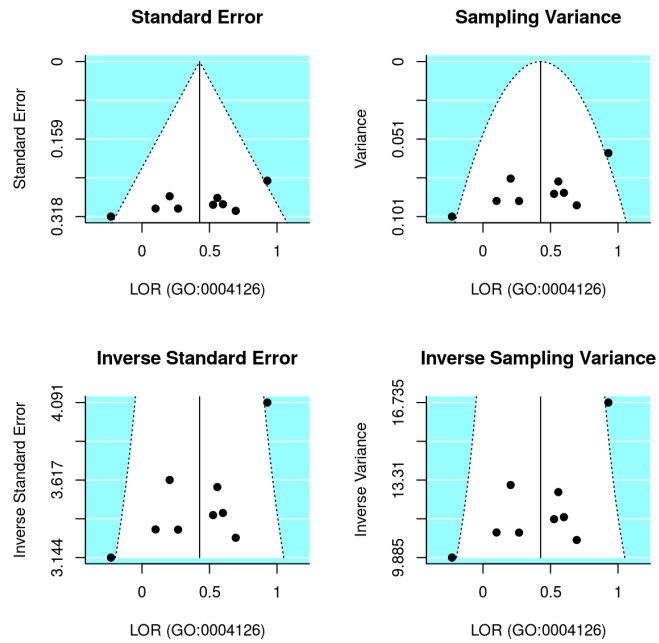


Figura A.21: Gráficos de embudo para el término GO:0004126.

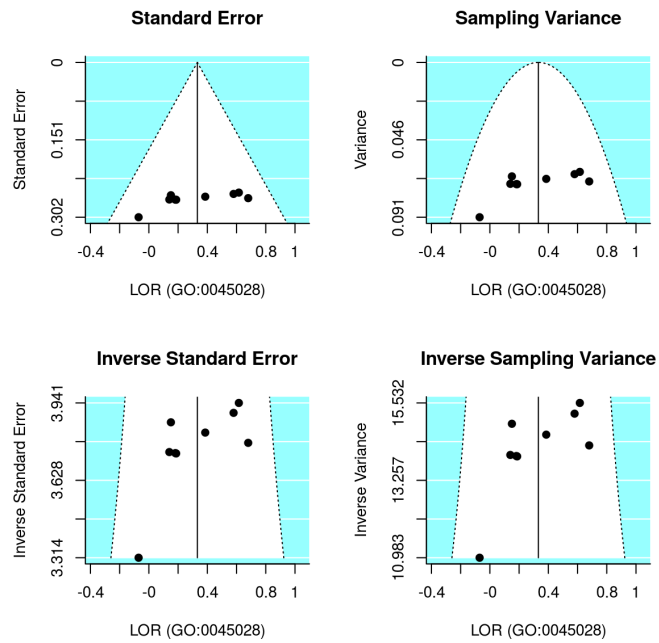


Figura A.22: Gráficos de embudo para el término GO:0045028.

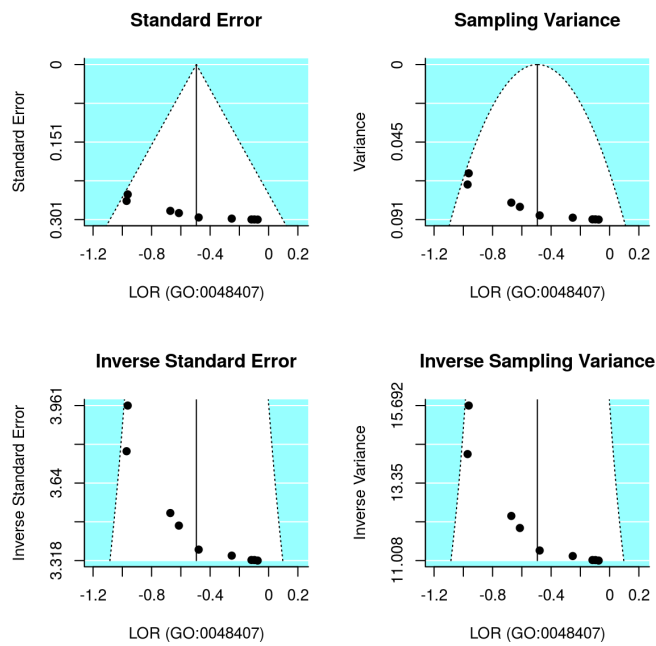


Figura A.23: Gráficos de embudo para el término GO:0048407.

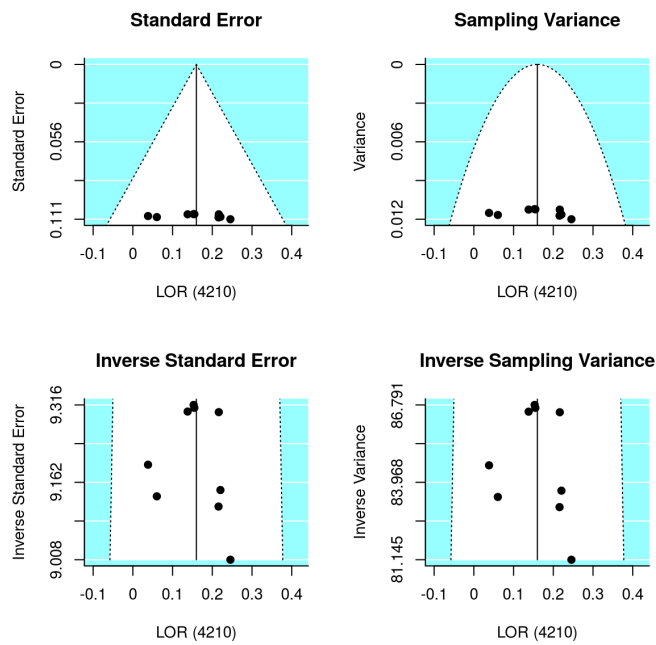


Figura A.24: Gráficos de embudo para la ruta hsa:04210.

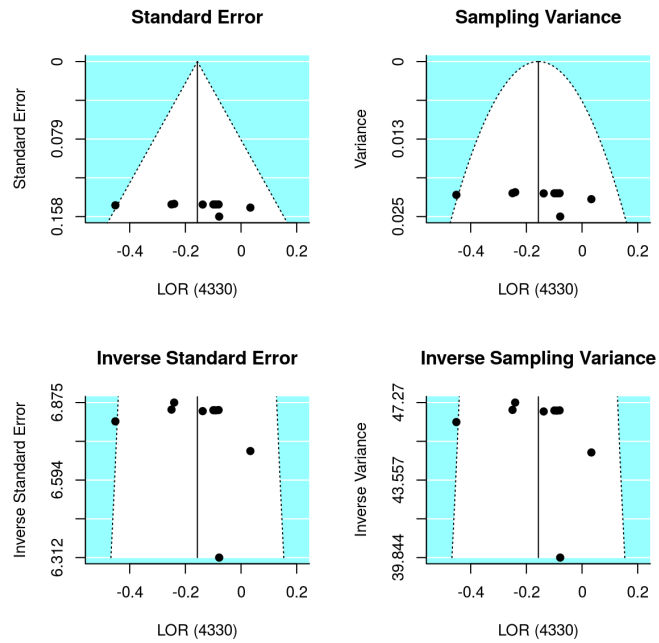


Figura A.25: Gráficos de embudo para la ruta hsa:04330.

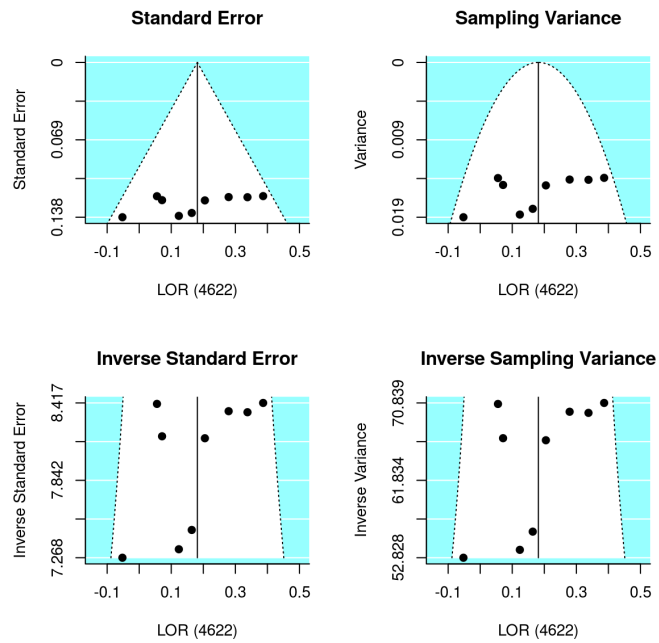


Figura A.26: Gráficos de embudo para la ruta hsa:04622.

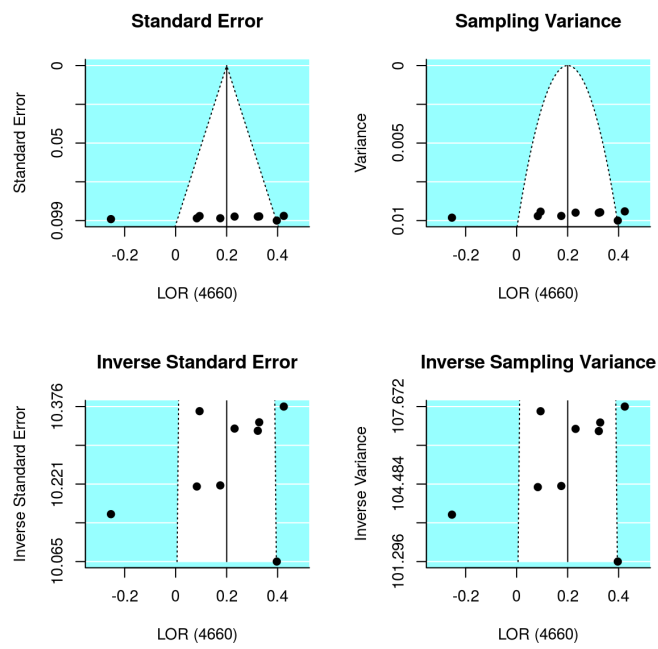


Figura A.27: Gráficos de embudo para la ruta hsa:04660.

B. Anexo B - Tablas

Tabla B.1: Versiones del software empleado en este trabajo.

Software	Versión
R	3.5.3

Paquete	Versión
AnnotationDbi	1.44.0
Biobase	2.42.0
biomaRt	2.38.0
edgeR	3.24.3
GEOquery	2.50.5
ggdendro	0.1-20
ggplot2	3.1.0
ggpubr	0.2
hgu133plus2.db	3.2.3
hipathia	1.4.1
illuminaHumanv3.db	1.26.0
KEGG.db	3.2.3
limma	3.38.3
mdgsa	1.14.0
metafor	2.1-0
methods	3.5.3
org.Hs.eg.db	3.7.0
reshape	0.8.8
stats	3.5.3
SummarizedExperiment	1.12.0
TCGAbiolinks	2.10.5
tidyverse	1.2.1
UpSetR	1.3.3
utils	3.5.3

Tabla B.2: Términos GO BP significativos en el metaanálisis. Se muestra el p-valor ajustado por BH, el LOR y su intervalo de confianza al 95%.

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0000226	<i>microtubule cytoskeleton organization</i>	-0,269	-0,203	-0,137	0
GO:0000491	<i>small nucleolar ribonucleoprotein complex assembly</i>	-0,699	-0,479	-0,259	0,009
GO:0001841	<i>neural tube formation</i>	-0,206	-0,142	-0,078	0,009
GO:0001843	<i>neural tube closure</i>	-0,221	-0,151	-0,082	0,009
GO:0001894	<i>tissue homeostasis</i>	0,05	0,103	0,157	0,024
GO:0002285	<i>lymphocyte activation involved in immune response</i>	0,092	0,226	0,359	0,046
GO:0002438	<i>acute inflammatory response to antigenic stimulus</i>	0,119	0,294	0,469	0,047
GO:0002524	<i>hypersensitivity</i>	0,271	0,637	1,003	0,038
GO:0002691	<i>regulation of cellular extravasation</i>	0,115	0,257	0,399	0,03
GO:0002701	<i>negative regulation of production of molecular mediator of immune response</i>	0,142	0,274	0,406	0,012
GO:0002705	<i>positive regulation of leukocyte mediated immunity</i>	0,092	0,204	0,315	0,029
GO:0002756	<i>MyD88-independent toll-like receptor signaling pathway</i>	0,142	0,317	0,492	0,03
GO:0002821	<i>positive regulation of adaptive immune response</i>	0,114	0,237	0,359	0,023
GO:0002824	<i>positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains</i>	0,112	0,237	0,362	0,027
GO:0002861	<i>regulation of inflammatory response to antigenic stimulus</i>	0,139	0,34	0,54	0,045
GO:0002863	<i>positive regulation of inflammatory response to antigenic stimulus</i>	0,317	0,6	0,883	0,01
GO:0002864	<i>regulation of acute inflammatory response to antigenic stimulus</i>	0,19	0,412	0,633	0,027
GO:0002888	<i>positive regulation of myeloid leukocyte mediated immunity</i>	0,111	0,239	0,368	0,027
GO:0002922	<i>positive regulation of humoral immune response</i>	0,167	0,338	0,51	0,023
GO:0002923	<i>regulation of humoral immune response mediated by circulating immunoglobulin</i>	0,169	0,364	0,559	0,027
GO:0003081	<i>regulation of systemic arterial blood pressure by renin-angiotensin</i>	0,12	0,248	0,376	0,023

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0006122	<i>mitochondrial electron transport, ubiquinol to cytochrome c</i>	-0,622	-0,429	-0,236	0,009
GO:0006183	<i>GTP biosynthetic process</i>	-0,685	-0,453	-0,222	0,023
GO:0006216	<i>cytidine catabolic process</i>	0,196	0,427	0,658	0,028
GO:0006228	<i>UTP biosynthetic process</i>	-0,818	-0,516	-0,215	0,043
GO:0006282	<i>regulation of DNA repair</i>	-0,25	-0,165	-0,081	0,023
GO:0006906	<i>vesicle fusion</i>	0,069	0,157	0,244	0,033
GO:0006957	<i>complement activation, alternative pathway</i>	0,155	0,331	0,507	0,027
GO:0007596	<i>blood coagulation</i>	0,065	0,122	0,18	0,01
GO:0007599	<i>hemostasis</i>	0,066	0,123	0,179	0,009
GO:0009148	<i>pyrimidine nucleoside triphosphate biosynthetic process</i>	-0,422	-0,268	-0,114	0,038
GO:0009972	<i>cytidine deamination</i>	0,196	0,427	0,658	0,028
GO:0010226	<i>response to lithium ion</i>	0,115	0,257	0,4	0,031
GO:0010744	<i>positive regulation of macrophage derived foam cell differentiation</i>	0,112	0,276	0,439	0,046
GO:0014020	<i>primary neural tube formation</i>	-0,208	-0,141	-0,074	0,01
GO:0014866	<i>skeletal myofibril assembly</i>	0,228	0,492	0,756	0,027
GO:0018200	<i>peptidyl-glutamic acid modification</i>	-0,473	-0,308	-0,144	0,027
GO:0018208	<i>peptidyl-proline modification</i>	-0,311	-0,202	-0,092	0,028
GO:0021892	<i>cerebral cortex GABAergic interneuron differentiation</i>	-0,761	-0,485	-0,209	0,036
GO:0030240	<i>skeletal muscle thin filament assembly</i>	0,263	0,549	0,836	0,024
GO:0030890	<i>positive regulation of B cell proliferation</i>	0,162	0,314	0,467	0,013
GO:0031033	<i>myosin filament organization</i>	0,162	0,394	0,626	0,045
GO:0031034	<i>myosin filament assembly</i>	0,199	0,426	0,654	0,027
GO:0032611	<i>interleukin-1 beta production</i>	0,08	0,195	0,309	0,045
GO:0032620	<i>interleukin-17 production</i>	0,126	0,286	0,446	0,033
GO:0032623	<i>interleukin-2 production</i>	0,102	0,252	0,401	0,046
GO:0032660	<i>regulation of interleukin-17 production</i>	0,133	0,304	0,475	0,033
GO:0032663	<i>regulation of interleukin-2 production</i>	0,125	0,261	0,396	0,024
GO:0032673	<i>regulation of interleukin-4 production</i>	0,139	0,337	0,534	0,045
GO:0032729	<i>positive regulation of interferon-gamma production</i>	0,109	0,259	0,408	0,04
GO:0032731	<i>positive regulation of interleukin-1 beta production</i>	0,095	0,214	0,333	0,033
GO:0032733	<i>positive regulation of interleukin-10 production</i>	0,152	0,271	0,391	0,009
GO:0032740	<i>positive regulation of interleukin-17 production</i>	0,166	0,358	0,55	0,027

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0032743	<i>positive regulation of interleukin-2 production</i>	0,176	0,328	0,481	0,009
GO:0032760	<i>positive regulation of tumor necrosis factor production</i>	0,087	0,207	0,328	0,042
GO:0032957	<i>inositol trisphosphate metabolic process</i>	0,228	0,485	0,741	0,027
GO:0032958	<i>inositol phosphate biosynthetic process</i>	0,093	0,222	0,35	0,04
GO:0035587	<i>purinergic receptor signaling pathway</i>	0,154	0,318	0,481	0,023
GO:0035588	<i>G protein-coupled purinergic receptor signaling pathway</i>	0,136	0,309	0,482	0,033
GO:0035589	<i>G protein-coupled purinergic nucleotide receptor signaling pathway</i>	0,158	0,332	0,505	0,025
GO:0035590	<i>purinergic nucleotide receptor signaling pathway</i>	0,182	0,333	0,484	0,009
GO:0036037	<i>CD8-positive, alpha-beta T cell activation</i>	0,185	0,379	0,573	0,023
GO:0036065	<i>fucosylation</i>	0,114	0,263	0,412	0,035
GO:0038061	<i>NIK/NF-kappaB signaling</i>	0,095	0,188	0,281	0,018
GO:0038095	<i>Fc-epsilon receptor signaling pathway</i>	0,169	0,291	0,412	0,005
GO:0042035	<i>regulation of cytokine biosynthetic process</i>	0,095	0,193	0,291	0,023
GO:0042089	<i>cytokine biosynthetic process</i>	0,087	0,18	0,274	0,023
GO:0042100	<i>B cell proliferation</i>	0,145	0,31	0,475	0,027
GO:0042107	<i>cytokine metabolic process</i>	0,09	0,185	0,281	0,023
GO:0042108	<i>positive regulation of cytokine biosynthetic process</i>	0,175	0,283	0,39	0,001
GO:0042226	<i>interleukin-6 biosynthetic process</i>	0,115	0,262	0,409	0,033
GO:0043171	<i>peptide catabolic process</i>	0,126	0,255	0,385	0,023
GO:0043302	<i>positive regulation of leukocyte degranulation</i>	0,115	0,254	0,392	0,028
GO:0043374	<i>CD8-positive, alpha-beta T cell differentiation</i>	0,241	0,479	0,716	0,018
GO:0043576	<i>regulation of respiratory gaseous exchange</i>	-0,434	-0,291	-0,148	0,016
GO:0044070	<i>regulation of anion transport</i>	0,062	0,152	0,242	0,046
GO:0044458	<i>motile cilium assembly</i>	-0,802	-0,513	-0,225	0,033
GO:0045006	<i>DNA deamination</i>	0,174	0,385	0,596	0,029
GO:0045076	<i>regulation of interleukin-2 biosynthetic process</i>	0,184	0,448	0,711	0,045
GO:0045086	<i>positive regulation of interleukin-2 biosynthetic process</i>	0,22	0,51	0,8	0,036
GO:0045332	<i>phospholipid translocation</i>	0,112	0,259	0,406	0,035
GO:0045408	<i>regulation of interleukin-6 biosynthetic process</i>	0,121	0,276	0,432	0,033

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0045410	<i>positive regulation of interleukin-6 biosynthetic process</i>	0,235	0,43	0,626	0,009
GO:0045414	<i>regulation of interleukin-8 biosynthetic process</i>	0,139	0,328	0,517	0,039
GO:0045739	<i>positive regulation of DNA repair</i>	-0,234	-0,15	-0,067	0,032
GO:0045807	<i>positive regulation of endocytosis</i>	0,059	0,126	0,193	0,027
GO:0045833	<i>negative regulation of lipid metabolic process</i>	0,071	0,166	0,261	0,038
GO:0046087	<i>cytidine metabolic process</i>	0,196	0,427	0,658	0,028
GO:0046635	<i>positive regulation of alpha-beta T cell activation</i>	0,143	0,316	0,489	0,029
GO:0046638	<i>positive regulation of alpha-beta T cell differentiation</i>	0,122	0,296	0,47	0,045
GO:0048305	<i>immunoglobulin secretion</i>	0,143	0,313	0,482	0,028
GO:0048873	<i>homeostasis of number of cells within a tissue</i>	0,085	0,204	0,323	0,043
GO:0050716	<i>positive regulation of interleukin-1 secretion</i>	0,123	0,272	0,42	0,029
GO:0050718	<i>positive regulation of interleukin-1 beta secretion</i>	0,143	0,306	0,469	0,027
GO:0050729	<i>positive regulation of inflammatory response</i>	0,08	0,198	0,316	0,047
GO:0050817	<i>coagulation</i>	0,065	0,121	0,177	0,009
GO:0050848	<i>regulation of calcium-mediated signaling</i>	0,059	0,129	0,199	0,028
GO:0050853	<i>B cell receptor signaling pathway</i>	0,193	0,429	0,665	0,03
GO:0050854	<i>regulation of antigen receptor-mediated signaling pathway</i>	0,173	0,368	0,563	0,027
GO:0050855	<i>regulation of B cell receptor signaling pathway</i>	0,265	0,484	0,703	0,009
GO:0050858	<i>negative regulation of antigen receptor-mediated signaling pathway</i>	0,152	0,367	0,583	0,044
GO:0050878	<i>regulation of body fluid levels</i>	0,045	0,098	0,15	0,027
GO:0050995	<i>negative regulation of lipid catabolic process</i>	0,106	0,254	0,402	0,042
GO:0051023	<i>regulation of immunoglobulin secretion</i>	0,122	0,285	0,448	0,037
GO:0051048	<i>negative regulation of secretion</i>	0,046	0,108	0,171	0,038
GO:0051957	<i>positive regulation of amino acid transport</i>	0,105	0,256	0,407	0,045
GO:0060513	<i>prostatic bud formation</i>	-0,625	-0,407	-0,19	0,027
GO:0060563	<i>neuroepithelial cell differentiation</i>	-0,28	-0,177	-0,075	0,04
GO:0060606	<i>tube closure</i>	-0,225	-0,156	-0,087	0,009
GO:0061756	<i>leukocyte adhesion to vascular endothelial cell</i>	0,101	0,214	0,328	0,027

ID	Nombre	LI 95%	LOR	LS 95%	FDR
GO:0070199	<i>establishment of protein localization to chromosome</i>	-0,477	-0,315	-0,152	0,023
GO:0072331	<i>signal transduction by p53 class mediator</i>	-0,136	-0,087	-0,039	0,03
GO:0072593	<i>reactive oxygen species metabolic process</i>	0,033	0,077	0,12	0,035
GO:0072606	<i>interleukin-8 secretion</i>	0,096	0,235	0,374	0,046
GO:0072676	<i>lymphocyte migration</i>	0,097	0,239	0,381	0,046
GO:0090174	<i>organelle membrane fusion</i>	0,059	0,14	0,22	0,038
GO:0090344	<i>negative regulation of cell aging</i>	-0,374	-0,239	-0,105	0,033
GO:0097090	<i>presynaptic membrane organization</i>	-0,542	-0,346	-0,15	0,035
GO:0097104	<i>postsynaptic membrane assembly</i>	-0,705	-0,482	-0,259	0,009
GO:0097105	<i>presynaptic membrane assembly</i>	-0,799	-0,516	-0,233	0,029
GO:0099054	<i>presynapse assembly</i>	-0,487	-0,309	-0,131	0,04
GO:0099172	<i>presynapse organization</i>	-0,459	-0,304	-0,148	0,023
GO:0140029	<i>exocytic process</i>	0,079	0,158	0,237	0,02
GO:1901070	<i>guanosine-containing compound biosynthetic process</i>	-0,493	-0,315	-0,137	0,035
GO:1901222	<i>regulation of NIK/NF-kappaB signaling</i>	0,087	0,19	0,292	0,028
GO:1901989	<i>positive regulation of cell cycle phase transition</i>	-0,227	-0,143	-0,059	0,044
GO:1902750	<i>negative regulation of cell cycle G2/M phase transition</i>	-0,239	-0,151	-0,062	0,045
GO:1902857	<i>positive regulation of non-motile cilium assembly</i>	-0,942	-0,591	-0,239	0,046
GO:1903975	<i>regulation of glial cell migration</i>	0,164	0,374	0,584	0,033
GO:1904666	<i>regulation of ubiquitin protein ligase activity</i>	-0,513	-0,364	-0,214	0,004
GO:1904668	<i>positive regulation of ubiquitin protein ligase activity</i>	-0,538	-0,347	-0,157	0,029
GO:1904994	<i>regulation of leukocyte adhesion to vascular endothelial cell</i>	0,175	0,33	0,486	0,01
GO:1904996	<i>positive regulation of leukocyte adhesion to vascular endothelial cell</i>	0,216	0,407	0,597	0,01
GO:2000341	<i>regulation of chemokine (C-X-C motif) ligand 2 production</i>	0,176	0,428	0,681	0,045
GO:2000377	<i>regulation of reactive oxygen species metabolic process</i>	0,037	0,093	0,148	0,048
GO:2000401	<i>regulation of lymphocyte migration</i>	0,13	0,25	0,37	0,012
GO:2000482	<i>regulation of interleukin-8 secretion</i>	0,14	0,303	0,466	0,027
GO:2001185	<i>regulation of CD8-positive, alpha-beta T cell activation</i>	0,129	0,316	0,503	0,045

Tabla B.3: Rutas KEGG significativas en el metaanálisis. Se muestra el p-valor ajustado por BH, el LOR y su intervalo de confianza al 95%.

ID	Nombre	LI 95%	LOR	LS 95%	FDR
380	<i>Tryptophan metabolism</i>	0,049	0,152	0,255	0,039
3440	<i>Homologous recombination</i>	-0,373	-0,229	-0,085	0,03
4060	<i>Cytokine-cytokine receptor interaction</i>	0,052	0,158	0,264	0,039
4620	<i>Toll-like receptor signaling pathway</i>	0,092	0,212	0,331	0,014
4623	<i>Cytosolic DNA-sensing pathway</i>	0,114	0,255	0,396	0,012
4640	<i>Hematopoietic cell lineage</i>	0,156	0,337	0,518	0,009
4650	<i>Natural killer cell mediated cytotoxicity</i>	0,063	0,183	0,302	0,036
4662	<i>B cell receptor signaling pathway</i>	0,093	0,226	0,359	0,018
4664	<i>Fc epsilon RI signaling pathway</i>	0,079	0,162	0,246	0,006
4930	<i>Type II diabetes mellitus</i>	0,066	0,162	0,258	0,018
4972	<i>Pancreatic secretion</i>	0,089	0,168	0,246	0,003
4973	<i>Carbohydrate digestion and absorption</i>	0,106	0,259	0,413	0,018
4976	<i>Bile secretion</i>	0,045	0,125	0,204	0,03
5143	<i>African trypanosomiasis</i>	0,087	0,255	0,423	0,037
5310	<i>Asthma</i>	0,138	0,462	0,786	0,048