VNIVERSIDAD DE SALAMANCA

FACVLTAD DE FARMACIA

INSTITVTO DE INVESTIGACIÓN BIOMÉDICA DE SALAMANCA

# TESIS DOCTORAL

## Expresión génica en mieloma múltiple: análisis de datos de RNA-seq y microarrays en combinación con estudios de metaanálisis y predicción de respuesta al tratamiento.

Memoria para optar al Grado de Doctor por el Programa de Doctorado en Farmacia y Salud, adscrito a la Facultad de Farmacia de la Universidad de Salamanca, presentada por

**Luis Antonio Corchete Sánchez**

Bajo la dirección de los doctores

**Francisco Javier Burguillo Muñoz**

**Norma Carmen Gutiérrez Gutiérrez**

2019

VNIVERSIDAD DE SALAMANCA

FACVLTAD DE FARMACIA

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL


**FRANCISCO JAVIER BURGUILLO MUÑOZ**

*Profesor Titular del Departamento de Química Física*

*Universidad de Salamanca*

**NORMA CARMEN GUTIÉRREZ GUTIÉRREZ**

*Profesora Asociada del Departamento de Medicina*

*Médico Adjunto del Servicio de Hematología*

*Universidad de Salamanca*


## CERTIFICAN:

Que el trabajo realizado por D. Luis Antonio Corchete Sánchez, Licenciado en Biología, realizado bajo su dirección en del Departamento de Química Física y titulado "***Expresión génica en mieloma múltiple: análisis de datos de RNA-seq y microarrays en combinación con estudios de metaanálisis y predicción de respuesta al tratamiento***", reúne las condiciones de originalidad y calidad científica requeridas para optar al Grado de Doctor.

Y para que así conste, firman el presente certificado en Salamanca, en junio de 2019

Fdo. Fco. Javier Burguillo Muñoz

Fdo. Norma C. Gutiérrez Gutiérrez

*A mis padres.*
*A mi sobrino Daniel.*

*No hay que temer a nada en la vida,*
*solo tratar de comprender.*

***Marie Sklodowska-Curie***

# Agradecimientos

***Agradecimientos***

No querría perder la ocasión que me brinda la escritura de este trabajo sin agradecer todo el esfuerzo y ayuda de quienes han hecho posible que finalmente se haya materializado.

En primer lugar, querría agradecer al Doctor Javier Burguillo por su entrega y dedicación y por haberme acogido en su Departamento bajo su tutela, pero por encima de todo, por haber sido un gran amigo que ha estado siempre dispuesto a ayudar en lo que fuese necesario y por despertarme el interés por la ciencia.

A la Doctora Norma Gutiérrez también querría agradecerle toda la ayuda que me ha brindado en estos cinco años, el haberme animado a iniciar este camino del Doctorado, el haberme dado la oportunidad de conocer la Hematología desde su grupo, y por haber sacado tiempo para echarme una mano a pesar de no disponer del mismo.

Por supuesto, quiero agradecer el apoyo que me ha dado todo el "Grupo Norma", Eli, Dalia, Patryk e Irena y los miembros más recientes, Cristina, Ignacio y María. Muchas gracias por vuestra ayuda y por darme vuestro tiempo cuando os lo he pedido. Tampoco podía olvidarme de Ana Belén, "*former member*", por toda la brasa que le di, primero en el Banco de Sangre y después en el Departamental, y porque siempre estás dispuesta a echar una mano.

Muchísimas gracias también a todo el laboratorio 12, al Departamento de Química-Física de la Facultad de Farmacia y al Servicio de Hematología, por vuestra confianza, ayuda y preocupación. Eva, al final sobreviví a la escritura jeje.

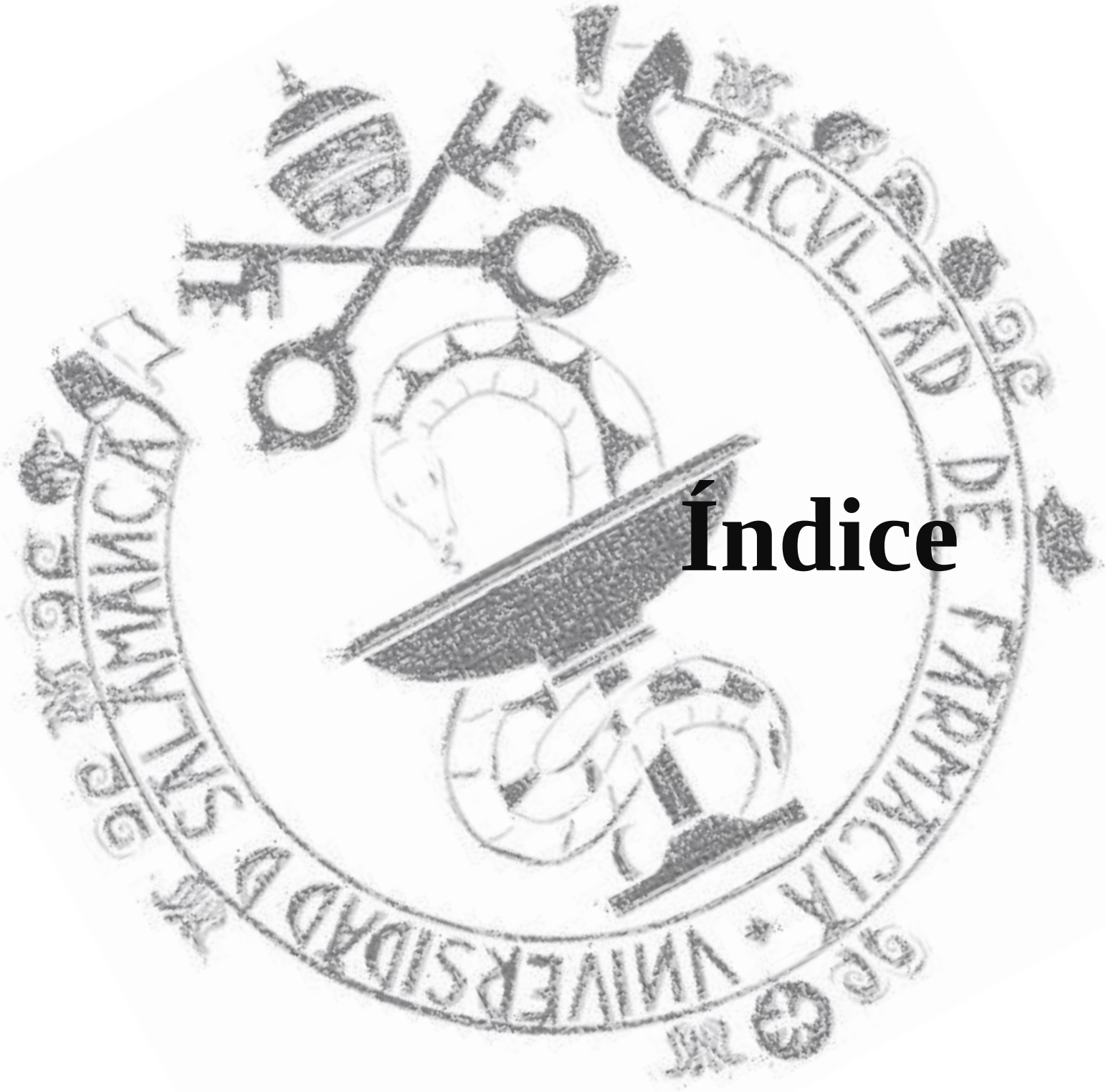
Gracias también al Doctor Bardsley por invitarme a compartir unos días con él que fueron una gran experiencia personal, por su ayuda y por su dedicación en el campo de la estadística.

Gracias a Diego y Javier de las Rivas, por su ayuda desinteresada y por poner su grano de arena en este trabajo.

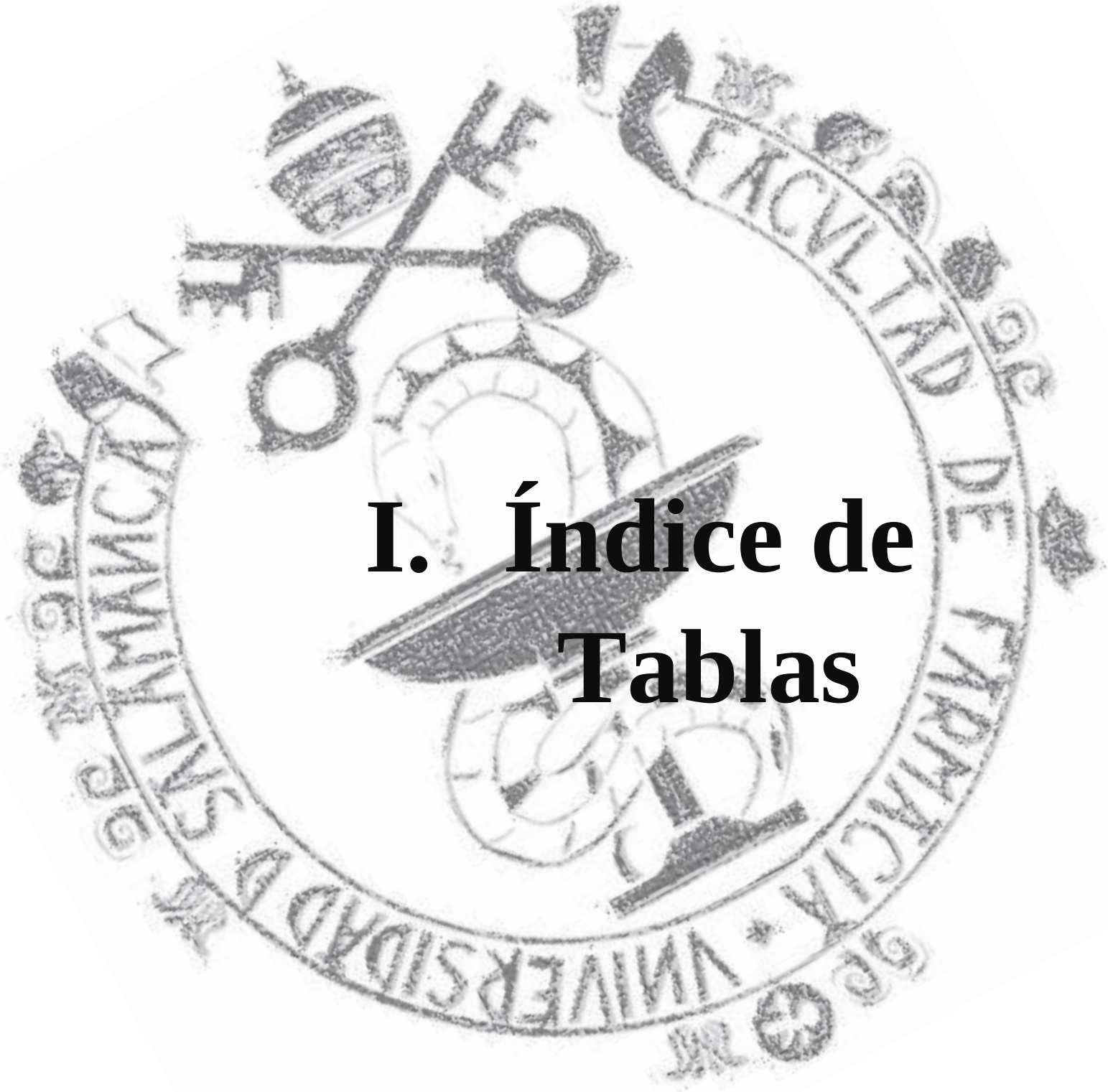
Gracias a la Doctora María Ángeles Castro, coordinadora del programa de doctorado, por su buena disposición y amabilidad.

Gracias a Manu, Rosario, Marta y Laura porque siempre han estado a mi lado sin desfallecer, a pesar de que últimamente no he podido dedicarles todo el tiempo que me hubiese gustado, ¡os debo mucho tiempo! Gracias a Jose, Joni, Lorena, Bea y Rebeca, por acompañarme toda la vida y por seguir acompañándome en este camino. *Muito obrigado* a Daniel, gracias a Kike, gracias a Miguel, gracias a Dani, *por seu conselho, por me ensinar que todo esforço tem a sua recompensa* y por ayudarme a sobrevivir durante todos estos años. En fin, gracias a todos mis amig@s y a toda la gente con la que he podido compartir parte de este periodo, por haberme ayudado en esta etapa de mi vida.

Y como no, gracias a lo más grande que tengo, mi familia. Gracias a mis padres por haberlo dado todo y más por mí, porque siempre me han apoyado y me han dado su amor y comprensión y por animarme a seguir adelante ¡os quiero! Gracias a mi hermana y a Clemen, por estar siempre de mi parte y ser una parte de mí. Gracias a mi abuela Lucía y a mis abuelos que ya no están, Segundo, Toño y Consola, por ser el ejemplo que quiero y debo seguir. Gracias a mi tía Pepi y a mis tíos Nacho, Tomás y Roberto, por acompañarme en este viaje y por todo vuestro cariño. Gracias a las primas de Oro (Sheila, Leticia y Pamela) y a mi primo Christian, por vuestra preocupación y apoyo incondicional. Gracias a Lucas y Valeria, la alegría de la casa. Gracias también a mis primos políticos, Mario, Óscar e Iñaki. ¡Muchas gracias a todos!




# Índice



# I. Índice de Tablas

## *Índice de Tablas*

# II. Índice de Figuras

## *Índice de Figuras*

## *Índice de Figuras*

## *Índice de Figuras*

# III. Índice de Anexos

***Los contenidos de los Anexos pueden consultarse en el CD adjunto al presente trabajo u online en el siguiente enlace:***

https://www.dropbox.com/sh/0fcenlc786wwca7/AADXXHuAmijYrhEEhrOUdeNVa?dl=0

*Índice de Anexos*

# IV. Abreviaturas

**AEC** *Expression Console* de Affymetrix

**ARE** Elementos ricos en adenilato-uridilato

**ARNc** ARN complementario

**ARNr** ARN ribosómico

**ASCT** Trasplante autólogo de células madre hematopoyéticas

**AUC** Área bajo la curva

**BA** BrainArray

**BAM** *Binary alignment map*

**BWT** Transformación de Burrows-Wheeler

**CC** Componente celular

**CCECC** Carcinoma de células escamosas de cabeza y cuello

**CCM** Coeficiente de correlación de Matthews

**COR** Característica operativa del receptor

**CoV** Coeficiente de variación no paramétrico

**CPM** *Count-per-million*

**cRC** Respuesta cercana a la respuesta completa

**DMSO** Dimetilsulfóxido

**DNMT** enzima ADN-Metil-Transferasa

**dNTPs** Desoxinucleótidos trifosfato

**DSMZ** *Deutsche Sammlung von Mikroorganismen and Zellkulturen GmbH*

**ED** Expresión diferencial

**EE** Enfermedad estable

**EM** Esperanza-Maximización

**EP** Enfermedad progresiva

**FC** *Fold change*

**FDR** False discovery rate

**FM** Funciones moleculares

**FN** Falsos negativos

**F**P Falsos positivos

**FPKM** *Fragments per Kilobase of transcript per Million mapped reads*

***Abreviaturas***

**FVN** Fracción de verdaderos negativos (especificidad)

**FVP** Fracción de verdaderos positivos (sensibilidad)

**GC** Guanina-Citosina

**GDE** Genes diferencialmente expresados

**GEO** *Gene Expression Omnibus*

**gl** Grados de libertad

**GLM** Modelos lineales generalizados

**GMSI** Gammapatia monoclonal de significado incierto

**GO** *Gene Ontology*

**GRE** Elementos de respuesta a glucocorticoides

**GWAS** Estudios de asociación de genoma completo (*Genome-wide association analysis*)

**HDACi** Inhibidores de las deacetilasas de histonas

**HGNC** *HUGO Gene Nomenclature Comitee*

**HKg** Gen Housekeeping o constitutivo

**HMCLs** Líneas celulares de mieloma múltiple

**HSP** Proteína de choque térmico

**HTA2.0** *Human Transcriptome Array 2.0*

**IFE** Inmunofijación

**IFN-γ** Interferón gamma

**Ig** Inmunoglobulina

**IGH** Cadena pesada de las inmunoglobulinas

**IL-2** Interleucina 2

**IMiDs** Agentes inmunomoduladores

**ISRE** Elemento de respuesta estimulados por interferón

**KNN** K vecinos más cercanos

**LCA** Línea celular A (KMS12-BM)

**LCB** Línea celular B (JJN-3)

**LCP** Leucemia de células plasmáticas

**LLA** Leucemia linfoblástica aguda

**LMA** Leucemia mieloide aguda

**LMC** Leucemia mieloide cronica

**LOOCV** *Leaver-One-Out Cross-Validation*

**MAD** Desviación absoluta de la mediana

**MBRP** Muy buena respuesta parcial

**MDS** Escalado multidimensional o *multidimensional scaling*

**MM** Mieloma múltiple

**MMRF** *Multiple Myeloma Research Foundation*

**NF** Número de factores

**NGS** Secuenciación masiva de nueva generación

**NK** *Natural killer*

**NR** No respondedores

**Ns** Nucleótidos no asignados

**OR** Pacientes respondedores

**ORA** Análisis de sobrerrepresentación

**oRC** Otras respuestas completas no inmunofenotípicas

**pb** Pares de bases

**PB** Procesos biológicos

**PCR** Reacción en cadena de la polimerasa

**PLS** Mínimos cuadrados parciales (*Partial Least Squares*)

**QMB** Modo basado en quasi-mapeo

**qRT-PCR** Reacción en cadena de la polimerasa cuantitativa con transcriptasa inversa

***r*** Coeficiente de correlación de Pearson

**RAE** Real academia de la lengua española

**RC** Respuesta completa

**RC-IF** Respuesta completa inmunofenotípica

**RCs** Respuesta completa estricta

**RF** Bosques aleatorios o *random forests*

**RLE** Expresión relativa en escala logarítmica

**RM** Respuesta mínima

**RP** Respuesta parcial

***Abreviaturas***

**RPKM** *Reads per kilobase million*

**SAM** *Sequence alignmnent map*

**SD** Desviación estándar

**SE** Error estándar

**SG** Supervivencia global

**shRNA** ARN cortos de interferencia

**SLP** Supervivencia libre de progresión

**SMM** Mieloma múltiple indolente (*smoldering multiple myeloma*)

**SQS** Enzima escualeno sintasa

**SRA** *Sequence Read Archive*

**SVM** Máquinas de soporte vectorial

**T0** Tratamiento 0 (control)

**T1** Tratamiento 1 (amilorida)

**T2** Tratamiento 2 (TG003)

**TAC** Tasa global de acierto

**TMM** Media recortada de M-valores

**TNFα** Factor de necrosis tumoral alfa

**TPM** *Transcripts per Million*

**UQ** Cuartil superior

**VAD** Vincristina / doxorrubicina (adriamicina) / dexametasona

**VIM** Importancia de la variable

**VIP** *Variance Influence in Projection*

**VN** Verdaderos negativos

**VP** Verdaderos positivos

**VPN** Valor predictor negativo

**VPP** Valor predictor positivo (precisión)

**VTD** Bortezomib / talidomida / dexametasona

**wSVM** Máquinas de soporte vectorial con pesos estadísticos

# 1. Introducción

# 1.1. Clínica y tratamiento del mieloma múltiple

## 1.1.1. Patología y prevalencia

El mieloma múltiple (MM) es una neoplasia hematológica caracterizada por la acumulación incontrolada de células plasmáticas malignas en la médula ósea, que generalmente producen una inmunoglobulina monoclonal detectable en el suero y/u orina[1]. Las manifestaciones clínicas más comunes del MM derivan del daño orgánico producido por la expansión de las células plasmáticas y por el exceso de inmunoglobulina clonal. Generalmente se agrupan bajo el acrónimo CRAB: hipercalcemia (C = *calcium*), fallo renal (R = *renal failure*), anemia (A = *anemia*) y lesiones óseas (B = *bone lesions*).

El MM representa aproximadamente el 10% de los cánceres hematológicos[2] y es el segundo más relevante tanto en número de nuevos casos como en muertes anuales tras el linfoma no Hodgkin[3]. En los países occidentales, la incidencia anual del MM es de 6,9 nuevos casos por cada 100.000 habitantes[4]. Concretamente en España, se diagnostican aproximadamente 2.000 nuevos casos cada año. En cuanto a la distribución por sexos, presenta una incidencia ligeramente superior en hombres que en mujeres, y se han observado también diferencias en la prevalencia de determinados eventos genéticos en función del sexo del paciente[5]. Respecto a la incidencia racial, el MM es dos veces más común en afroamericanos que en caucásicos[6]. La edad mediana de diagnóstico del MM es de 70 años, aunque en torno al 35% de los casos se diagnostican por debajo de los 65 años[1, 4, 7]. En pacientes que presentan una edad inferior a 60 años, la supervivencia a 10 años es aproximadamente del 30%[8].

## 1.1.2. Historia

La historia del MM no se puede entender sin conocer la historia de la célula plasmática. El descubrimiento de esta célula se remonta a finales del siglo XIX, cuando Santiago Ramón y Cajal logró dibujarlas con precisión mientras realizaba una investigación sobre lesiones sifilíticas. Las denominó "*células cianófilas"* debido a su afinidad por los colorantes azules o verdosos. Sin embargo, no sería hasta 1903 cuando Paul Unna introdujo el nombre de célula plasmática para referirse a ellas, nombre que ya había sido propuesto previamente por Waldeyer en 1875 para referirse a otro tipo de células descritas en el tejido conjuntivo sin relación con las descubiertas por Cajal[9-11].

En el año 1845, unos años antes del hallazgo de Cajal, un médico y químico inglés, Henry Bence Jones, descubrió en la orina de un enfermo la presencia de una sustancia que al añadirle ácido nítrico se disolvía con el calor y volvía a precipitar al enfriarse. Esta sustancia, a la que llamó "*deutóxido hidratado de albúmina",* fue obtenida estudiando muestras de orina de un paciente de 45 años llamado Thomas Alexander McBean diagnosticado de *mollites ossium,* como se conocía entonces al MM, ya que el término MM no sería acuñado hasta 1873 por el médico ruso von Rustizky[12]. A juicio de Bence Jones sería esta la sustancia clave en el diagnóstico del MM[13]. De manera paralela en 1846, Johann Heller observó que cuando las muestras de orina de pacientes con MM se calentaban ligeramente por encima de los 50ºC, se formaba un precipitado proteico que desaparecia si se aplicaba un calentamiento adicional. Heller consiguió distinguir esta

proteína de la albúmina y de la caseína, y aunque no reconoció su precipitación cuando la orina fue enfriada, se cree que esta podría ser la misma sustancia de la que hablaba Bence-Jones[14]. Casi 35 años después de la observación de este fenómeno, en 1880, Richard Fleischer consiguió aislar una proteína procedente de muestras normales de médula ósea que respondía a las características descritas por Bence-Jones, y en su honor se refirió a esta sustancia como *proteina de Bence-Jones*. Este término acuñado entonces se sigue utilizando en la actualidad y es considerado el primer biomarcador de cáncer de la historia[15]. En 1900, 10 años después de que Cajal describiese las células plasmáticas, Homer-Wright descubrió que estas células eran las células implicadas en el MM cerrando así el capítulo sobre el origen celular del MM[11]. Un episodio más en el papel de las células plasmáticas en el MM se escribió en 1909, cuando Weber propuso que estas células podrían ser la causa de las lesiones óseas que se producen en esta enfermedad[16].

Algunos descubrimientos inmunológicos y los avances técnicos de los años posteriores contribuyeron a perfeccionar el diagnóstico del MM. En este sentido, algunos de los hallazgos más relevantes fueron: el descubrimiento de los anticuerpos[17]; la introducción de nuevas técnicas como la ultracentrifugación y la electroforesis de proteínas que permitieron la separación de las globulinas séricas en tres componentes designados como alfa, beta y gamma[18]; la localización de la actividad de los anticuerpos en la fracción de globulina gamma de las proteínas plasmáticas[19]; la identificación del pico monoclonal característico del MM aplicando técnicas de electroforesis y de inmuno-electroforesis[11, 20, 21]. En el año 1956 se produjo otro gran hallazgo gracias al trabajo de los investigadores Korngold y Lipari, ya que lograron demostrar que el antisuero contra proteínas de Bence-Jones también reaccionaba frente a las proteínas séricas del MM, sugiriendo de esta manera la relación entre ambas. Además, identificaron tres proteínas Bence-Jones diferentes de las que dos siempre se encontraban juntas en la orina del mismo paciente a las que designaron como Kappa y Lambda[22]. Por otro lado, la implantación de los aspirados de médula ósea como técnica diagnóstica permitió la visualización de las células plasmáticas tumorales y su caracterización morfológica[23].

En 1958 se produjo el primer gran hito a nivel farmacológico en el tratamiento del MM con el descubrimiento del potencial terapéutico de la sarcolisina o melfalán. Los estudios de Blokhin y posteriormente de Bergsagel en 1962 demostraron la gran eficacia de este compuesto, capaz de conseguir remisiones del MM hasta en un tercio de los pacientes[24]. Estudios posteriores como el de Alexanian en 1969, en los que se combinaba este compuesto con glucocorticoides como la prednisona, lograron demostrar unos resultados terapéuticos superiores a los del melfalán aplicado en monoterapia[25]. Las terapias combinadas en el MM han sido una constante a lo largo de la historia de esta patología. En el periodo comprendido entre 1976 y 1992 se realizaron múltiples ensayos clínicos combinando distintos agentes quimioterapéuticos, pero ninguno mostró una eficacia superior a la combinación del melfalán con la prednisolona tal y como se recogió en el metaanálisis llevado a cabo por Walter M. Gregory en 1992 sobre 18 estudios que recogían un total de 3.814 pacientes[26]. En estos años, las opciones de tratamiento empiezan a multiplicarse, de manera que en 1984 aparece la terapia VAD (vincristina/adriamicina/dexametasona)[27], se empiezan a llevar a cabo los primeros trasplantes alogénicos y comienzan los estudios para evaluar la quimioterapia a altas dosis con trasplante autólogo de células madre, cuya superiroridad frente a la quimioterapia convencional quedó demostrada a finales de los años 90[28].

En las dos últimas décadas, la aplicación de nuevos tratamientos en el MM no ha dejado de diversificarse con la introducción de nuevos modelos terapeúticos, como por ejemplo los agentes inmunomoduladores (IMiDs) como la talidomida y la lenalidomida, los inhibidores del proteasoma de primera generación como el bortezomib y de segunda generación como el carfilzomib, la aparición de fármacos dirigidos a vías celulares como la deacetilación de histonas como es el caso del panobinostat, o más recientemente la inclusión de las terapias basadas en el uso de anticuerpos monoclonales como el daratumumab.

Estos avances farmacológicos, junto con los grandes descubrimientos que se han realizado en el campo del MM a nivel químico, histológico, citológico, inmunológico y genómico[29] han traído consigo un aumento de la calidad de vida de los pacientes con MM. Sin embargo, la progresión de la enfermedad aún es inevitable y el MM sigue siendo una patología incurable[30], por lo que sigue siendo necesaria la investigación de nuevas moléculas diana para el desarrollo de futuras modalidades terapeúticas.

### 1.1.3. Etiología

La comprensión de la etiología del MM hasta la fecha es muy limitada debido, entre otras causas, a la baja frecuencia poblacional de esta patología. Sin embargo, han sido varios los estudios realizados sobre factores concretos que podrían influir en la aparición y desarrollo del MM. Entre ellos se encuentran factores de riesgo intrínsecos inalterables, como la edad superior a los 65 años o el género masculino[31]. Además, se ha tratado de buscar una posible relación entre factores de riesgo ambientales y el desarrollo del MM. Así, múltiples estudios han tratado de dilucidar la posible influencia de agentes como el tabaquismo[31, 32], el consumo de alcohol[31, 32], la dieta[31, 33-35] o la obesidad[31, 36-40], entre otros, sobre la aparición del MM. Aunque muchos de estos estudios han suscitado gran controversia, no se ha logrado demostrar una asociación concluyente con el desarrollo del MM[41]. Recientemente se ha añadido a la lista de posibles factores ambientales predisponentes la exposición a los ataques del "World Trade Center" de los bomberos de la ciudad de Nueva York[42].

Otra de las posibles causas que se han barajado como posible factor desencadenante del MM son los factores hereditarios. Se ha demostrado la existencia de evidencias de que una historia familiar en la que aparece el MM produce un incremento del riesgo de padecerlo de 1,5 a 5 veces[31]. En este sentido, trabajos como el de Koura y Langston[43] han aportado pruebas que apuntan a la existencia de una predisposición genética a la aparición del MM así como de otras discrasias de células plasmáticas. Esta evidencia viene sustentada tanto por estudios de asociación de genoma completo (GWAS, del inglés *Genome-Wide Association Analysis*), en los que se identificaron posibles loci de susceptibilidad[44, 45], como por estudios de dianas antigénicas de paraproteínas, concretamente las forma hiper-fosforilada de la diana de las paraproteínas 7 y 8 (pP-7 y pP-8), que son heredadas en una forma autosómica dominante, generando propensión a producir respuestas inmunes que llevan a la transformación maligna de las células plasmáticas[46, 47]. Al igual que los factores ambientales, estos resultados han generado un gran debate, lo cuál demuestra la necesidad de estudios moleculares más amplios sobre

la etiología del MM para poder comprender mejor los factores de riesgo y mecanismos que predisponen al desarrollo de esta patología.

### 1.1.4. Biología y fisiopatología

En general, los estudios epidemiológicos han demostrado que el MM evoluciona desde un estadio previo llamado "gammapatía monoclonal de significado incierto (GMSI)" que suele progresar a un segundo estadio conocido como "mieloma múltiple indolente (SMM, del inglés "smoldering MM))". Estas dos entidades cursan de una forma asintomática hasta que se produce la progresión al MM sintomático. El MM sintomático podría perder la dependencia de las señales de supervivencia procedentes del micromedioambiente de la médula ósea y convertirse en un mieloma extramedular o bien evolucionar a una forma más agresiva conocida como leucemia de células plasmáticas (LCP), caracterizada por la presencia de más de un 20% de células plasmáticas en la sangre periférica[48, 49] (**Figura 1.1**).

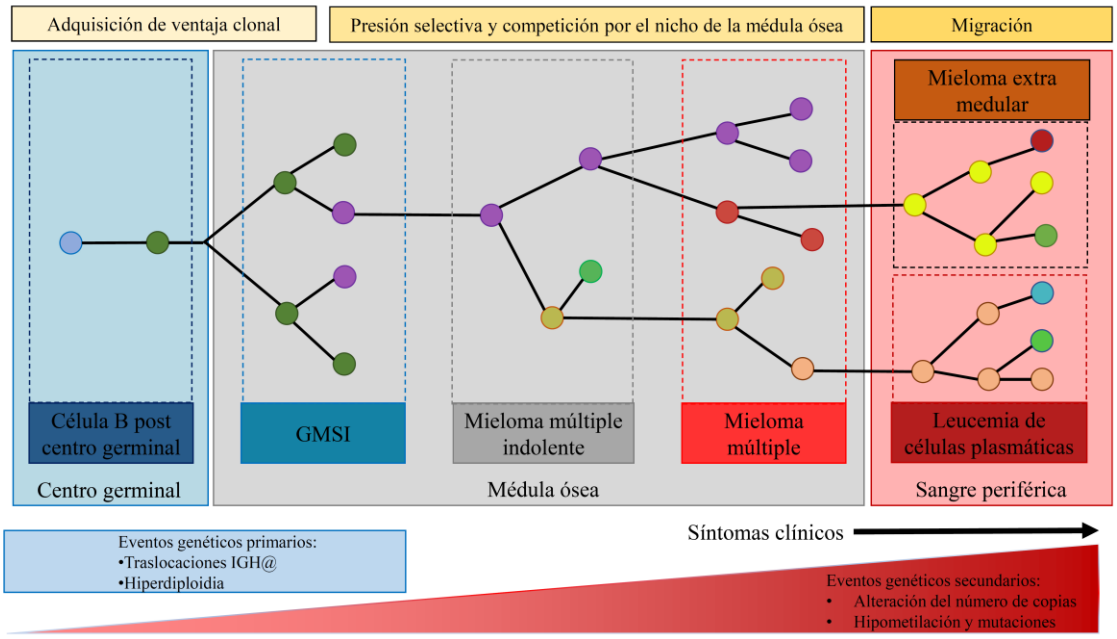



***Figura 1.1.*** *Esquema de la iniciación y el progreso del mieloma múltiple (MM). Este proceso se inicia con los estadios precursores como son la "gammapatía monoclonal de significado incierto (GMSI") y el "mieloma múltiple indolente (SMM)". El MM puede finalmente progresar a enfermedades independientes de la médula ósea, como el mieloma extramedular y la "leucemia de células plasmáticas (LCP)". Los círculos representan la evolución clonal de la célula plasmática partiendo de la célula plasmática premaligna en el centro germinal.*

El desarrollo del MM es un proceso multifactorial que requiere la concurrencia de lo que se conoce como eventos genéticos primarios y de eventos genéticos secundarios para producir el fenotipo tumoral del MM[50]. Los eventos primarios son los responsables del desarrollo de la célula B clonal maligna y están constituidos por las lesiones iniciales que están presentes en la totalidad de las células clonales y marcan el fondo genético de los subgrupos moleculares de MM[51, 52]. Los principales eventos genéticos primarios son las traslocaciones del gen de la cadena pesada de las inmunoglobulinas (*IGH*) y la hiperdiploidía, que generalmente suelen ser mutuamente excluyentes[53]. Las

traslocaciones del gen *IGH*, localizado en la banda cromosómica 14q32, están presentes en casi el 60% de los casos de MM. Aunque este gen puede estar traslocado con diversas regiones cromosómicas, las traslocaciones más frecuentes involucran a los genes *CCND1* en la t(11;14), *NSD2* (*MMSET*) y *FGFR3* en la t(4;14), *MAF* en la t(14;16), *MAFB* en la t(14;20) y *CCND3* en la t(6;14). El resultado de la fusión del potenciador o *enhancer* del locus del gen de la cadena pesada de Igs con los genes acompañantes de la traslocación será expresión anormalmente elevada de estos últimos genes, lo que modificará, entre otros procesos, la proliferación celular[54, 55]. La hiperdiploidía se observa en aproximadamente el 50% de los pacientes con MM[55]. Los cariotipos hiperdiploides se caracterizan por la presencia de trisomías de los cromosomas impares 3, 5, 7, 9, 11, 15, 19 y 21 y, aunque su origen no está claro, se baraja la hipótesis que podría ser el resultado de una mitosis catastrófica[53]. Existe una gran heterogeneidad en cuanto a los efectos que puede tener la presencia de la hiperdiploidía sobre la regulación de otros genes. De hecho, se han encontrando hasta cuatro grupos de pacientes con MM hiperdiploide en los que los patrones de expresión génica estarían bien diferenciados[56]. Tanto las traslocaciones del gen *IGH* como los cariotipos hiperdiploides aparecen en las células plasmáticas clonales de las entidades precursoras del MM, como el SMM y la GMSI.

En lo que respecta a los eventos genéticos secundarios, se cree que estarían implicados en la progresión desde GMSI a MM. Estos eventos son los que proporcionan a un subclon determinado una ventaja selectiva concreta sobre el resto de subclones favoreciendo su prevalencia[53]. Entre los eventos secundarios se incluyen amplificaciones y delecciones recurrentes, como la ganancia del brazo cromosómico 1q o las pérdidas del brazo 1p y de los genes *RB1* o *TP53*, así como fenómenos epigenéticos como la metilación o la regulación de genes por microARNs. Otros eventos secundarios más tardíos son las traslocaciones secundarias que involucran al oncogén *MYC*, o las mutaciones activadoras en genes como *TP53*, *RAS*, *FGFR3*, *CYLD* o *TRAF3*[57].

El micromedioambiente también juega un papel fundamental en la biología del MM. Este esta formado por múltiples tipos de células hematopoyéticas como células B, células T o células *Natural Killer* (NKs), entre otras, además de células que no proceden de la estirpe hematopoyética, como pueden ser las células endoteliales o los osteoblastos. Todos estos tipos celulares son esenciales en el desarrollo del MM ya que, mediante la secreción de determinados factores y citocinas, contribuyen a la migración y proliferación de la célula mielomatosa[54].

Como queda patente en este breve resumen, la biología del MM es extraordinariamente compleja, especialmente debido a la diversidad genómica inter e intratumoral. La investigación incesante de todos estos aspectos clave en la patogenia del MM contribuirá sin duda al desarrollo de nuevas terapias más efectivas contra esta enfermedad.

## 1.2. Terapias en el tratamiento del mieloma múltiple

Aunque el MM aún es una enfermedad incurable, la introducción en los últimos años de nuevas estrategias terapéuticas como el trasplante autólogo de progenitores hematopoyéticos, así como la disponibilidad de cada vez más agentes terapéuticos, ha contribuido a prolongar notablemente la supervivencia global de las personas afectadas por esta patología[1, 7, 8, 58, 59]. De este modo, se cuentan por decenas las diferentes opciones terapéuticas disponibles actualmente para el tratamiento del MM, abarcando desde la quimioterapia convencional, introducida en los años 50 del siglo XX, hasta la más reciente introducción de los tratamientos con anticuerpos monoclonales[60].

### 1.2.1. Quimioterapia convencional

La quimioterapia consiste en el uso de drogas citotóxicas para eliminar las células cancerígenas. Se suele aplicar como un único agente o en combinación con otros fármacos como corticosteroides o agentes inmunomoduladores para incrementar su eficacia. La quimioterapia con melfalán oral unida al glucocorticoide prednisona fue el primer tratamiento eficaz utilizado en pacientes con MM[61].

*Melfalán (Alkeran ®)*

El melfalán (**Figura 1.2**) es un derivado de fenilalanina de la mostaza nitrogenada con actividad alquilante bifuncional que actúa inhibiendo la replicación celular. Este compuesto fue sintetizado por primera vez en 1953 sustituyendo el grupo metilo de la mostaza nitrogenada por el aminoácido fenilalanina[62]. El melfalán actúa como agente alquilante inespecífico del ADN, bloqueando su capacidad replicativa y conduciendo a la aparición de un efecto citotóxico en las células tratadas con este compuesto[63]. El proceso de alquilación también afecta de manera significativa a la transcripción, ya que la formación de aductos mediante la unión de los grupos alquilo a las bases de ADN resulta en la fragmentación del ADN por las enzimas de reparación que tratan de reemplazar estas bases alquiladas, afectando este proceso a la capacidad transcripcional de este ADN[64], conduciendo a la parada del ciclo celular[65] y en último término a la apoptosis celular[66]. Las altas dosis de melfalán continúan siendo el regimen de acodicionamiento más utilizado en el trasplante autólogo de progenitotes hematopoyéticos, procedimiento que forma parte del tratamiento de los pacientes con MM menores de 65-70 años.

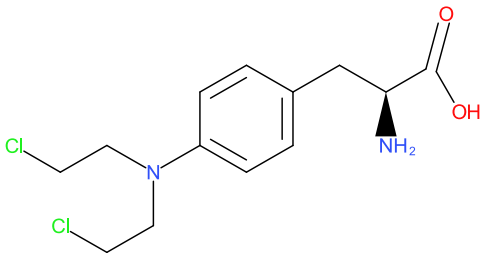



***Figura 1.2**. Estructura química del melfalán*

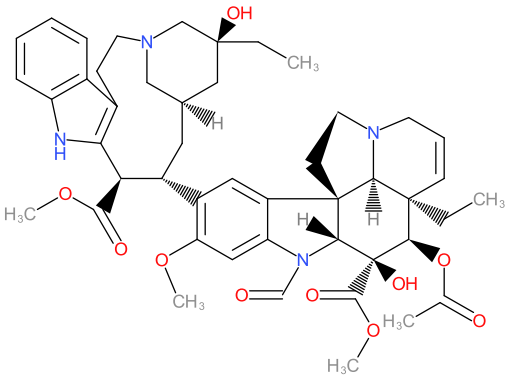


***Figura 1.3.*** *Estructura química de la vincristina*

*Vincristina (Oncovin®)*

La vincristina (**Figura 1.3**) es un alcaloide vegetal aislado de la planta *Vinca rosea* L. de la familia de las Apocináceas. Este compuesto se une de manera irreversible a los microtúbulos y a otras proteínas implicadas en la mitosis celular deteniendo a las células tumorales en metafase[67]. El efecto final de esta parada será la muerte de la célula a través de la activación de las vías apoptóticas[68, 69]. Se han descrito otros efectos asociados a este alcaloide sobre el metabolismo de aminoácidos[70], sobre la actividad $Ca^{2+}$-ATPasa dependiente de calmodulina[71] y sobre la síntesis de ácidos nucleícos[72].

*Ciclofosfamida (Cytoxan®)*

La ciclofosfamida (**Figura 1.4**) es un agente alquilante perteneciente al grupo de las oxazafosforinas[73]. Las principales propiedades de este compuesto son su poder antimitótico y antiproliferativo. Se han demostrado también potentes propiedades inmunosupresoras[74]. La ciclofosfamida necesita ser activada enzimáticamente en el hígado para tener propiedades citotóxicas, formando dos moléculas alquilantes del ADN. Estas actúan uniéndose a una de las dos hebras del ADN de las células cancerosas impidiendo la división celular[75].

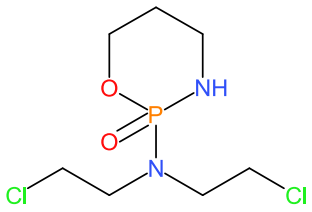


***Figura 1.4.*** *Estructura química de la ciclofosfamida*

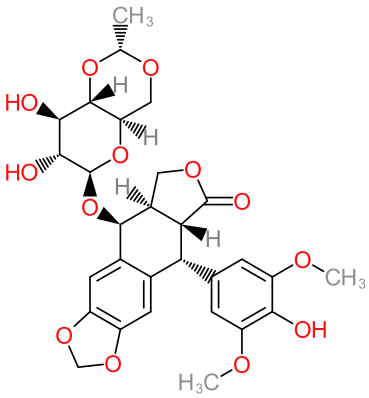


***Figura 1.5.*** *Estructura química del etopósido*

*Etopósido (VP-16, VePesid®)*

El etopósido (**Figura 1.5**) es un alcaloide vegetal[76] inhibidor de la ADN topoisomerasa II. El etopósido estabiliza la unión de la topoisomerasa II al ADN, de manera que impide la unión de las hebras rotas de ADN y por tanto la correcta condensación cromosómica[77]. Esto llevará a la célula a cometer errores en la síntesis de ADN y por tanto se promoverá la apoptosis celular[78].

*Doxorrubicina (Adriamicina o Rubex®)*

La doxorrubicina (**Figura 1.6**) es un compuesto de la familia de las antraciclinas aislado del hongo *Streptomyces peucetius*[79], que inhibe la síntesis de ADN, ARN y proteínas, potenciando de esta manera su efecto citotóxico. Este compuesto se intercala entre las bases de ADN logrando que las ARN y las ADN polimerasas sean inhibidas. También actúa impidiendo que la topoisomerasa II pueda progresar, inhibiendo de esta forma que la cadena de ADN pueda transcribirse[77]. La doxorrubicina también favorece la formación de especies reactivas de oxígeno, lo que favorece su actividad citotóxica[80].

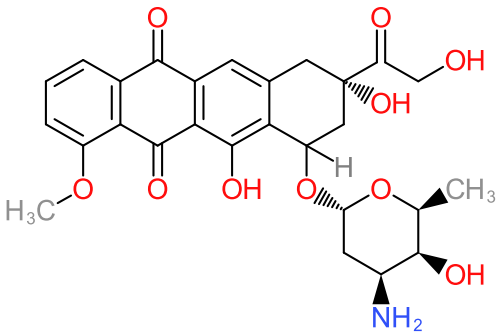



***Figura 1.6.*** *Estructura química de la doxorrubicina*

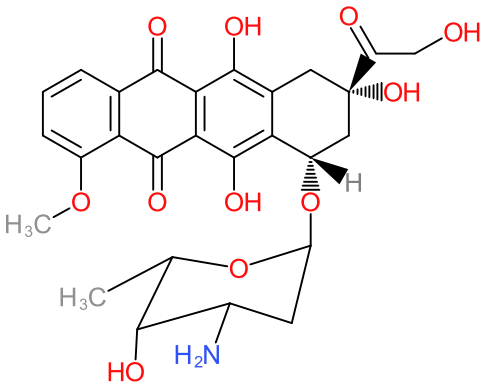



***Figura 1.7.*** *Estructura química de la doxorrubicina liposomal*

*Doxorubicina liposomal (Doxil®)*

La doxorubicina liposomal (**Figura 1.7**) es una forma pegilada de la doxorubicina encapsulada en una esfera lipídica o liposoma. La principal ventaja de este compuesto frente a la forma no encapsulada de doxorrubicina es su bajo efecto cardiotóxico[81].

*Bendamustina (Treanda®)*

La bendamustina (**Figura 1.8**) es un agente antineoplásico alquilante del grupo de las mostazas nitrogenadas[82]. Este compuesto actúa induciendo la apoptosis de las células tumorales gracias a su actividad alquilante dependiente de p53, aunque también ha sido descrito un mecanismo de muerte no dependiente de p53[83]

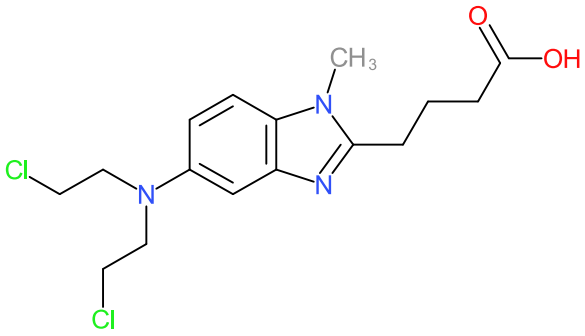



***Figura 1.8.*** *Estructura química de la bendamustina*

### 1.2.2. Corticosteroides

Los corticosteroides son fármacos antiinflamatorios que imitan el efecto de las hormonas producidas de manera natural por las glándulas suprarrenales. Estos compuestos están implicados en múltiples procesos fisiológicos, entre los que destacan los procesos antiinflamatorios y la inmunosupresión. Los corticosteroides son uno de los compuestos más comúnmente utilizados para el tratamiento del cáncer[84]. El MM no es una excepción, ya que desde los años 70 se utiliza por su actividad citotóxica frente a las células plasmáticas[85].

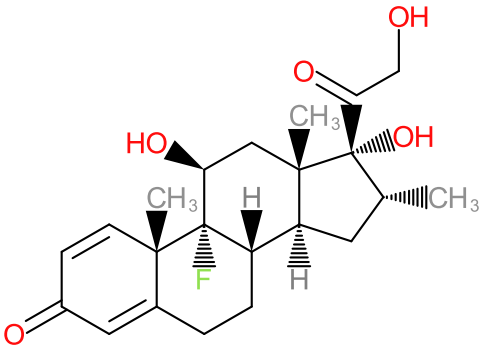



***Figura 1.9.*** *Estructura química de la dexametasona*

*Dexametasona*

La dexametasona (**Figura 1.9**) es un glucocorticoide sintético que actúa como antiinflamatorio[86] e inmunosupresor[87]. La dexametasona se une a su receptor formando un complejo que penetra en el núcleo de la célula uniéndose al ADN formando un complejo que se unirá a los elementos de respuesta a glucocorticoides del ADN (GRE)[88], modulando la transcripción y por tanto la síntesis proteica[86].

*Prednisona*

La prednisona (**Figura 1.10**) es un glucocorticoide sintético también con acción antiinflamatoria e inmunosupresora. Al igual que la dexametasona, forma un complejo receptor-glucocorticoide que entra en el núcleo de la célula e interacciona con el ADN modulando la transcripción génica.

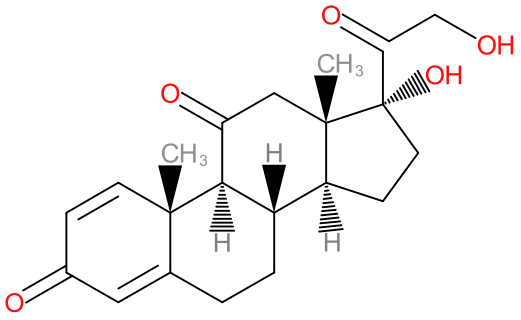



***Figura 1.10.*** *Estructura química de la prednisona*

### 1.2.3. Agentes inmunomoduladores

Los agentes inmunomoduladores (IMiDs) son compuestos que afectan a la capacidad del sistema inmune para llevar a cabo sus funciones, de modo que pueden tener función inmunoestimuladora o inmunosupresora. Además, se ha observado en el MM que estos compuestos presentan también capacidad antiproliferativa, antiangiogénica y antiinflamatoria[89]. Por todo esto, el descubrimiento de los IMiDs ha supuesto un gran avance en el tratamiento del MM ya que ha contribuido a mejorar notablemente tanto la supervivencia como la calidad de vida de los pacientes[90, 91].

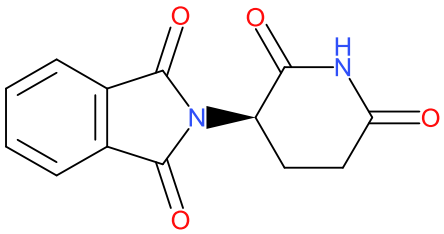



***Figura 1.11.*** *Estructura química de la talidomida*

*Talidomida (Thalomid®)*

La talidomida (**Figura 1.11**) es un derivado del ácido glutámico con efecto inmunomodulador y potencialmente antineoplásico. Este compuesto modula la síntesis de citocinas, principalmente del factor de necrosis tumoral alfa (TNFα)[92], reduciendo la vida media de su ARNm. Este compuesto también tiene un papel inhibitorio en la ruta de fagocitosis[93] y angiogénesis[94].

*Lenalidomida (Revlimid®)*

La lenalidomida (**Figura 1.12**) es un compuesto con actividad antineoplásica análogo de la talidomida, con mayor potencia y menor toxicidad que esta última[95]. El principal mecanismo de acción de la lenalidomida es la ubiquitinación y posterior destrucción de sustratos proteicos como Ikaros o Aiolos a través de su unión con la proteína codificada por el gen *CRBN*[96], aunque la lenalidomida también actúa a nivel transcripcional[97]. Asímismo, este compuesto inhibe la producción de citocinas proinflamatorias como TNFα o IL-1, IL-6 e IL-12, además de promover la síntesis de citocinas antiinflamatorias[95, 98, 99]. La lenalidomida inhibe también la angiogénesis y promueve la parada en G1 y la apoptosis de las células tumorales[100].

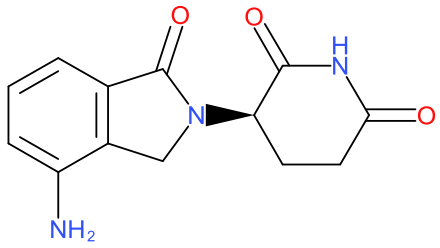



***Figura 1.12.*** *Estructura química de la lenalidomida*

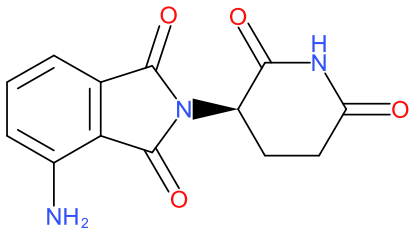



***Figura 1.13.*** *Estructura química de la pomalidomida*

*Pomalidomida (Pomalyst®)*

La pomalidomida (**Figura 1.13**) es un agente inmunomodulador derivado de la talidomida con actividad antimieloma[101]. Este compuesto inhibe directamente la angiogénesis y también el crecimiento[101] y la proliferación de las células tumorales, además de inducir su apoptosis[102]. A estas actividades contribuye el incremento de la producción del interferón gamma y las interleucinas 2 y 10, así como la reducción de la producción de citocinas proinflamatorias como la IL-6[98]. Se ha descrito que este compuesto además favorece la respuesta inmune mediada por células T y células NK[103]. Sin embargo, el mecanismo de acción de la pomalidomida no se limita a estos procesos, ya que también se ha registrado un efecto modulador sobre la expresión génica, bien a través de mecanismos inhibitorios como es el caso de la prostaglandina-endoperóxido sintasa 2 (*PTGS2*) en monocitos[104], o bien a través de mecanismos de inducción de la expresión génica como ocurre con el gen *CDKN1A* en líneas celulares de linfoma de Burkitt[105]. Este efecto sobre la expresión génica fue evaluado en este trabajo en MM mediante técnicas de metaanálisis

### 1.2.4. Inhibidores del proteasoma

Descubierta en los años 80, la vía de la ubiquitina-proteasoma tiene un papel esencial en la maquinaria de degradación proteica de la célula[106]. El papel central en esta vía lo juega un complejo multiproteico llamado proteasoma, cuya principal función es la destrucción de las proteínas marcadas previamente con ubiquitina. Se ha visto que las células cancerígenas presentan una mayor actividad de este complejo que las células normales[107], lo que ha sido una de las razones para pensar en el proteasoma como posible diana terapeútica frente a algunos tipos de cáncer como el MM[108]. Esto ha llevado al desarrollo de múltiples compuestos como bortezomib, carfilzomibo o ixazomib, cuyo mecanismo de acción provoca la inhibición de este complejo, causando la inducción en la célula tumoral de la cascada de apoptosis como resultado de la acumulación de proteínas mal plegadas en la célula[109].

*Bortezomib (Velcade®, PS-341)*

El bortezomib (**Figura 1.14**) es un ácido borónico dipeptídico modificado. Se trata de un inhibidor reversible del proteasoma que ha sido utilizado tanto en monoterapia como en combinación con otros fármacos en el tratamiento del MM. Este fármaco actúa sobre la ruta ubiquitina-proteasoma de homeostasis de proteínas celulares bloqueando la acción del proteasoma 26S, que es una enzima multi-catalítica que degrada las proteínas mal plegadas o aberrantes[110]. De manera particular, el bortezomib inhibe la activación de las subunidades β5 y β1 del núcleo 20S en el complejo del proteasoma 26S. El bortezomib se une con el sitio activo de estas subunidades formando un complejo que produce el bloqueo reversible la actividad tipo quimotripsina[111] y de la actividad tipo caspasa[112] de ambas subunidades, respectivamente, lo que conduce a la inducción de la apoptosis de las células tratadas con este compuesto[113]. El proteasoma 26S también juega un papel relevante en la regulación de la transcripción[114], por lo que su inhibición podría conducir a la desregulación de múltiples vías y funciones en la célula tratada con bortezomib. El bortezomib además inhibe la activación del factor de necrosis tumoral kappa B (NF-κB), bloquea la producción y señalización intracelular de la interleucina 6 e inhibe la angiogénesis, entre otros efectos[115].

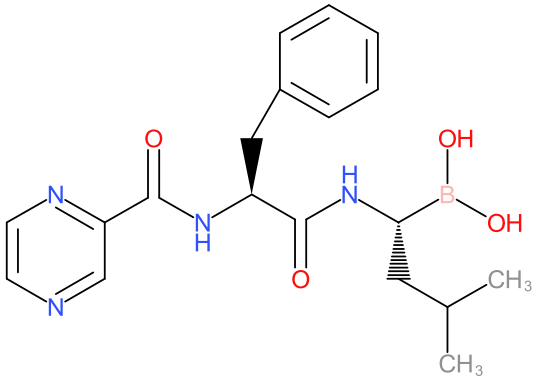



***Figura 1.14.*** *Estructura química del bortezomib*

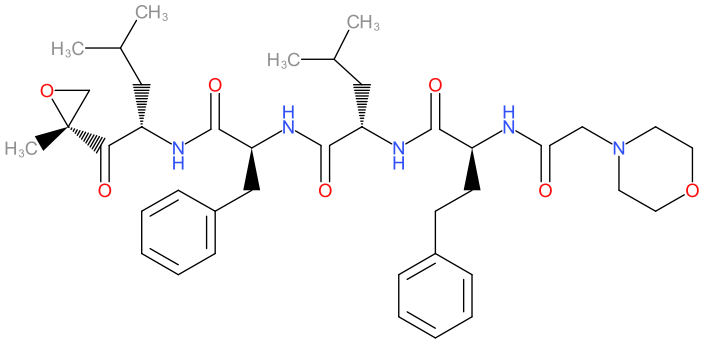


***Figura 1.16.*** *Estructura química del carfilzomib*

*Carfilzomib (Kyprolis®)*

El carfilzomib (**Figura 1.16**) es un inhibidor irreversible del proteasoma que pertenece al grupo de las epoxicetonas. El efecto de inhibición irreversible lo lleva a cabo mediante la unión específica e irreversible a los sitios activos treonina N-terminales del proteasoma[116, 117]. Se ha observado que el tratamiento con carfilzomib tiene una potente actividad antiproliferativa y proapoptótica sobre células malignas en cánceres hematológicos[118].

*Ixazomib (Ninlaro®)*

El ixazomib (**Figura 1.17**) es un inhibidor del proteasoma tipo boronato de segunda generación, de administración oral a diferencia del bortezomib y del carfilzomib. Se trata de un inhibidor reversible del proteasoma que se une e inhibe selectivamente a la subunidad beta tipo 5 (PSMB5) del proteasoma 20S[119]. Estudios de este compuesto en líneas celulares de MM han demostrado su capacidad para inducir la apoptosis[120]. Esta actividad se ha comprobado que es dependiente de caspasas involucrando además otras vías de señalización, como p53-p21 o la vía de estrés del retículo endoplasmático[121].

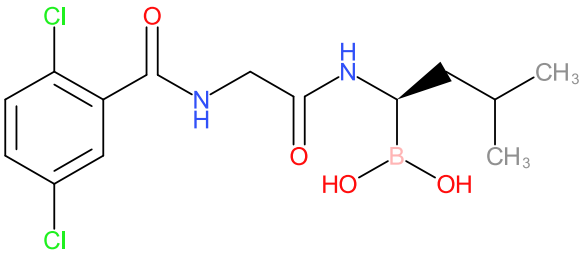


***Figura 1.17.*** *Estructura química del ixazomib*

### 1.2.5. Inhibidores de las deacetilasas de histonas

El desarrollo de los compuestos inhibidores de las deacetilasas de histonas (HDACi) como terapia antineoplásica es la respuesta a multitud de estudios en los que se ha reportado que la hiperacetilación y la hipoacetilación de las histonas es un proceso importante y frecuente en cáncer[122, 123], incluido el MM[124]. De manera general, los HDACi actúan inhibiendo las enzimas que participan en la desacetilación de las histonas presentes en la cromatina, proteínas que son de gran importancia en procesos como la transcripción celular, produciendo finalmente la detención del ciclo celular y la diferenciación y la muerte celular[125].

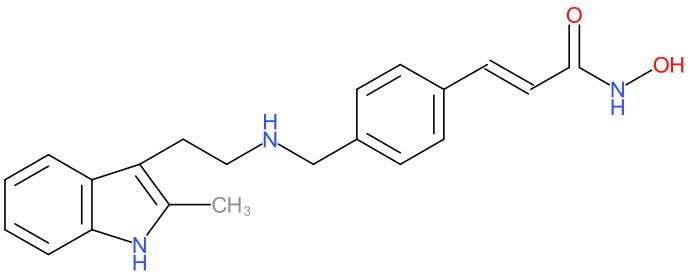



***Figura 1.18.*** *Estructura química del panobinostat*

*Panobinostat (Farydak®)*

El panobinostat (**Figura 1.18**) es un derivado del ácido hidroxámico que produce la inhibición de las desacetilasas de clase I y II, que son responsables de la regulación de múltiples procesos en la célula como la transcripción o la apoptosis[126]. La inhibición de estas enzimas por panobinostat evita la desacetilación de las histonas haciendo que se produzcan cambios en la expresión génica debidos a los daños producidos en el ADN[127] por la relajación de la cromatina[128]. Se ha demostrado en estudios *in vitro* su poder sinérgico con otros compuestos promoviendo la actividad celular proapoptótica y la detención del ciclo celular en la fase G2/M, mediante la acumulación de histonas y otras proteínas acetiladas[129, 130].

### 1.2.6. Agentes hipometilantes

Los agentes hipometilantes son una familia de compuestos utilizados en el tratamiento de diversos tipos de cánceres hematológicos, como los síndromes mielodisplásicos o el MM, cuyo principal efecto es la redución de grupos metilo que están unidos de forma natural al ADN[131]. Estos fármacos se incorporan al ADN reemplazando la citosina y formando uniones irreversibles con la enzima ADN-Metil-Transferasa (DNMT). Al restringir la presencia de dicho enzima, el ADN que se sintetice presentará un menor grado de metilación y se reexpresará[132].

*Azacitidina (Vidaza ®)*

La azacitidina (**Figura 1.19**) es una pirimidina análogo químico del nucleósido citidina. Aunque originalmente su destino fue el uso como droga citotóxica, la azacitidina actúa sobre las células tumorales siguiendo un doble mecanismo. Por un lado, se ha comprobado que a bajas dosis este compuesto produce su actividad antineoplásica mediante la inhibición de la ADN metiltranferasa, lo que produce a la hipometilación del ADN[133, 134], conduciendo a la activación de los genes silenciados por metilación[135]. Por otra parte, se ha comprobado que la azacitinida produce citotoxicidad directa en células hematopoyéticas cancerígenas, ya que al ser un análogo del nucleósido citosina puede ser incorporado directamente al ADN o al ARN[136, 137]. Mediante este segundo mecanismo, la azacitidina consigue activar de forma selectiva la expresión de ciertos genes de las células sometidas a tratamiento[138]. Además, la azacitidina puede conducir mediante su incorporación al ARN al desensamblaje de los ribosomas ocasionando la inhibición de la producción de proteínas[139].

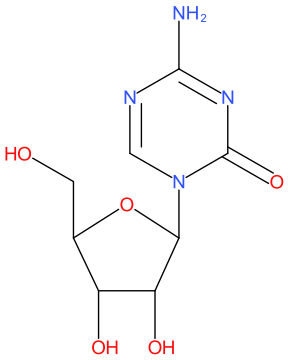



***Figura 1.19.*** *Estructura química de la azacitidina*

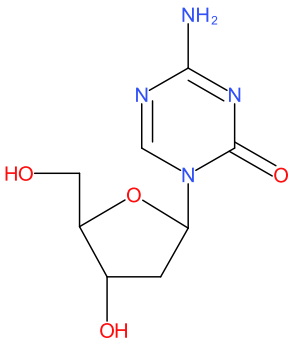



***Figura 1.20.*** *Estructura química de la decitabina*

*Decitabina (Dacogen ®)*

La decitabina (**Figura 1.20**) es un agente hipometilante desoxiderivado de la azacitidina. Al igual que la azacitidina, la decitabina produce hipometilación en el ADN a través de la inhibición de la enzima ADN metiltransferasa[140]. Esta acción la lleva a cabo mediante su incorporación directa en la hebra de ADN, sin embargo, a diferencia de la azacitidina no se incorpora a cadena de ARN debido a su estructura de desoxirribonucleótido. Como resultado de esta hipometilación, la decitabina produce la restauración de la función de genes críticos para la diferenciación y proliferación celular, como ocurre con el gen *TNFRSF10A* en células de leucemia mieloide aguda (LMA)[141]. La decitabina, como la azacitidina, también tiene un mecanismo de acción dual, de manera que a bajas dosis produce la hipometilación del ADN, pero a altas dosis conduce a la inhibición de la proliferación celular[139]. No obstante, se ha demostrado en algunos tipos de cáncer que la decitabina presenta efectos diferentes a los producidos por la azacitidina tanto a nivel de expresión génica, como de ciclo celular, daño en el ADN y apoptosis[142].

### 1.2.7. Anticuerpos monoclonales

Los anticuerpos monoclonales son proteínas recombinantes complejas que reaccionan específicamente contra epítopos concretos de sus proteínas diana. Estas moléculas han sido ampliamente utilizadas como agentes inmunoterapeúticos en numerosos tumores. En el MM, el uso de anticuerpos monoclonales se ha convertido en una alternativa de tratamiento debido a los numerosos antígenos presentes en la superficie de la célula plasmática, como la glicoproteina CD38, el receptor CS1 y el antígeno BCMA, entre otros[143]. Por todo esto, los anticuerpos monoclonales están llamados a ser la nueva revolución en el tratamiento de esta patología[144].

*Daratubumab (Darzalex®)*

Daratumumab es un anticuerpo monoclonal humano contra la glicoproteína transmembrana CD38. Este compuesto induce citotoxicidad dependiente de anticuerpo y dependiente de complemento[145]. Otros efectos desencadenados por este anticuerpo es la fagocitosis celular dependiente de anticuerpo[145, 146] y la inhibición de la actividad enzimática de CD38[147].

*Elotuzumab (Empliciti®)*

Elotuzumab es un anticuerpo monoclonal humanizado que se une a la proteína SLAMF7, proteína expresada en MM y en células *natural killer* (NK)[148]. El elotuzumab ejerce un efecto doble, ya que por un lado activa directamente las células NK y por otro induce la citotoxicidad dependiente de anticuerpos mediada por células a través de CD16[149].

### 1.2.8. Interferón γ

El interferón γ (IFN-γ) es una citocina que es secretada por las células T activas y células NK, de gran importancia en la respuesta inmune, principalmente en procesos de infección vírica o bacteriana. Esta molécula goza de un gran potencial terapéutico en procesos cancerígenos, lo que ha conducido a su desarrollo como agente farmacológico[150]. El MM es uno de los procesos tumorales en los que se ha podido comprobar este potencial, ya que se ha demostrado su poder inhibitorio sobre la interleucina 6 (IL-6), esencial en el desarrollo del MM[151], conduciendo el tratamiento con IFN-γ a la supresión del crecimiento celular dependiente de la propia IL-6[152].

### 1.2.9. Trasplante de células progenitoras hematopoyéticas

El trasplante de células progenitoras hematopoyéticas consiste en la administración al paciente con MM de nuevas células madre productoras de sangre, que pueden provenir tanto de la médula ósea como de la sangre periférica. Existen dos tipos de trasplante en función del individuo del que procedan las células trasplantadas:

a) *Trasplante autólogo*: las células madre proceden del propio paciente con MM
b) *Trasplante alogénico*: las células madre proceden de otra persona (donante), este tipo de trasplante es de mayor riesgo ya que puede producir lo que se conoce como *enfermedad injerto contra huésped*, en la que las nuevas células inmunes identifican las células del paciente como extrañas y son atacadas.

## 1.3. Agentes terapéuticos en fase de investigación en mieloma múltiple

### 1.3.1. Inhibidores de bromodominio

Los bromodominios son una familia de motivos proteicos que se unen a los residuos de lisina acetilados en las colas de las histonas, los cuales juegan un papel muy relevante en los procesos de remodelado de la cromatina y la activación transcripcional[153, 154]. Su relevancia en el campo del MM viene dada por el descubrimiento de su potencial acción reguladora sobre la expresión génica de *MYC*[155], oncogén directamente involucrado en el desarrollo del MM[156]. El uso de inhibidores de bromodominio se ha propuesto como una potencial estrategia terapéutica en el MM, ya que se ha demostrado en ensayos *in vitro* e *in vivo* su capacidad reguladora de *MYC*, induciendo un efecto antiproliferativo en las células mielomatosas[155].

*JQ1*

El JQ1 (**Figura 1.21**) es una tienotriazoldiazepina con actividad inhibidora de proteínas de bromodominio extraterminal (BET). Este compuesto se une de forma competitiva a los motivos de reconocimiento de acetil lisina, también llamados bromodominios[157]. El JQ1 aún no ha sido utilizado en ensayos clínicos en humanos debido a su corta vida media, sin embargo, estudios en modelos *in vitro* e *in vivo* han demostrado su eficacia sobre células malignas. De esta manera, estudios en líneas celulares de MM han demostrado que este compuesto es capaz de reducir la liberación y la transcripción de interleucina 6 inducida por lipopolisacáridos[158]. También se ha comprobado que este compuesto regula la transcripción de *MYC* y las dianas dependientes de este gen, así como su papel sobre la inhibición de la proliferación celular en MM, a través de la potente inhibición que ejerce sobre la proteína Brd4, responsable del control transcripcional del oncogén *MYC*[155]. Así, esta regulación transcripcional de *MYC* implica que la aplicación del JQ1 también tiene efectos sobre el transcriptoma celular, por lo que en este trabajo se buscó dilucidar mediante técnicas de metaanálisis la modulación transcriptómica de este compuesto

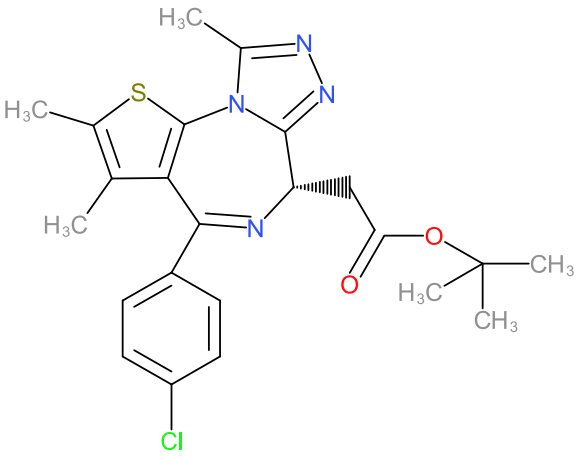



***Figura 1.21.*** *Estructura química del JQ1*

### 1.3.2. Agentes moduladores del *splicing* alternativo

El *splicing* alternativo es un proceso posttranscripcional de reordenamiento combinatorio que ocurre en el núcleo de la célula, en el que, mediante la unión selectiva de exones, partes de exones y/o partes de intrones procedentes de un único preARNm, se pueden producir multiples moléculas de ARNm maduro estructuralmente diferentes. Este proceso es crítico para multitud de funciones de la célula[159], y también se ha observado su implicación en múltiples patologías como el cáncer[160]. En el MM, se ha comprobado que la desregulación del *splicing* alternativo tiene una repercusión en la respuesta al tratamiento y en el pronóstico[161-164]. Sin embargo, aún queda mucho por investigar en este campo buscando posibles dianas que permitan el desarrollo de terapias que modulen del *splicing* alternativo en el MM.

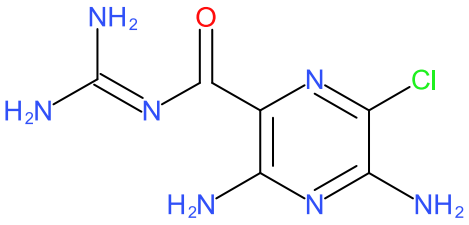



***Figura 1.22.*** *Estructura química de la amilorida*

*Amilorida (Midamor®)*

La amilorida (**Figura 1.22**) es una piracina-carbonil-guanidina sinterizada en los años 60 del siglo XX que funciona como agente antidiurético. Se trata de un compuesto inhibidor de canales de $Na^{+}$[165] aunque también se ha podido comprobar su eficacia sobre otros procesos de transporte de iones[166]. El hallazgo de sus propiedades antimetastásicas y antitumorales, debidas principalmente a esta inhibición de las proteínas intercambiadoras de protones[167] y a su capacidad de modulación del *splicing* alternativo a través de la fosforilación de factores de regulación del *splicing*[168], ha conducido a la investigación de este compuesto como una potencial terapia en MM[169] y leucemia mieloide crónica (LMC)[170]. Este fármaco ha sido elegido en el presente trabajo para la puesta a punto de un flujo de trabajo o *pipeline* utilizando datos de RNA-seq.

*TG003*

TG003 (**Figura 1.23**) es un benzotiazol con función de inhibidor de la familia de quinasas Clk. Se ha descrito su efecto sobre la regulación del *splicing* alternativo modulado por la fosforilación de las proteínas SR a través de la inhibición de las quinasas Clk1 y Clk4, lo que conduce esta modulación del *splicing*[171]. El hecho de que este fármaco actúe sobre las proteínas SR, concretamente sobre SRSF1 le confiere además un especial interés en la investigación del cáncer, ya que SRSF1 es un protooncogén que contribuye a la progresión tumoral mediante el cambio de los patrones de *splicing* de genes como el supresor tumoral *BIN1*[172]. En MM se ha podido observar que el tratamiento con este compuesto provoca la inhibición del crecimiento celular y la supervivencia celular[173]. Este compuesto, junto con la amilorida, ha sido seleccionado para la puesta a punto de un *pipeline* de RNA-seq en el presente trabajo.

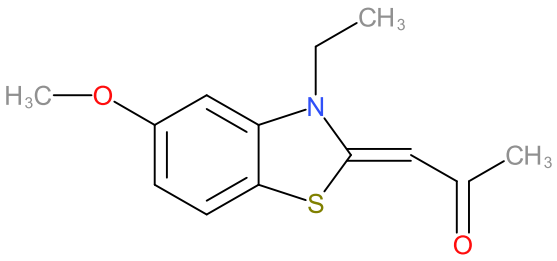
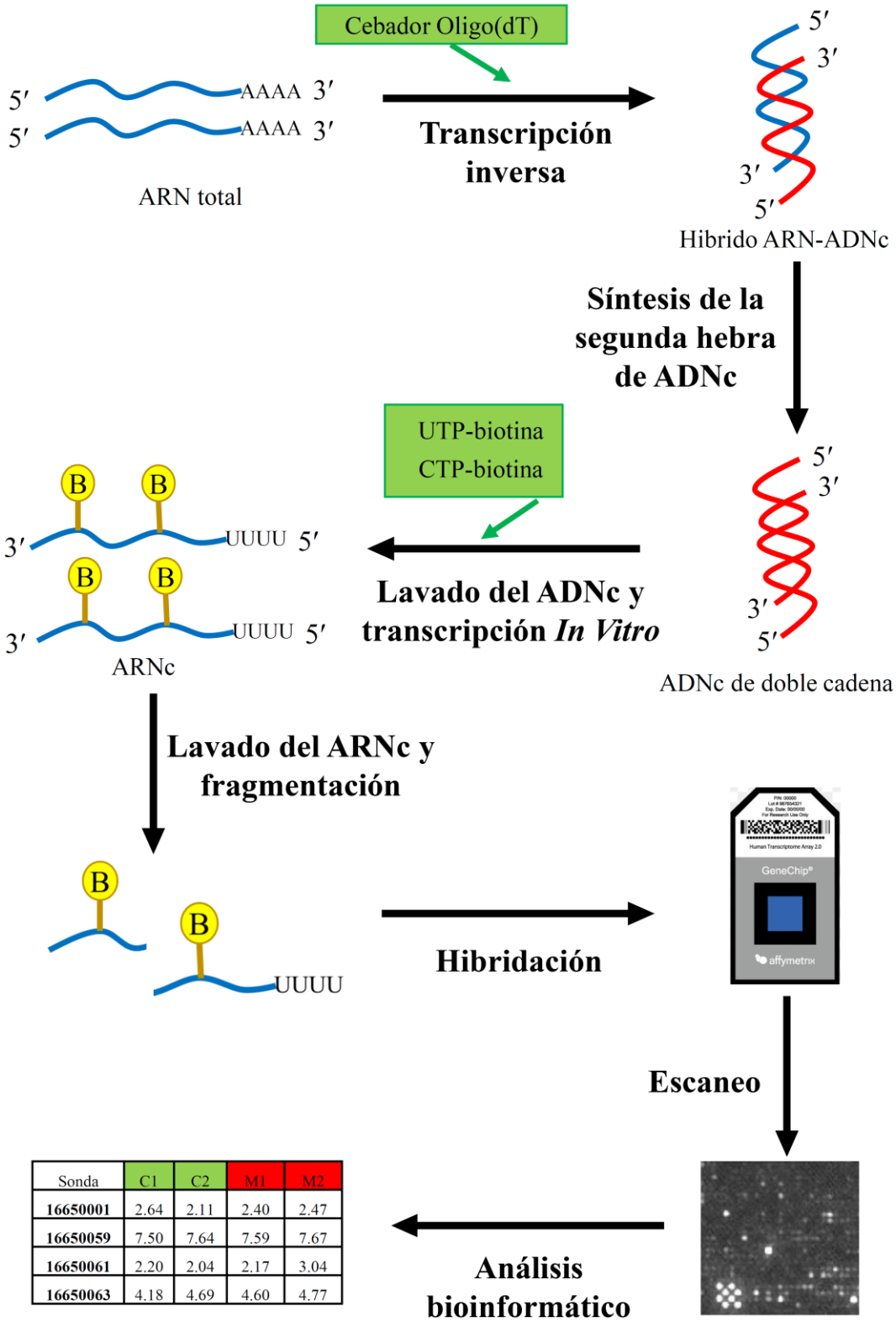



***Figura 1.23.*** *Estructura química del TG003*

## 1.4. Microarrays en la investigación del mieloma múltiple

El origen de los microarrays se remonta a la década de los 70 del siglo XX, cuando en 1975, Grunstein y Hogness, diseñaron un método de hibridación molecular de ADN en cultivos de *Escherichia coli*, utilizando placas de agar cubiertas con filtros de nitrocelulosa. Posteriormente se aplicaba sobre este filtro una sonda etiquetada con radiación que hibridaba en cada muestra con ADN complementario[174]. Años más tarde, en 1979, Gergen y colaboradores adaptarían este método para producir matrices en las que se encontrasen más de 1.000 colonias de bacterias[175]. A finales de los 80, se llevó a cabo la robotización del proceso, produciendo una notable reducción del tiempo de procesamiento y un gran aumento de la precisión de los resultados obtenidos. Esto junto con el desarrollo en los 90 de la clonación del ADN complementario, supusieron un punto de partida importante para el desarrollo de los microarrays. La miniaturización de estas matrices en el año 1995 condujo a la utilización por primera vez del concepto microarray en estudios de expresión génica por Schema y colaboradores[176]. Todo esto supuso un gran avance en los estudios genómicos que terminó de materializarse en el año 1997 cuando por primera vez se publicó el primer estudio de expresión génica mediante un microarray de genoma completo de levaduras[177]. Desde entonces, el uso de los microarrays se ha generalizado a múltiples especies, para el estudio de múltiples condiciones y patologías y con numerosas variantes, como pueden ser los estudios de expresión génica, de metiloma, número de copias, microRNAs, etc.

Las aplicaciones más extendidas de los microarrays en estos años han sido su utilización en estudios de expresión génica para identificar posibles biomarcadores tumorales, diferenciar subtipos moleculares en una patología determinada o para averiguar los efectos de compuestos químicos sobre la expresión génica[178]. Actualmente, los microarrays de expresión génica consisten en una matriz sólida de vidrio, plástico o silicona, con miles de pequeñas áreas o *spots* que contienen una secuencia conocida de oligonucleótidos, a la que se hibrida una librería de fragmentos de ARN complementario (ARNc) generada a partir del ARNm procedente de la muestra en la que se quiere conocer la expresión génica. Aunque la hibridación del ARNc es el punto clave en el procesamiento del microarray[179, 180], el proceso experimental y analítico del microarray consta de muchos más pasos, que se resumen brevemente en la **Figura 1.24**.




| Sonda | C1 | C2 | M1 | M2 |
|---|---|---|---|---|
| **16650001** | 2.64 | 2.11 | 2.40 | 2.47 |
| **16650059** | 7.50 | 7.64 | 7.59 | 7.67 |
| **16650061** | 2.20 | 2.04 | 2.17 | 3.04 |
| **16650063** | 4.18 | 4.69 | 4.60 | 4.77 |

***Figura1.24.*** *Esquema del procesamiento general para los microarrays de expresión génica de Affymetrix. Adaptado de The Affymetrix GeneChip Platform: An Overview, Dalma-Weishausz, 2006.*

El objetivo final de todo este proceso es recopilar los datos de expresión génica del conjunto de muestras sujetas a análisis, para finalmente crear perfiles de expresión génica que muestren cambios simultáneos de la expresión de varios genes en función de las diferentes condiciones objeto del estudio. Sin embargo, aunque analizar simultáneamente varios genes pueda parecer *a priori* una ventaja, esto no ha sido visto así siempre. El hecho de ser una técnica capaz de generar una cantidad tan masiva de datos, al interrogar simultáneamente la expresión de miles de genes, ha generado cierto escepticismo en la comunidad científica[181] debido a lo que se conoce como "maldición de la dimensionalidad" (término acuñado por Bellman en 1957[182]). A pesar de ello, esta técnica goza actualmente de un amplio reconocimiento en estudios biomédicos, y es de gran utilidad en multitud de áreas de investigación como el cáncer.

El primer trabajo que utilizó microarrays para el estudio de la expresión génica en el campo de la oncología se realizó en el área de la hematología en 1999. En este estudio pionero, Golub y colaboradores[183] demostraron el poder de los microarrays en la clasificación y predicción de pacientes con leucemia mieloide aguda (LMA) y leucemia linfoblástica aguda (LLA) en función de los niveles de expresión génica de 6.817 genes. Desde entonces, el número de estudios publicados en cáncer en los que la expresión génica ha sido estudiada mediante microarrays ha aumentado progresivamente, hasta las dos últimas décadas donde se observa cierto estancamiento **Figura 1.25a**. Estos patrones de publicación también aparecen en el campo de la oncohematología, donde de manera particular en el MM podemos observar un comportamiento análogo **Figura 1.25b**.

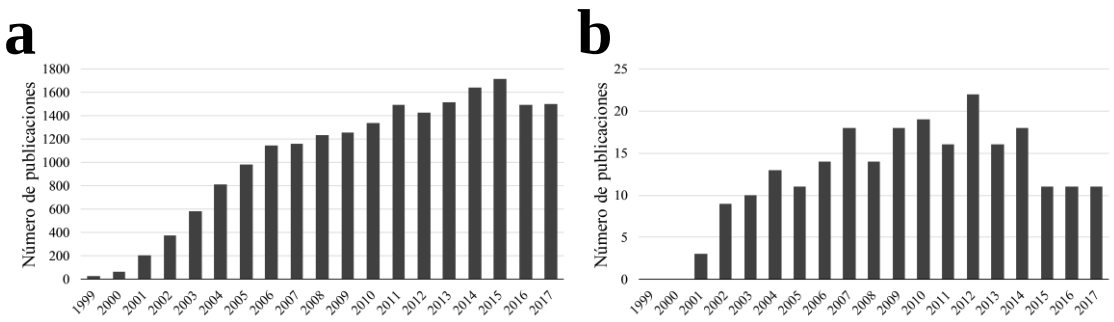



***Figura 1.25.*** *Número de publicaciones por año en Pubmed para estudios de expresión génica con microarrays en* ***a)*** *cáncer y* ***b)*** *mieloma múltiple. Las búsquedas se llevaron a cabo a través de la página web http://dan.corlan.net/medline-trend.html, utilizando como términos de búsqueda en* ***a)*** *"microarray AND cancer AND gene AND expression" y en el caso de* ***b)*** *"microarray AND myeloma AND gene AND expression".*

En estos años, los estudios transcriptómicos en MM han generado información relevante sobre la patogenia de esta enfermedad[184]. En este sentido, se descubrieron ocho subgrupos de pacientes con comportamientos clínico-biológicos bien diferenciados, lo que demuestra la enorme heterogeneidad molecular del MM. Aunque sin duda, uno de los hallazgos más relevantes y, en cierta manera inesperado, que revelaon los estudios de expresión génica en el MM fue la expresión significativamente aumentada de una de las ciclinas D (*CCND1, CCND2 o CCND3*) en todos los MM e incluso en las GMSI[185-187]. El desarrollo de la mielomagénesis ha sido otro de los puntos destacados en el análisis transcriptómico del MM, donde se han estudiado cohortes de pacientes que abarcan todos o parte de los estadios evolutivos de la enfermedad, logrando identificar firmas génicas

relacionadas con el riesgo de progresión del MM[188], o vías de señalización desreguladas entre los diferentes estadios del MM[189].

Los estudios de expresión génica mediante microarrays han tenido también gran relevancia en el campo de la farmacología en el MM. Una de las aplicaciones más comunes de esta técnica ha sido el estudio del efecto del tratamiento con fármacos en modelos *in vitro* en líneas celulares de MM (HMCLs, del inglés *Human Myeloma Cell Lines*) sensibles a estos compuestos. La meta de estos estudios ha sido la determinación de una firma de expresión génica en respuesta a un determinado tratamiento y así, elucidar sus posibles mecanismos de acción en la célula mielomatosa. El número de estudios publicados con este objetivo se cuenta por decenas, abarcando un gran abanico de fármacos, ya que podemos encontrar estudios que comprenden desde la quimioterapia convencional con fármacos como el melfalán[190, 191], hasta estudios que buscan la posible aplicación de fármacos no utilizados actualmente en la clínica para el tratamiento del MM[155, 192].

Otra de las aplicaciones dentro de la farmacología es la de la identificación de una firma de expresión génica especifica en el contexto de un determinado tratamiento en pacientes con MM. De este modo, podemos encontrar estudios con pacientes de nuevo diagnóstico tratados con quimioterapia en altas dosis[193], pacientes tratados con combinaciones de fármacos, como pueden ser bortezomib, talidomida y dexametasona (VTD)[194], pacientes tratados con IMiDs y dexametasona además de autotransplante[195], o pacientes tratados con regímenes basados en bortezomib[196]. La determinación de esta firma génica puede ser de gran interés para el hematólogo con el fin de determinar qué pacientes pueden beneficiarse de unas u otras terapias. Además, este tipo de estudios puede tener una segunda aplicación, como es la predicción de la respuesta de los pacientes a estos tratamientos. Estudios previos han demostrado que la predicción de la respuesta completa en pacientes con MM tratados con terapias basadas en bortezomib y VAD (vincristina, doxorrubicina y dexametasona), utilizando únicamente datos de expresión génica puede tener una capacidad predictora limitada, y probablemente sea necesaria la incorporación al modelo de otros elementos génomicos tales como microARNs o datos procedentes de técnicas de secuenciación masiva (NGS)[197]. Sin embargo, estudios más recientes han revelado el poder predictor de la expresión génica en MM considerando múltiples niveles de respuesta, lo que indicaría la necesidad de que la respuesta debe ser tratada como un factor con múltiples niveles en lugar de considerarla dicotomizada[198]. No obstante, en ambos estudios las series analizadas son muy variables en cuanto a los fármacos aplicados a los pacientes. A la vista de estos resultados, parece necesaria la ampliación de la investigación sobre el poder de predicción de la expresión génica en MM, incluyendo un mayor número de estudios y estratificando adecuadamente los grupos de tratamiento.

## 1.5. RNA-seq en la investigación del mieloma múltiple

Aunque la primera vez que se acuñó el término RNA-seq fue en el año 2008 por Lister y colaboradores[199], los orígenes de la RNA-seq se remontan al año 1975 cuando Sanger y Coulson desarrollaron el método "menos-más", mediante el cual fue posible la secuenciación de fragmentos de ADN de cadena sencilla utilizando una ADN polimerasa de *Escherichia coli*[200]. Ya en 1977, Maxam y Gilbert publicaron un método de secuenciación química, que, mediante el marcaje radioactivo del ADN seguido de cuatro reacciones de digestión química, determinaba la secuencia de la molécula de ADN tras la separación de los fragmentos obtenidos por electroforesis en geles de acrilamida[201]. A finales de ese mismo año, Sanger y colaboradores propusieron una nueva técnica de secuenciación basada en el uso de enzimas ADN polimerasas e inhibidores de la síntesis de la cadena de ADN para lograr una terminación específica de la misma, resolviendo estos fragmentos en geles de poliacrilamida una vez se completase la reacción de secuenciación[202]. En los años siguientes se llevaron a cabo varias mejoras respecto al protocolo original, como por ejemplo el uso de desoxinucleótidos trifosfato (dNTPs), cebadores marcados con fluorescencia y el uso de electroforesis capilar, pero el mayor hito se produjo en 1986 cuando Smith, Hood y colaboradores lograron la automatización del proceso[203]. Esto se tradujo en la aparición en 1987 del secuenciador *ABI 370A DNA sequencer*, que junto con las continuas mejoras que se le fueron aportando, sería el secuenciador dominante durante las dos siguientes décadas hasta la llegada de la tecnología conocida como *Next Generation Sequencing* (NGS) en 2005. La primera aproximación con éxito de NGS fue la secuenciación 454 que es una técnica de pirosecuenciación masiva en paralelo[204]. Esta técnica está basada en la liberación de pirofosfato durante la síntesis *de novo* de una nueva cadena de ADN, lo cual permite medidas en tiempo real de la síntesis de ADN[205]. Al mismo tiempo que esta técnica iba ganando auge, se veía también superada por la secuenciación patentada por la compañía Illumina. Esta nueva técnica de secuenciación combinaba la tecnología de terminación de cadena o método de Sanger, con la inmovilización del molde de secuenciación en una superficie de cristal. En este nuevo sistema propuesto por Illumina, las moléculas de ADN inmovilizadas eran amplificadas mediante amplificación en puente seguida de las síntesis de ADN, usando cuatro terminadores de cadena marcados con fluorescencia[206]. Este método, que empezó secuenciando fragmentos de 25 nucleótidos es la tecnología dominante en el mundo de la secuenciación, puesto que en la actualidad es capaz de secuenciar hasta 300 nucleótidos y generar millones de lecturas independientes en secuenciadores como el *MiSeq* de Illumina.

Además del desarrollo de la tecnología de secuenciación, los estudios de RNA-seq se están enfrentando a otra serie de retos relacionados fundamentalmente con el desarrollo de métodos y técnicas para su análisis. En la última década, se han propuesto cientos de algoritmos y flujos de trabajo o *pipelines* para analizar este tipo de datos, pero existe un gran debate acerca de cuáles de estos métodos obtienen los mejores resultados, por lo que aún es preciso que se ahonde en la investigación comparativa de las diferentes metodologías de análisis de la RNA-seq. Por esta razón, uno de los desafíos a los que se enfrenta actualmente la RNA-seq es la falta de consenso a la hora de realizar el análisis de las secuencias obtenidas.

El análisis de RNA-seq normalmente consta de varios pasos[207]: recortado de las secuencias (*trimming*), alineamiento o mapeo (*alignment*), contaje (*counting*) y normalización (*normalization*) y en muchos casos, análisis de la expresión diferencial (ED). Estos pasos se han representado esquemáticamente en la **Figura 1.26**.

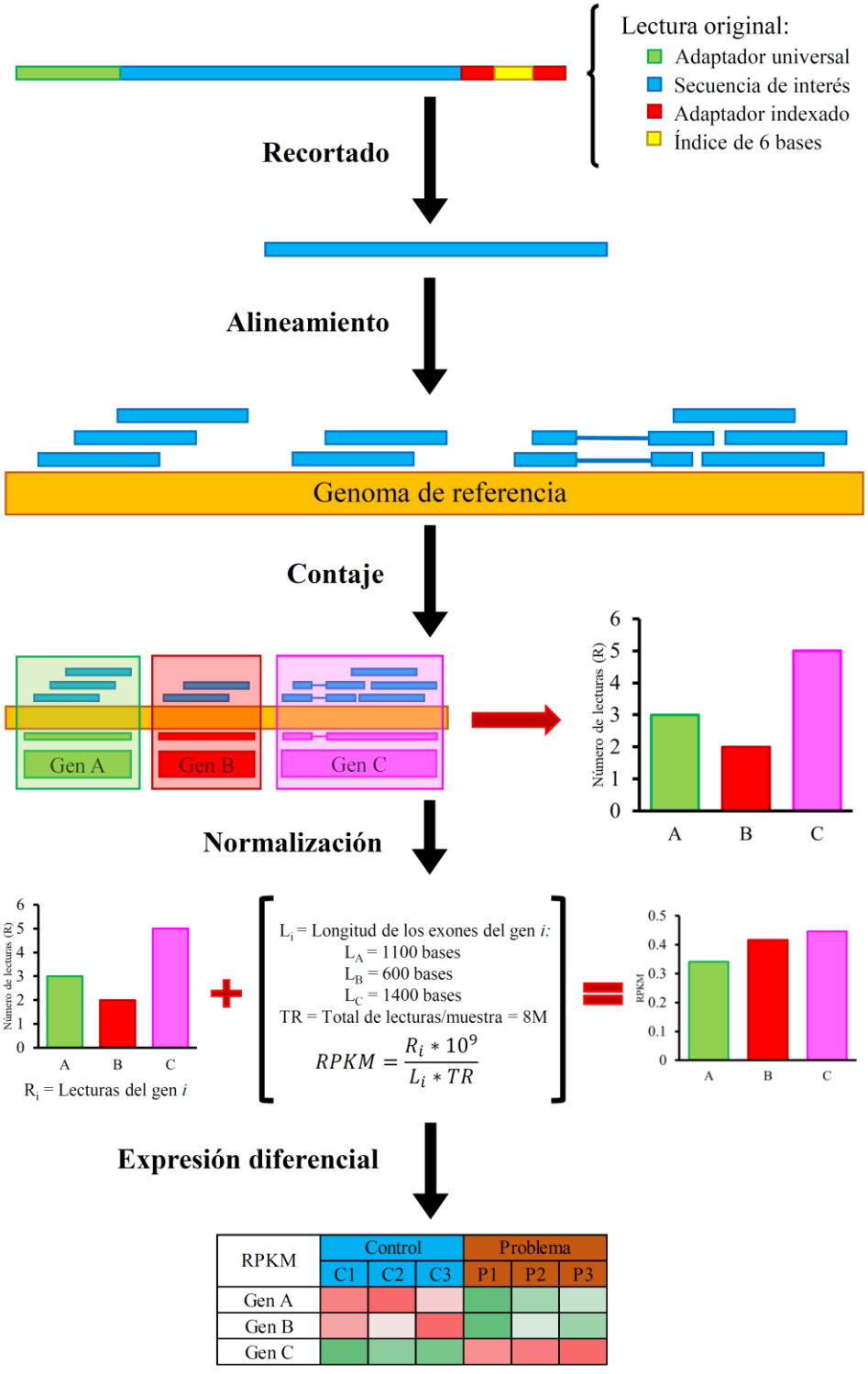



***Figura 1.26.*** *Esquema del flujo de trabajo típico de un experimento de RNA-seq. En el paso de normalización se representa como ejemplo la normalización mediante RPKM (Reads Per Kilobase Million). El cálculo de las RPKM del ejemplo está basado en un experimento realizado con una profundidad de secuenciación de 8 millones de lecturas.*

El recortado de secuencias es una técnica que suele aplicarse sobre las secuencias en bruto de RNA-seq. Consiste en eliminar las secuencias de los adaptadores, que fueron añadidas para llevar a cabo la secuenciación, y de los nucleótidos de mala calidad en las zonas terminales de las lecturas. Este proceso conduciría a la mejora de la calidad de las lecturas resultantes. Sin embargo, esta técnica tiene muchos detractores en el mundo de la bioinformática ya que, si se aplica de manera agresiva, podría sesgar las estimaciones a nivel de expresión génica[208] o de ensamblaje de transcritos[209]. Sin embargo, sus beneficios han sido ampliamente evaluados tanto en estudios de RNA-seq cuyo alineamiento ha sido llevado a cabo contra un genoma de referencia[210, 211], como en estudios de RNA-seq con alineamineto *de novo*[212, 213]. Por tanto, el recortado se secuencias podría ser en muchos casos recomendable con el fin de mejorar la fiabilidad de los resultados finales.

El segundo paso en el análisis de la RNA-seq es el alineamiento o mapeo de las lecturas obtenidas del proceso de recortado. El alineamiento se realiza habitualmente contra un genoma o un transcriptoma de referencia. El objetivo del alineamiento es obtener la localización de la secuencia de cada lectura respecto a la referencia contra la que se está alineando. Las herramientas y algoritmos utilizados en este paso están bajo constante desarrollo, por lo que en la literatura es posible encontrar múltiples opciones para llevar a cabo este proceso. Debido a esta gran variedad de algoritmos y a la relevancia del alineamiento de las lecturas en un estudio de RNA-seq, son múltiples los trabajos en los que se ha procedido a la evaluación y comparación de los algoritmos empleados en este paso[214-220]. Cada uno de los estudios ha evaluado grupos diferentes de algoritmos de alineamiento, en condiciones experimentales diferentes y con una finalidad distinta, como puede ser la cuantificación de la expresión génica cruda o la ED de genes o transcritos en dos condiciones diferentes, como pueden ser tratamiento *vs.* control.

Una vez completado el procedimiento de alineamiento, se genera un archivo SAM, del inglés *Sequence Alignment Map*, que contiene la información del mapeo de las lecturas sobre el genoma o transcriptoma de referencia. Es común que este archivo tenga que ser procesado con otras herramientas como *Samtools*[221] o *Picard* (http://broadinstitute.github.io/picard), con el fin de ordenar las lecturas, eliminar lecturas no alineadas o transformar el archivo a una versión binaria (BAM, del inglés *Binary Alignment Map*). Con estos archivos BAM se procederá a continuación a la asignación de las lecturas a un gen o transcrito determinado, proceso conocido como contaje o cuantificación. El contaje representa un gran desafío en el análisis de la RNA-seq debido a las propiedades intrínsecas de los datos analizados, como la longitud efectiva del gen que se analiza y la longitud de las lecturas con la que se ha realizado la secuenciación. Por este motivo, se han propuesto múltiples aproximaciones para realizar la cuantificación a nivel génico, cada una de ellas produciendo resultados diferentes, entre las que se pueden encontrar las planteadas por los algoritmos *Cufflinks*[222], *eXpress*[223], *HTSeq*[224], *RSEM*[225] o *Stringtie*[226], cuya descripción detallada se recoge en la **Sección de Material y métodos**. Otra de las estrategias que se ha propuesto más recientemente para abordar el contaje de las lecturas ha sido el desarrollo de los algoritmos conocidos como pseudoalineadores. Los pseudoalineadores se caracterizan por integrar, en un mismo algoritmo, los procesos de alineamiento y contaje, y además por proponer una nueva técnica llamada pseudoalineamiento[227], que de modo general, puede definirse como la asignación de lecturas a una secuencia diana sin alineamiento de secuencias al nivel de

nucleótido. Una de las grandes ventajas de los pseudoalineadores, además de la integración de dos de los pasos de análisis, es su menor coste computacional y el ahorro de tiempo de análisis, lo que los convierte en una alternativa muy atractiva a la hora de realizar análisis exploratorios de datos de RNA-seq.

Tras el proceso de contaje, el investigador normalmente se encuentra con una matriz de datos de tamaño *n***m*, donde *n* es el número de genes y *m* es el número de muestras analizadas, que contiene el número de lecturas asignadas a cada uno de los genes del genoma considerado. Sobre esta matriz se lleva a cabo el siguiente paso del análisis de RNA-seq: la normalización. Este paso es de gran relevancia en los análisis de RNA-seq ya que, aunque no existe la posibilidad de la presencia de sesgos debido al ruido de fondo producido por el proceso de hibridación como ocurre en los microarrays[228], la RNA-seq no está exenta de otros tipos de sesgo asociados a la técnica de secuenciación. Los principales sesgos detectados han sido tanto a nivel posicional[229], debido a que los fragmentos secuenciados suelen localizarse con mayor frecuencia al inicio o final del gen, como a nivel específico de secuencia[230, 231], donde la secuencia que rodea el inicio o el final de los potenciales fragmentos secuenciados afecta a la posibilidad de que estos fragmentos sean seleccionados para su secuenciación. El proceso de normalización trataría de solucionar estos problemas propios de la RNA-seq a través del desarrollo de múltiples estrategias, que abarcan desde los métodos de estimación no basados en abundancia, como el método del cuartil superior, a métodos que utilizan factores de escalado en función del tamaño de la librería, como son TMM, *DESeq* o FPKM. La normalización, junto con el proceso de contaje, son pasos esenciales en el establecimiento de la expresión de un gen mediante RNA-seq. Por esta razón se han llevado a cabo en los últimos años multitud de estudios en los que han sido comparadas las diferentes estrategias empleadas en ambos pasos[225, 227, 232-238], con el fin de elucidar la mejor estrategia para establecer el valor de expresión génica en RNA-seq.

La normalización podría ser el paso definitivo en trabajos cuyo objetivo fuese la cuantificación de la expresión génica. Sin embargo, en la mayoría de los casos, el objetivo va más allá con la detección de genes cuya abundancia ha cambiado de manera significativa entre dos o más condiciones experimentales. Estos son los llamados análisis de expresión génica diferencial, que tampoco están exentos de dificultades[239]. El abordaje de este tipo de análisis se ha llevado a cabo de forma tradicional, bien utilizando métodos paramétricos, o siguiendo una segunda alternativa con métodos no paramétricos. Los métodos paramétricos se basan en la asunción de que cada valor de expresión génica cae dentro de una distribución particular, como puede ser la distribución binomial negativa, tal y como consideran los algoritmos *baySeq*[240], *Cuffdiff*[241], *DESeq2*[242], *EBSeq*[243] y *edgeR*[244]; o una distribución log-normal, como asumen los métodos *limma*[245] y *Ballgown*[246]. Por otra parte, los llamados métodos no paramétricos, no asumen ningún tipo de distribución de los datos, de manera que no imponen un modelo rígido al que ajustar estos datos. En el presente trabajo se evaluarán dos de estos métodos no paramétricos: *SAMseq*[247] y *NOIseq*[248].

A pesar de todos los esfuerzos que se han realizado para mejorar el estudio de la expresión génica diferencial, y de los numerosos autores que han propuesto estudios comparativos en los que se trata de determinar los métodos más apropiados para la determinación de la expresión génica diferencial en RNA-seq[235, 248-258], actualmente sigue

sin haber consenso sobre qué aproximación asegura la validez de los resultados en cuanto a robustez, reproducibilidad y exactitud[255], aspectos que constituyen una parte importante en el presente trabajo.

Si bien la determinación de la metodología óptima en cada uno de los pasos anteriores ya es un gran reto en el análisis de la RNA-seq, el mayor desafío es combinar los pasos anteriores para generar un *pipeline* óptimo que permita la determinación robusta tanto de la expresión cruda de los genes de una muestra, como de la expresión génica diferencial entre dos o más condiciones experimentales. Como se expuso anteriormente, el rendimiento de los métodos y algoritmos de análisis en cada paso individual del análisis ha sido evaluado en varios trabajos, sin embargo, solamente un pequeño grupo de autores ha evaluado el rendimiento de los diferentes métodos empleados en cada paso en combinación con el resto que constituyen un *pipeline* completo[253, 259-261]. El estudio del efecto combinado de cada paso de un *pipeline* es un proceso complejo, ya que depende del escenario de análisis de RNA-seq, de manera que la metodología óptima en cada paso del *pipeline* va a depender en gran medida del objetivo y del diseño experimental que el investigador haya planteado[262, 263]. Por esta razón, es necesario realizar una evaluación de diferentes *pipelines* en base a su rendimiento, precisión y exactitud, propiedades que serán abordadas en detalle en la presente investigación.

A pesar de esta complejidad que envuelve el proceso global de análisis, la RNA-seq es una técnica cada vez más popular en el panorama científico, gracias en buena medida a la versatilidad que ofrece, permitiendo tanto el estudio de los niveles de expresión génica como de transcritos, así como estudios de fusión de genes, inserción de genomas víricos o estudios de *splicing* alternativo[264]. Otro de los motivos de la amplia difusión de la RNA-seq ha sido la democratización de su precio, debido en gran medida a la aparición de nuevos métodos de preparación de librerías y también al surgimiento de un gran número de compañías que ofrecen la RNA-seq como servicio.

El campo de la oncología no ha sido ajeno a este *boom* de las técnicas de secuenciación de ARN. Como se observa en la **Figura 1.27a**, el número de publicaciones relacionadas con el estudio del cáncer en las que se ha utilizado RNA-seq se encuentra en pleno ascenso. Esta progresión se debe principalmente a las ventajas que ofrece sobre otras técnicas ya que es capaz de dar una resolución a nivel de nucleótido, tiene muy poco ruido de fondo, requiere una baja cantidad de ARN y ofrece varios niveles de detección, a nivel de gen, transcrito, etc., con un solo análisis. Uno de los estudios que demostró la gran ventaja de la RNA-seq se llevó en cáncer de mama. En este estudio, la RNA-seq logró la detección de más de 2.000 ARN asociados con el riesgo de recurrencia del cáncer de mama[265], mientras que un estudio realizado previamente sobre las mismas muestras utilizando qRT-PCR solamente logró detectar 21 genes[266].

La RNA-seq también ha sido considerada de gran utilidad en el campo de la oncohematología, sobre todo por su capacidad de identificar biomarcadores de subtipos patológicos y de estadios de progresión. Así, por ejemplo, se han descubierto multitud de defectos como mutaciones puntuales, inserciones, deleciones, traslocaciones, omisión de exones y fusiones génicas con quinasas en pacientes con leucemia linfoblástica aguda de células T[267]. Dentro de este campo, de manera particular en el MM, la investigación utilizando RNA-seq está empezando a despegar, tal y como se puede deducir de la **Figura**

**1.27b**. Así, hasta este momento, se ha logrado con la RNA-seq la identificación de moléculas capaces de modular la vía de MAPK[268], la identificación de nuevos agentes potencialmente terapeúticos[169] y la determinación de una firma genética capaz de distinguir qué pacientes va a responder mejor a una terapia concreta[269]. En cualquier caso, el futuro de la investigación del MM con RNA-seq es prometedor, ya que unido al reciente desarrollo de técnicas como la *single cell RNA-seq* o la edición de genomas con CRISPR-Cas9, podría llevar al descubrimiento de procesos hasta ahora desconocidos en la biología de esta patología.

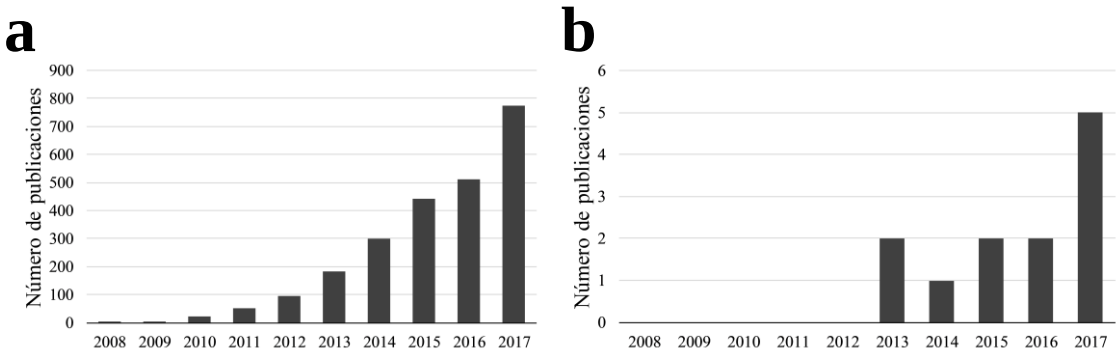



***Figura 1.27.*** *Número de publicaciones por año en Pubmed para estudios de expresión génica con RNA-seq en* ***a)*** *cáncer y* ***b)*** *mieloma múltiple. Las búsquedas se llevaron a cabo a través de la página web http://dan.corlan.net/medline-trend.html, utilizando como términos de búsqueda en* ***a)*** *"RNA-seq AND cancer" y en el caso de* ***b)*** *"RNA-seq AND myeloma".*

## 1.6. Análisis comparativo de microarrays y RNA-seq

El microarray ha sido la técnica de análisis elegida en la gran mayoría de estudios de expresión génica de alto rendimiento desde finales de la década de los 90 hasta la primera década del siglo XXI. Esta predominancia ha sido debida a la capacidad del microarray frente a otras técnicas como la PCR de proporcionar al investigador, en un solo análisis, datos de expresión de miles de genes y transcritos al mismo tiempo. Sin embargo, a pesar del éxito cosechado por esta técnica en ámbitos como la investigación de biomarcadores en cáncer[270-272] o el estudio de los cambios de expresión génica frente a fármacos[273], los microarrays presentan algunas limitaciones que son difícilmente resolubles. Una de estas limitaciones se debe a su naturaleza estática, ya que solo es posible determinar la expresión de genes o transcritos que cuentan con sondas en la plataforma. Esto supone una gran limitación sobre estudios de organismos o patologías en los que continuamente se están descubriendo nuevos genes o transcritos, ya que el microarray, por su propio diseño, al ser una matriz rígida con unas sondas predefinidas no puede ser actualizado y terminaría siendo una estructura obsoleta, siendo la única solución a este problema el diseño de un nuevo chip de análisis. Otra de las limitaciones del microarray aparece en el proceso de hibridación, ya que en ocasiones se observan procesos de hibridación no específica de ARN, como la hibridación cruzada de secuencias de genes con secuencias similares, y de saturación de la señal emitida por las sondas, lo que puede afectar a la detección de la señal de genes con baja y alta expresión[274]. Por

estos motivos, la RNA-seq se ha visto desde principios de esta década como una alternativa a los análisis transcriptómicos con microarrays[264]. Sin embargo, el uso de la RNA-seq también tiene sus correspondientes desventajas. Los principales obstáculos que se encuentra actualmente el investigador que realiza RNA-seq son debidos a la metodología empleada en la preparación de librerías. De este modo, en uno de los primeros pasos como es el proceso de fragmentación del ARN, las técnicas de digestión enzimática, sonicación o el uso de cationes divalentes para fragmentar el ARN pueden incluir sesgos a nivel posicional (los fragmentos se sitúan preferentemente al inicio o al final del transcrito) o específicos de secuencia (la secuencia que rodea el inicio o el final de los fragmentos potenciales afecta a la probabilidad de ser seleccionados para su secuenciación) que afecten a la estimación de la expresión génica[275]. Además, también se han detectado sesgos en los pasos de transcripción inversa y de amplificación mediante PCR[230, 264]. No obstante, la RNA-seq es un método que se encuentra en la actualidad en desarrollo, con lo que muchos de los problemas que han surgido durante sus primeros años de vida se han ido solucionando de manera satisfactoria. De esta manera, las dificultades asociadas a que la RNA-seq puede ser costosa computacionalmente y además requerir mucho tiempo de procesamiento se están empezando a resolver a través de las nuevas implementaciones algorítmicas. Otra de las limitaciones propias de la RNA-seq es el efecto que la elevada abundancia de algunos ARN, como pueden ser los ARNr, puede tener sobre la población global de ARNs provocando su dilución en el conjunto global de ARNs estudiados. Este artefacto se ha corregido en los últimos años mediante el desarrollo de técnicas para la selección de colas de ARN poliadeniladas[276].

Estos hechos expuestos han llevado a la comunidad científica a plantearse estudios comparativos en los que se evalúe de una manera precisa las ventajas e inconvenientes de ambas tecnologías en diferentes escenarios experimentales. Hasta la fecha, se han publicado varios trabajos en los que se ha tratado de aclarar si los resultados provistos por RNA-seq o el microarray son comparables o si, por el contrario, alguna de las dos tecnologías se muestra superior a la otra. Una lista de estos trabajos comparativos, junto con sus principales características, se recoge en la **Tabla 1.1**.

***Tabla1.1.*** *Estudios comparativos entre RNA-seq y microarray publicados hasta la fecha.*

| Referencia | Organismo | Procedencia | Plataforma de microarray | Plataforma de RNA-seq |
|---|---|---|---|---|
| Grabowiecka et al., 2018[277-] | *Escherichia coli* | Cepa O157:H7 | Agilent #G4813A-020097 | Illumina Hiseq2000 |
| Wolff et al., 2018[278] | *Homo sapiens* | Lineas celulares de linfoma de Burkitt y pacientes de cáncer de recto | Affymetrix Human Gene 1.0 ST | Illumina HiSeq2000, single-end |
| Keck et al., 2018[279] | *Homo sapiens* | Tumores neuroendocrinos intestino delgado | Affymetrix Human Transcriptome Array 2.0 | Illumina HiSeq4000, paired-end, 75 pb |
| Romero et al., 2018[280] | *Homo sapiens* | Lineas celulares de cáncer de mama | Affymetrix Human Transcriptome Array 2.0 | Illumina HiSeq2000, paired-end, 100 pb |
| Chen et al., 2017[281] | *Homo sapiens* | Carcinoma pulmonar de células escamosas | Affymetrix HG-U133A y Agilent G4502A y Affymetrix Human Exon 1.0 ST | Illumina HiSeq2000 |
| Nazarov et al., 2017[282] | *Homo sapiens* | Carcinoma pulmonar de células escamosas | Affymetrix Human Transcriptome Array 2.0 | Illumina HiSeq2000 |
| Dapas et al., 2017[283] | *Homo sapiens* | Glioblastoma multiforme y Carcinoma pulmonar de células escamosas | Affymetrix Human Exon 1.0 ST | Illumina HiSeq2000 |
| Li et al., 2016[284] | *Homo sapiens* | Psoriasis | Agilent Whole Human Genome Oligo Microarray one-color | Illumina, single-end, 85 bp |
| Zhang et al., 2016[285] | *Homo sapiens* | Tejido sano normal | Varias plataformas | Varias plataformas |
| Yu et al., 2015[286] | *Homo sapiens* | Tejido sano normal | Agilent Whole Human Genome Microarray 4x44K v1 y Illumina HumanHT-12 v4 Expression BeadChip y Affymetrix Human Gene 1.0 ST y Affymetrix Human Transcript Array 2.0 | Illumina HiSeq2500, single-end, 50 pb |
| Zhang et al., 2015[287] | *Homo sapiens* | Neuroblastoma | Agilent 4x44k oligonucleotide microarrays | Illumina HiSeq2000, paired-end, 90 y 100 pb |
| Robinson et al., 2015[288] | *Saccharomyces cerevisiae* | Cepa haploide DBY 12000 FY | Agilent Yeast Expression 8 × 15K arrays | Illumina HiSeq2500, 141 pb |
| Nault et al., 2015[289] | *Mus musculus* | Hígado | Agilent 4×44 K microarray | Illumina HiSeq2500, single-end, 50 pb |
| Zhang et al., 2014[290] | *Mus musculus* | Neuroesferas | Affymetrix GeneChip Mouse Gene 1.0 ST arrays | Illumina Hiseq2000, paired-end, 101 pb |
| Fumagalli et al., 2014[291] | *Homo sapiens* | Cáncer de mama | Affymetrix HG-U133 Plus 2.0 | Illumina HiSeq2000, paired-end, 50 pb |
| Zhao et al., 2014[292] | *Homo sapiens* | Cáncer de mama | Agilent whole genome microarrays | Illumina HiSeq2000, paired-end, 48 pb |
| Perkins et al., 2014[293] | *Rattus norvegicus* | Nervio espinal | Affymetrix Rat Exon 1.0 ST arrays | Illumina GAIIx, paired-end, 34 pb |
| Zhao et al., 2014[294] | *Homo sapiens* | Células T de memoria CCR6+ CD4 | Affymetrix GeneChip HT HG-U133+ PM arrays | Illumina HiSeq2000, paired-end, 90 pb |
| Black et al., 2014[295] | *Rattus norvegicus* | Hígado | Affymetrix HT Rat230+PM microarrays | SOLiD 5500xl, single-end, 50 pb |
| Zwemer et al., 2014[296] | *Homo sapiens* | Fluido amniótico | Affymetrix Human Genome U133 Plus 2.0 Array | Illumina Hiseq2000, paired-end, 50 pb |

***Tabla1.1 (continuación).*** *Estudios comparativos entre RNA-seq y microarray publicados hasta la fecha.*

| Referencia | Organismo | Procedencia | Plataforma de microarray | Plataforma de RNA-seq |
|---|---|---|---|---|
| Xu et al., 2013[297-] | *Homo sapiens* y similación | Lineas celulares de cáncer de colon | Affymetrix HG-U133 Plus 2.0 | Illumina Hiseq2000, paired-end, 100 pb |
| Sekhon et al., 2013[298] | *Zea mays* | Distintas partes de la planta | NimbleGen microarray | Illumina HiSeq2000, single-end, de 35 a 101 pb |
| Mooney et al., 2013[299] | *Canis familiaris* | Linfoma de células B | Affymetrix GeneChip Canine Genome V2.0 Array | Illumina HiSeq2000, paired-end, 100 pb |
| Giorgi et al., 2013[300] | *Arabidopsis thaliana* | Distintas partes de la planta | Affymetrix ATH1 | Illumina, varios protocolos |
| Raghavachari et al., 2012[301] | *Homo sapiens* | Anemia de células falciformes | Affymetrix Human Exon 1.0 ST | Illumina, paired-end |
| Kogenaru et al., 2012[302] | *Xanthomonas citri* subsp. citri | Cepas mutantes *hrpX* | Agilent 8-by-15-K DNA microarray chips | Illumina, single-end, 74 bp |
| Sirbu et al., 2012[303] | *Drosophila melanogaster* | Embrión | Affymetrix single-channel microarrays y Microarray no comercial de doble canal | Roche 454 GS FLX, single-end |
| van Delft et al., 2012[304] | *Homo sapiens* | Carcinoma hepatocelular | Affymetrix HG-U133 Plus 2.0 | Illumina, paired-end |
| Bottomly et al., 2011[305] | *Mus musculus* | Cuerpo estriado | Affymetrix MOE 430 2.0 array /Illumina MouseRef-8 v2.0 array | Illumina, single-end, 43 pb |
| Toung et al., 2011[306] | *Homo sapiens* | Lineas celulares inmorataizadas de células B | Affymetrix Human HG-Focus Target Array | Illumina HiSeq200, paired-end, 50 pb |
| Su et al., 2011[307] | *Rattus norvegicus* | Riñón | Affymetrix Rat Genome 230 2.0 Array | Illumina GA II, single-end, 36 pb |
| Malone et al., 2011[308] | *Drosophila melanogaster* | Cabeza | Nimblegen custom array | Ilumina GA I |
| Liu et al., 2011[309] | *Homo sapiens*, *Pan troglodytes* y *Macaca mulatta* | Cerebelo | Affymetrix GeneChip Human HJAY array | Illumina GA II, single-end, 36-bp |
| Bradford et al., 2010[310] | *Homo sapiens* | Lineas celulares de cáncer de mama | Affymetrix Human Exon 1.0ST arrays | SOLiD v3, 50 pb |
| Griffith et al., 2010[311] | *Homo sapienes* | Linea celular de cáncer de colon | Affymetrix Human Exon 1.0 ST y custom NimbleGen microarray | Illumina GA II, paired-end, 42 pb |
| Bullard et al., 2010[312] | *Homo sapiens* | Cerebro | Affymetrix HG-U133 Plus 2.0 | Illumina GA I, single-end, 35 pb |
| Agarwal et al., 2010[313] | *Caenorhabditis elegans* | Segundo estado larval | Affymetrix C. elegans Tiling 1.0R Array | Illumina GA I |
| Fu et al., 2009[314] | *Homo sapiens* | Cerebro | Affymetrix Human Exon 1.0 ST | Illumina, single-end, 36 pb |
| Bloom et al., 2009[315] | *Saccharomyces cerevisiae* | Híbrido diploide de dos cepas | Agilent two color arrays | Illumina GA I |
| Marioni et al., 2008[228,] | *Homo sapiens* | Hígado y riñón | Affymetrix HG-U133 Plus 2.0 | Illumina, 32 pb |

La mayor parte de los trabajos recogidos en la **Tabla 1.1** trata de comparar la reproducibilidad y la repetibilidad de la cuantificación de la expresión génica, así como determinar el rendimiento en la detección de genes diferencialmente expresados de ambas técnicas. En los primeros estudios comparativos llevados a cabo en 2008[228], se pudo observar que, a grandes rasgos, los resultados obtenidos en la determinación de la expresión génica y la expresión génica diferencial utilizando RNA-seq eran comparables a los obtenidos mediante microarrays, incluso se describió cierta ventaja de la detección de RNA-seq cuando se contrastaron los resultados de las dos tecnologías con qRT-PCR. Sin embargo, debido a las limitaciones de la técnica, propias de una tecnología que en aquel año aún estaba comenzado a desarrollarse y a que el flujo de trabajo bioinformático aún no estaba puesto a punto, serían necesarios nuevos estudios comparativos a medida que la técnica fuese evolucionando. Desde la publicación de este primer trabajo se han realizado numerosos análisis comparativos en los que se ha evaluado la capacidad de cuantificación de la expresión génica y el nivel de detección de la ED de las dos tecnologías, pero además, también se han valorado otros aspectos como la determinación de ontologías génicas[285], el rango dínámico[286] o la detección de eventos de splicing alternativo[280], por citar algunos ejemplos. En estos trabajos, recogidos en la **Tabla 1.1**, se encontraron de manera general mejores prestaciones de la RNA-seq frente a los microarrays, entre las que se destacan el buen rango dinámico de esta técnica[286], así como el alto poder de detección de genes diferencialmente expresados[284, 297, 304, 309] y la superioridad en la detección cuantitativa de la expresión génica y transcriptómica[287-289, 292-294, 299, 313, 314]. Sin embargo, también son numerosos los trabajos en los que se detecta una alta concordancia entre las dos tecnologías, incluso en condiciones en las que el microarray ha estado tradicionalmente por debajo de la RNA-seq, como puede ser la detección del splicing alternativo[280] o la evaluación de cambios de expresión en isoformas[283]. Por este motivo, no son pocos los autores que abogan, si las circunstancias lo permiten, por realizar los estudios de expresión génica con ambas tecnologías, ya que presentan un alto grado de complementariedad y ortogonalidad, proponiendo la necesidad de su coexistencia, evolución y mejora[279, 285, 301, 302].

## 1.7. Uso de líneas celulares en la farmacoterapia del mieloma múltiple

Una línea celular es un cultivo celular permanente que prolifera de manera indefinida en un medio de cultivo y un espacio apropiados[316]. La primera línea celular fue desarrollada en 1912 por Alexis Carrel y sus colaboradores a partir de explantos de corazón del embrión de pollo (*Gallus gallus*)[317]. Sin embargo, el cultivo celular propuesto por Carrel careció de éxito al ser de difícil mantenimiento, no tener replicabilidad y casi carecer de aplicabilidad[318]. Estos primeros estudios, junto con el descubrimiento de los antibióticos, que permitirían la comercialización de medios de cultivo estériles, supusieron el sembrado de la semilla para el desarrollo de lo que conocemos hoy como líneas celulares. De esta manera, la primera línea celular permanentemente estable, conocida como línea celular "L", fue establecida en 1943 a partir de fibroblastos de ratón

(*Mus musculus*)[319]. En los años siguientes se produjo una continua propagación de los cultivos celulares, hasta que, en 1951, George Gey y sus colaboradores generaron la primera línea celular humana, derivada de células de cáncer cervical de la señora Henrietta Lacks, en cuyo honor se bautizó esta línea celular como HeLa[320]. Desde el establecimiento de la línea HeLa, las líneas celulares inmortalizadas se han utilizado intensamente como modelos biológicos en la investigación científica. Actualmente, las líneas celulares constituyen una alternativa experimental a las células primarias obtenidas directamente del organismo objeto de estudio. Esto se debe principalmente a que aportan ciertas ventajas sobre el uso de células primarias, como son su coste económico, su facilidad de uso y la capacidad de proveer al investigador de una fuente ilimitada de material de investigación, sin los problemas éticos que conlleva el uso de animales de experimentación o tejidos humanos[321]. Sin embargo, los ensayos experimentales realizados en líneas celulares no están exentos de dificultades, ya que se ha podido comprobar a través del examen de líneas celulares de cáncer, que en muchos casos, no son representativas de la patología de la que proceden debido a la heterogeneidad que presenta el cáncer en los pacientes[322]. Otra limitación de las líneas celulares es que están desprovistas del medio ambiente local del organismo en el que se estudia la patología, lo que impide estudiar las interacciones que pueden tener las células patológicas con otras células o componentes en el organismo de interés[321]. A pesar de estos obstáculos, resulta evidente que las líneas celulares, utilizadas en unas condiciones adecuadas[323], han permitido, permiten y permitirán a los investigadores adquirir un amplio conocimiento sobre la biología del cáncer. Su uso como modelos preclínicos en diversos tipos de cáncer ha posibilitado la predicción de la respuesta celular a fármacos, así como la comprensión de los mecanismos de acción de los fármacos en la célula tumoral, sin la necesidad de emplear muestras clínicas, muchas veces costosas de conseguir[324].

El estudio del MM también se ha beneficiado ampliamente del uso de las líneas celulares. Desde el establecimiento de la línea celular RPMI-8226 en 1966 por Matsuoka y colaboradores[325], se han publicado multitud de trabajos basados en el uso de HMCLs dedicados a desentrañar los efectos de distintos compuestos farmacológicos o a revelar mecanismos esenciales en la biología del MM.

En lo que respecta al campo farmacológico, las HMCLs han sido de especial relevancia en los ensayos preclínicos de fármacos, de manera que han jugado un papel esencial en el descubrimiento de los efectos sobre las células mielomatosas de compuestos como bortezomib[326], los IMiDs[327] o panobinostat[328]. Además, el uso de las HMCLs también ha sido clave en el descubrimiento y comprensión de los mecanismos de la resistencia celular a drogas, gracias al establecimiento de líneas celulares resistentes a determinados fármacos a partir de líneas celulares parentales sensibles. Esto ha sido especialmente útil en el caso de los fármacos aprobados para el tratamiento del MM como el bortezomib, para el que se han descrito mecanismos de resistencia asociados al gen *PSMB5*[329], o la lenalidomida, donde se ha desvelado el papel que juega la disminución de la expresión del gen *CRBN* en el mecanismo de resistencia a este compuesto[330]. Las HMCLs también han sido muy útiles para comprender la biología del MM. De esta manera, gracias al establecimiento de distintos modelos de líneas celulares que representan los diferentes patrones moleculares y estructurales de las células tumorales del MM, se ha conseguido revelar la diferencia de expresión génica que explica la transformación maligna de la célula plasmática[272]. Con todo esto, y a pesar de que, como

se indicaba anteriormiente, los estudios con HMCLs tienen serias limitaciones a la hora de ser extrapolados al campo clínico[331], las HMCLs siguen siendo la primera vía de abordaje en cualquier trabajo de investigación en el campo del MM.

## 1.8. Metaanálisis y mieloma múltiple

El metaanálisis ha sido descrito como una técnica de análisis esencial para el progreso de la ciencia[332], gracias a su capacidad de encontrar una respuesta global a un problema dado a través de la síntesis de los resultados de todos los estudios realizados hasta una fecha determinada en un campo concreto del conocimiento. El metaanálisis siempre ha ido de la mano del término "revisión sistemática", sin embargo, ambos términos no deben ser confundidos, ya que el metaanálisis es la parte cuantitativa de una revisión sistemática. De esta manera es posible encontrar revisiones sistemáticas sin metaanálisis, pero no lo contrario, ya que carecería de sentido. Así, cuando una revisión sistemática viene seguida de la aplicación de metodología estadística para combinar dos o más estudios, se conoce como revisión sistemática cuantitativa o revisión sistemática con metaanálisis[333]. La revisión sistemática con metaanálisis es un proceso que consta de múltiples pasos, cada uno de los cuales es esencial para poder alcanzar una conclusión de forma robusta[334]. El primero de los pasos es establecer la pregunta concreta que se desea responder. Esta pregunta deberá ser lo más corta y descriptiva posible. La siguiente etapa consiste en fijar cómo se va a medir el resultado de la intervención, utilizando para ello medidas del efecto como pueden ser el "*odds ratio"* y el "riesgo relativo", en el caso de variables dicotómicas, o la "diferencia de medias" y el "ratio de medias log-transformado", en el caso de variables continuas. A continuación, hay que desarrollar una estrategia de búsqueda exhaustiva de trabajos publicados en bases de datos, repositorios, bibliotecas online, etc. Para ello, habrá que definir unos términos de búsqueda apropiados en función de las características del estudio que se esté realizando. En la búsqueda en bases de datos es de vital importancia el uso apropiado de los *operadores booleanos AND, OR* y *NOT*. En una siguiente etapa hay que definir una serie de criterios de inclusión y de exclusión para poder identificar todos los ensayos que sean relevantes en el área de la hipótesis planteada, así como elegir un método de medida de la calidad de los estudios. Uno de los métodos comunmente empleados para la revisión de la calidad en trabajos recogidos en la base de datos Cochrane (https://www.cochranelibrary.com/), es la escala de Jadad[335]. La escala de Jadad determina un sistema de puntuación, que a modo de cuestionario con 7 preguntas, evalúa la bondad de cada uno de los ensayos clínicos candidatos para el metaanálisis. El siguiente paso consiste en la extracción de los datos de los estudios seleccionados, dependiendo este paso del tipo de información que se quiera recopilar en función de la pregunta que se haya planteado en el primer paso. Tras la extracción de los datos, se ha de proceder con el análisis de la heterogeneidad de los estudios seleccionados utilizando las pruebas estadísticas Q e $I^2$. En función del resultado de las pruebas de heterogeneidad se decidirá qué método estadístico se utilizará para combinar los resultados de los estudios elegidos. De esta manera, el investigador puede decidir entre aplicar un método de "efecto fijo", que solamente considera la variabilidad intraestudios, o un método de "efecto aleatorio", en el que se valora tanto la variabilidad

intraestudios como la variabilidad interestudios. En ambos casos el objetivo es la combinación ponderada de los resultados de los estudios seleccionados, dando más peso estadístico a los estudios más precisos. Para realizar estos cálculos se dispone de múltiples herramientas y paquetes estadísticos como *metafor*[336] en el software R, el paquete estadístico SIMFIT[337], o el programa CMA[338]. El último paso de la revision sistemática con metaanálisis es la interpretación clínica de los resultados y su publicación en revistas científicas[339].

El desarrollo de todo este proceso ha requerido de tiempo y una gran investigación en términos estadísticos y metodológicos. De este modo, el origen del metaanálisis se remonta al siglo XVIII, cuando astrónomos y matemáticos como Gauss y Laplace, abordaron las ideas propuestas años antes por Blaise Pascal sobre el desarrollo de técnicas que permitiesen lidiar de forma cuantitativa con datos procedentes de diferentes observaciones[340]. Sin embargo, no sería hasta el siglo XX cuando Karl Pearson, finalmente, aplicaría estas ideas al campo de los ensayos clínicos, para comparar la infección y la mortalidad en soldados que voluntariamente habían sido inoculados contra la fiebre tifoidea[341]. Desde este momento, la evolución del metaanálisis ha sido constante, introduciéndose diferentes metodologías y procedimientos estadísticos para batallar contra algunos de los problemas propios de este tipo de análisis. Así, podemos encontrar la introducción de los modelos de efectos fijos y de efectos aleatorios en 1954[342], el desarrollo de un método para el cálculo de la varianza entre estudios en 1986[343], el establecimiento de metodologías para evaluar el sesgo de publicación en 1997[344] y la proposición de un nuevo índice de medida de la heterogeneidad mediante el estadístico $I^2$ en 2002[345]. Aunque el metaanálisis sigue presentando algunas limitaciones, como puede ser la presencia de heterogeneidad entre los trabajos sujetos a estudio o la dificultad de solucionar los sesgos de publicación[346], las mejoras que se han desarrollado en estos años han conseguido que el metaanálisis sea una técnica capaz de ofrecer información fiable basada en la evidencia experimental. Esto tiene su reflejo en el gran incremento de publicaciones en las que se han realizado trabajos de metaanálisis en general, que se viene produciendo en los últimos años (**Figura 1.28**).

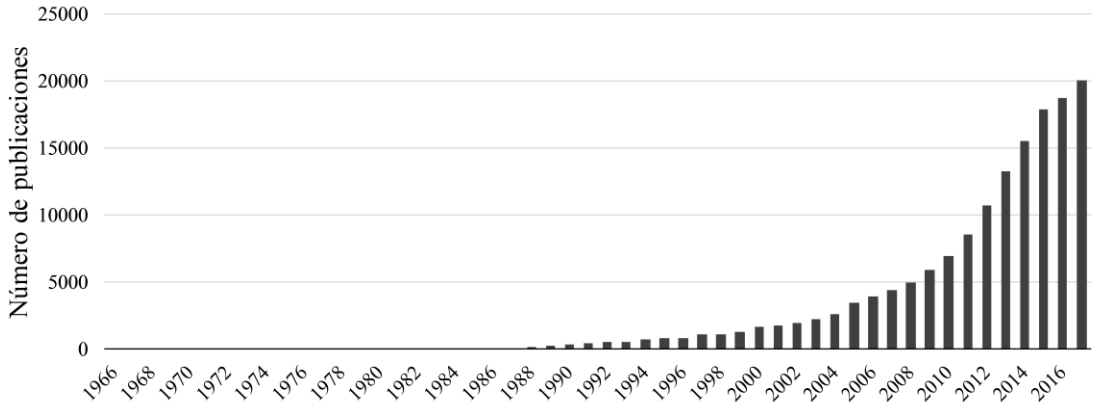



***Figura 1.28.*** *Número de publicaciones por año en Pubmed para trabajos de metaanálisis. Las búsquedas se llevaron a cabo a través de la página web [http://dan.corlan.net/medline-trend.html](http://dan.corlan.net/medline-trend.html), utilizando como término de búsqueda la palabra "meta-analysis".*

Esta explosión de publicaciones también se ha registrado de manera análoga en el campo de la investigación del cáncer. Así, en el periodo comprendido entre 2008 y 2013 se registró un incremento de cinco veces el número de publicaciones anuales que utilizaron metaanálisis en temas relacionados con la oncología, la mayor parte focalizadas en el estudio del riesgo de padecer cáncer y en el estudio genético de la enfermedad[347]. De manera particular, esta tendencia alcista también se ha producido en el estudio del MM (**Figura 1.29**). Uno de los principales objetivos del metaanálisis en MM ha sido la evaluación de la eficacia de distintos regímenes de tratamiento sobre parámetros como la supervivencia global (SG) o la supervivencia libre de progresión (SLP). No obstante, esta no ha sido la única aplicación del metaanálisis en el MM, ya que también se han publicado numerosos estudios en los que se ha evaluado la eficacia de la vía de administración del fármaco, la eficacia de los tratamientos en subgrupos específicos de MM, y la toxicidad y la seguridad de los diferentes tratamientos empleados. Sin embargo, un tipo de metaanálisis que aún no ha tenido gran desarrollo es el realizado a partir de datos de expresión génica de microarrays en MM con la finalidad de combinar diferentes estudios publicados para alcanzar una síntesis cercana a la evidencia. A pesar de que el uso de este tipo de estudios ha tenido gran éxito en otros desórdenes oncogénicos[348], en el caso del MM solamente se ha realizado un trabajo de metaanálisis con microarrays utilizando muestras de sangre periférica[349]. Los estudios con metaanálisis en líneas celulares con el fin de determinar la firma de expresión génica de un determinado fármaco tampoco ha sido una vía de investigación explotada en el MM. Todos estos hechos abren una prometedora vía de investigación para la posible detección tanto de genes de respuesta a diferentes fármacos, así como para el descubrimiento de nuevos biomarcadores de respuesta a tratamiento. Este tipo de investigación ha formado parte de los objetivos del presente trabajo.

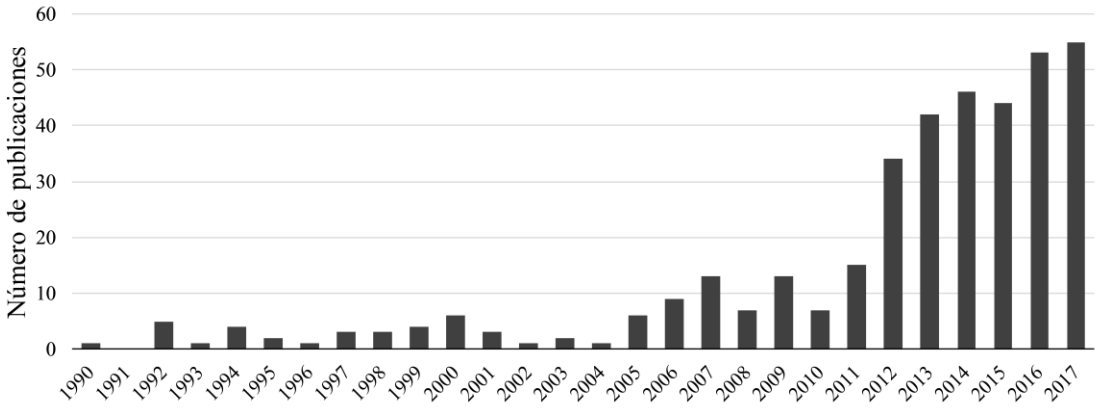



***Figura 1.29.*** *Número de publicaciones por año en Pubmed para trabajos de metaanálisis en mieloma múltiple. Las búsquedas se llevaron a cabo a través de la página web http://dan.corlan.net/medline-trend.html, utilizando como término de búsqueda las palabras "meta-analysis AND mieloma".*

## 1.9. Modelos predictivos de respuesta al tratamiento en pacientes con mieloma múltiple

La palabra predecir, del latín *praedicĕre*, significa, según el diccionario de la Real Academia Española (RAE), "anunciar por revelación, conocimiento fundado, intuición o conjetura algo que ha de suceder". Por tanto, predecir consiste en conocer, con el conocimiento disponible *a priori*, un evento que sucederá en el futuro. En el campo de la medicina clínica la predicción de la evolución de la enfermedad de un paciente o de la respuesta que este va a tener frente a un determinado tratamiento siempre ha sido un objetivo de suma importancia. Uno de los pasos críticos en el proceso de predicción es el descubrimiento de potenciales biomarcadores que, medidos en el momento del diagnóstico, determinen si el paciente va a responder a un tratamiento y cómo va a progresar su enfermedad. Este es el primer paso hacia el desarrollo de terapias personalizadas que eviten a los pacientes tratamientos potencialmente dañinos que no supongan ningún beneficio para la cura de su afección[350], además de suponer para el sistema sanitario un ahorro de costes innecesarios.

En el campo del tratamiento del cáncer, el descubrimiento de nuevos biomarcadores de respuesta a tratamiento siempre ha sido un gran reto, ya que permitiría predecir el desarrollo de la enfermedad y detectar su evolución en un estadío temprano, a la vez que podría guiar a los profesionales sanitarios a la hora de decidir la terapia adecuada para cada paciente; además podría ayudar en la investigación del cáncer identificando nuevas dianas para el desarrollo de nuevos fármacos[351]. Los biomarcadores tumorales tienen una naturaleza muy diversa, pudiendo ser hormonas, enzimas, glicoproteínas o receptores celulares, e incluso alteraciones que se producen en la célula tumoral como las mutaciones genéticas, amplificaciones, delecciones, traslocaciones o cambios en la expresión génica. Debido a esto, existen multitud de estrategias empleadas para la identificación de marcadores, muchas de ellas basadas en ensayos proteómicos, como son la electroforesis en gel bidimensional, el ELISA o la citometría de flujo[352], y otras basadas en tecnologías de alto rendimiento, como la secuenciación masiva de ARN y ADN, la espectroscopía de masas o los microarrays[353]. Estas últimas se han convertido en una herramienta atractiva y de uso muy extendido debido a la gran cantidad de potenciales biomarcadores que se pueden analizar de manera simultánea. Como un caso particular, las tecnologías de alto rendimiento de análisis de la expresión génica han tenido una especial relevancia en la búsqueda de biomarcadores en cáncer. Esto se debe a la importancia que las variaciones en los perfiles de expresión génica pueden tener en la progresión del cáncer[354], en el diagnóstico de los pacientes y para la predicción clínica de la evolución de la enfermedad[355]. Por estos motivos, los biomarcadores de expresión génica se han empleado en diversos estudios predictivos en diferentes patologías como el cáncer de mama[355-357], el cancer de próstata[358] o el cancer colorectal[359], por poner algunos ejemplos.

En MM también se han llevado a cabo múltiples trabajos predictivos utilizando datos de expresión génica procedentes de tecnologías de alto rendimiento. Una de las primeras aproximaciones investigó[360] la predicción de los tipos de cadena ligera y pesada de Igs, a partir de datos de expresión génica de microarrays de Affymetrix, obteniendo en todos los casos resultados satisfactorios. La expresión génica medida mediante

microarrays también ha servido para la predicción de subgrupos citogenéticos en MM[361]. Por último, como ya se adelantó en el **Apartado 1.4**, otra de las aplicaciones de la expresión génica determinada con microarrays en MM, ha sido la predicción de la respuesta a distintos regímenes de tratamiento[197, 198]. Sin embargo, en este último caso, debido a la contraposición entre estos dos trabajos, no existen evidencias concluyentes de la capacidad predictora de los datos de expresión génica sobre la respuesta al tratamiento.

En todos estos trabajos, la capacidad de predicción de un determinado acontecimiento viene también determinada por la elección de los algoritmos predictivos. En los últimos años, los métodos de predicción no paramétricos, como los algoritmos basados en aprendizaje de máquinas o *machine learning*, han ganado gran relevancia en el campo de la genómica y son vistos como una herramienta potencial en medicina. Estos métodos entrenan un modelo clasificador de acuerdo con las características de biomarcadores previamente definidos en una serie de muestras de entrenamiento, para finalmente validar el modelo predictivo en una nueva serie de muestras de validación. La necesidad en la investigación científica de estos modelos ha llevado al desarrollo de decenas de métodos estadísticos, entre los que destacan, por su popularidad[362] los algoritmos de modelos lineales generalizados (GLM, del inglés *Generalized Linear Models*), las máquinas de soporte de vectores (SVM, del inglés *Support Vector Machines*)[363], los K-vecinos más cercanos (KNN, del inglés *K-Nearest Neighbours*)[364] y los bosques aleatorios (RF, del inglés *Random Forests*)[365]. Otro de los algoritmos que ha obtenido muy buenos resultados en predicción de la expresión génica medida con microarrays ha sido los mínimos cuadrados parciales (PLS, del inglés *Partial Least Squares*)[366]. Los estudios predictivos suelen llevarse a cabo utilizando una combinación de varios de estos algoritmos, comparando finalmente su eficacia en la predicción. El porcentaje o tasa de aciertos va a depender de múltiples factores, como la cantidad de variables predictoras de que se dispone, la colinelalidad existente entre las variables y también del objetivo del predictor, dependiendo este de si se valora más la velocidad o la exactitud en la predicción[367]. Por tanto, otro punto clave en cualquier estudio de predición será la comparación de diferentes algoritmos, con el fin de determinar cuál es el que mejor se ajusta a unas condiciones experimentales concretas.

Por todo lo expuesto, una finalidad del presente trabajo ha sido la investigación de distintos algoritmos preditivos con pacientes de MM, a pesar de la dificultad que presentan la heterogeneidad genética de los pacientes y de los tratamientos aplicados, así como los diferentes tipos de respuesta al tratamiento alcancadas por los pacientes. Fruto de este estudio será determinar la bondad de un conjunto de potenciales biomarcadores en la predicción de la respuesta a un determinado régimen de tratamiento, así como determinar aquellos algoritmos que llegan a un mayor porcentaje de acierto en las predicciones.

# 2. Objetivos

**1.** Encontrar un flujo de trabajo o *pipeline* óptimo para el análisis de RNA-seq con datos reales procedentes de líneas celulares de mieloma múltiple, tanto a nivel de expresión génica cruda como a nivel de expresión génica diferencial.

**2.** Realizar un estudio comparativo entre las dos técnicas de análisis masivo de datos de expresión génica más utilizadas en la actualidad, RNA-seq y microarrays.

**3.** Determinar mediante estudios de metaanálisis los perfiles de expresión génica del mieloma múltiple asociados a los fármacos más utilizados en su tratamiento, analizando los posibles mecanismos de acción.

**4.** Determinar mediante técnicas de metaanálisis los perfiles de expresión génica de pacientes con mieloma múltiple asociados a la respuesta a diferentes regímenes de tratamientos.

**5.** Predecir la respuesta a diferentes regímenes de tratamiento del mieloma múltiple utilizando datos de expresión génica.

# 3. Material y métodos

## 3.1. Muestras y fármacos utilizados

### 3.1.1. Líneas celulares de mieloma múltiple: KMS12-BM y JJN-3

Para el estudio de RNA-seq y su comparación con microarrays se utilizaron dos líneas celulares bien establecidas de mieloma, KMS12-BM procedente de un MM con traslocación t(11;14), y JJN-3 proveniente de una leucemia de células plasmáticas con traslocación t(14;16). Estas líneas celulares fueron adquiridas desde el *Deutsche Sammlung von Mikroorganismen and Zellkulturen GmbH* (DSMZ). El cultivo de ambas líneas celulares se realizó en medio RPMI1640 suplementado con 10% de FBS y antibióticos (Gibco). De manera rutinaria se llevaron a cabo chequeos en busca de la presencia de micoplasma utilizando el kit MycoAlert (Lonza). La identidad de las líneas celulares se confirmó también de manera periódica mediante análisis STR con el kit PowerPlex 16 HS de Promega.

### 3.1.2. Fármacos: amilorida y TG003

Ambas líneas celulares fueron sometidas a tratamiento con dos compuestos potencialmente moduladores del *splicing* alternativo como son la amilorida y el TG003. Ambos compuestos fueron adquiridos a Sigma-Aldrich. En el caso de la amilorida, la concentración aplicada fue de 0,1 mM en las dos líneas celulares durante 24 horas. La concentración aplicada del compuesto TG003 fue de 0,4 mM en ambas líneas celulares, también durante 24 horas. Como controles se trataron ambas líneas celulares con dimetilsulfóxido (DMSO). En todos los casos se realizaron tres réplicas biológicas. Como resultado final se obtuvieron seis tripletes (18 muestras) procedentes de las seis combinaciones entre las líneas celulares y los compuestos farmacológicos empleados en el estudio. Un esquema de la preparación de las muestras mencionadas aparece recogido en la **Figura 3.1**.

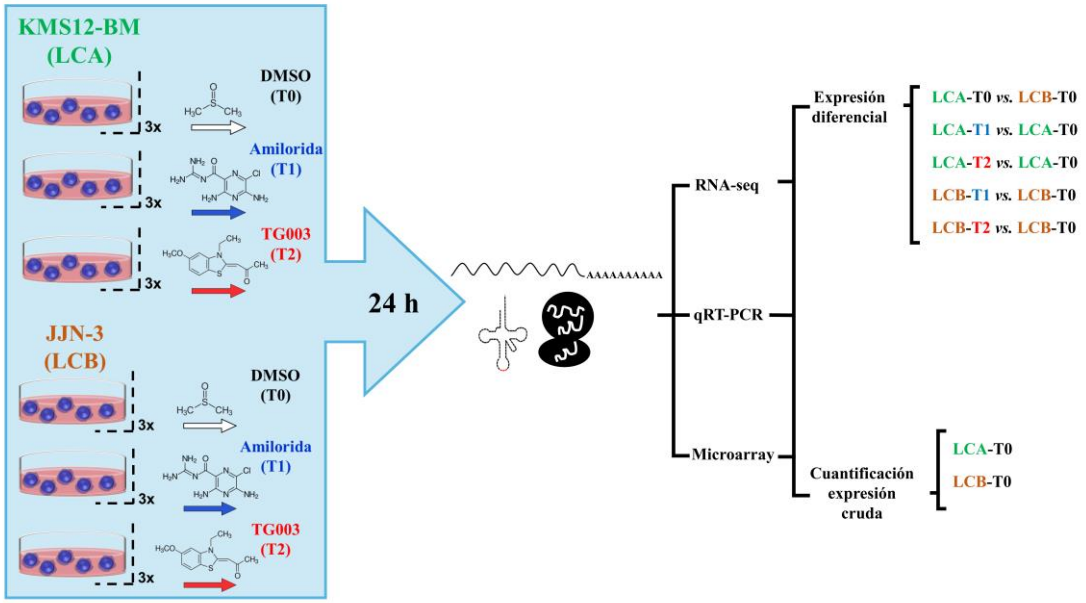



***Figura 3.1.*** *Procedimiento experimental del estudio de RNA-seq y microarray. Se utilizaron dos líneas celulares de mieloma múltiple (KMS12-BM [LCA] y JJN-3 [LCB]) y dos compuestos farmacológicos (amilorida [T1] y TG003 [T2]) para llevar a cabo los estudios del presente trabajo. Las muestras control, tratadas con DMSO (T0) fueron utilizadas en los estudios de cuantificación de la expresión génica cruda, mientras que el total de 18 muestras fue utilizado para los estudios de expresión génica diferencial.*

### 3.1.3. Extracción y secuenciación del ARN

El ARN poliadenilado de las 18 muestras fue extraído utilizando el kit RNeasy Plus Mini de Qiagen y su integridad fue medida con el Agilent 2100 Bioanalyzer, alcanzando todas las muestras los estándares mínimos (concentración > 200 ng/µL y valores de RIN > 8,0). El ARN total fue posteriormente convertido por transcripción inversa a ADNc utilizando el kit High-Capacity cDNA Reverse Transcription de Applied Biosystems. A continuación, se procedió a la construcción de las librerías siguiendo el protocolo TruSeq Stranded mRNA Sample Preparation Guide de Illumina. Las librerías resultantes fueron normalizadas y secuenciadas usando un secuenciador HiSeq™ 2500 en el laboratorio Lifesequencing S.L. (Valencia, España). La secuenciación se hizo mediante lecturas de extremos pareados de 101 pares de bases (pb), con un rango de 36.240.231-77.906.369 lecturas pareadas.

## 3.2. Metodologías en el análisis de RNA-seq

El análisis de los datos de RNA-seq se realizó partiendo de los archivos FASTQ obtenidos directamente del secuenciador. Todos los FASTQ fueron sometidos a control de calidad utilizando el programa FASTQC (v0.11.3)[368], donde se determinaron los siguientes parámetros de calidad:

a) Calidad de la secuencia por base: mediante un diagrama de cajas se da una visión general de la calidad de cada base en cada posición de la secuencia considerando todas las lecturas del archivo FASTQ.
b) Valores de calidad por secuencia: determina el número de secuencias a un nivel determinado de calidad (Phred Score)
c) Contenido de la secuencia por base: determina el porcentaje de lecturas que tiene cada uno de los cuatro nucleótidos (adenina, timina, guanina o citosina) en cada posición de las 101 pb que tienen las secuencias analizadas.
d) Contenido de guanina-citosina (GC) por secuencia: muestra el contenido de GC a lo largo de toda la secuencia en un FASTQ y lo compara contra una distribución normal que ajusta el algoritmo.
e) Contenido de nucleótidos no asignados (Ns) por base: determina el número de nucleótidos que no ha podido ser asignado a uno de los cuatro nucleótidos del ADN en cada base de la lectura.
f) Distribución de la longitud de las secuencias: muestra el número de lecturas en función de su longitud.
g) Secuencias duplicadas: número de lecturas que se encuentran más de una vez en el archivo FASTQ. La presencia de estas secuencias duplicadas puede ser debida a una sobreamplificación en la etapa de PCR al preparar las librerías.
h) K-meros sobrerrepresentados: se determina el número de k-meros que aparecen con más frecuencia en el archivo FASTQ.
i) Calidad por "baldosas": muestra si alguna región específica en el secuenciador ("baldosa" o *tile*) está enriquecida en secuencias de mala calidad.

Un ejemplo de estas representaciones de calidad de los archivos FASTQ para la muestra LCA-T0-M1, se muestran en la **Figura 3.2**.

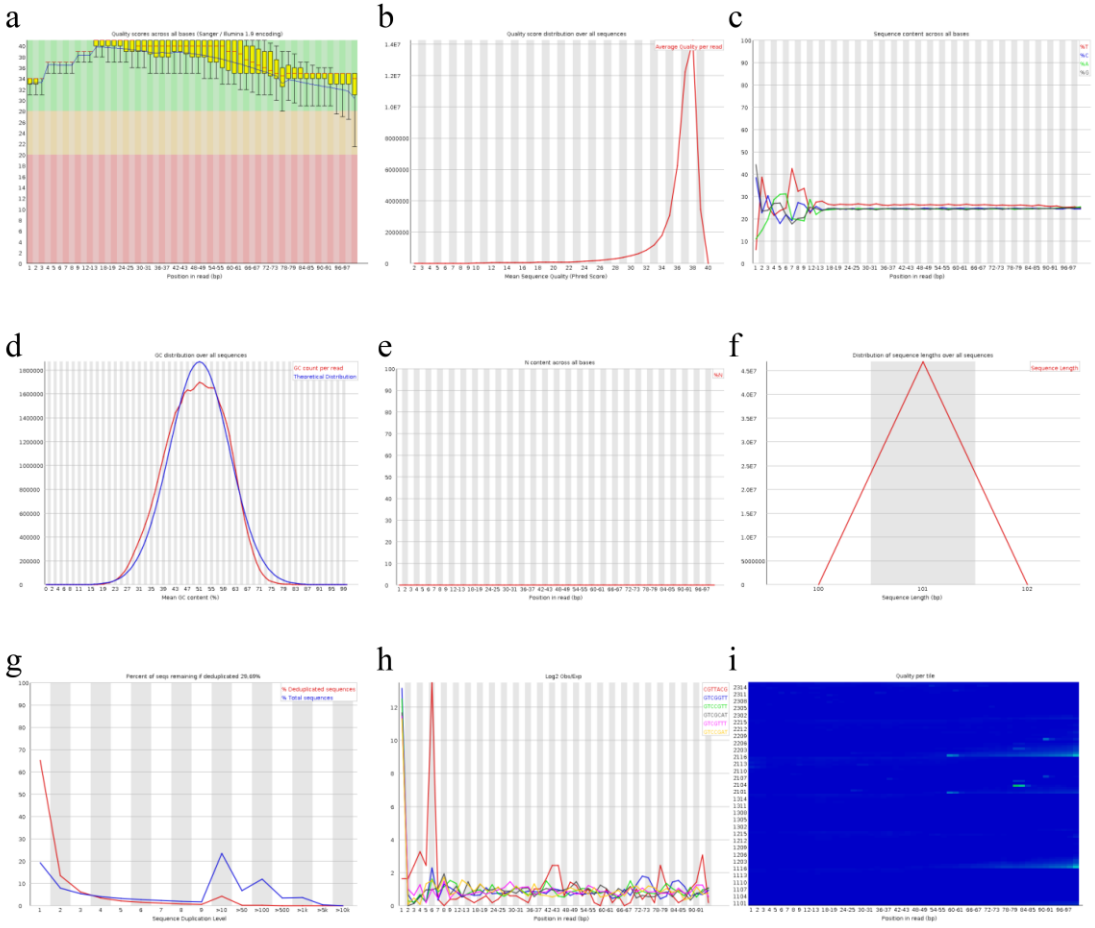



***Figura 3.2.*** *Resultado del análisis de control de calidad utilizando el algoritmo FASTQC sobre las lecturas del canal r1 de la muestra 141048 (línea celular KMS12-BM [LCA] tratada con DMSO [T0]). Las letras de los paneles siguen la nomenclatura expuesta en el texto arriba comentado.*

### 3.2.1. Recortado de las lecturas (*Trimming*)

Tras la determinación de la calidad de las secuencias se procedió a la aplicación de cada *pipeline*, comenzando por el proceso de recortado de las lecturas o *trimming*. Este procedimiento es un paso beneficioso en experimentos de RNA-seq[211], aunque debe aplicarse de una forma no agresiva y en conjunción con una selección de longitud mínima de lectura adecuada para evitar cambios impredecibles en la expresión génica[209]. Los algoritmos utilizados en este proceso fueron:

a) *Trimmomatic* (v0.35)[369].
b) *Cutadapt* (v1.12)[370].
c) *BBDuk* (v.Oct.,23,2015), incluido en la suite *BBTools*[371].

En todos los algoritmos, se realizó el recorte de las secuencias de los adaptadores de Illumina sobre cada una de las lecturas, así como la eliminación de nucleótidos cuya calidad no tuviese una puntuación de Phred q > 20. El tamaño mínimo de lectura

permitido fue de 51 pb. El resto de los parámetros aplicados en cada uno de los algoritmos se detallan en los archivos de órdenes o *scripts* adjuntos en el **Anexo 1**. El número de lecturas supervivientes arrojado por cada uno de los algoritmos se extrajo del informe de resultados provisto por cada algoritmo de recortado. Por su parte, las ratios de alineamiento fueron obtenidas tras el proceso de alineamiento utilizando la herramienta *samtools* (v1.3.1)[221].

Los análisis estadísticos realizados sobre los porcentajes de lecturas supervivientes, las ratios de alineamiento, y sobre su influencia en la determinación de la expresión génica cruda se llevaron a cabo utilizando la prueba de Kruskal-Wallis seguida del test *post-hoc* de Dunn con el paquete *dunn.test* (v.1.3.5)[372] en R (v.3.5.1)[373]. Se consideraron como estadísticamente significativas todas las comparaciones cuyo FDR < 0,05.

### 3.2.2. Alineamiento o mapeo (*alignment* o *mapping*)

Una vez aplicado el procedimiento de recortado se procedió al alineamiento de las lecturas contra el genoma y/o el transcriptoma humano. Las lecturas alineadas de manera satisfactoria contra la referencia son lo que se conoce como alineamientos o lecturas mapeadas. Para llevar a cabo el mapeo de las lecturas se utilizaron cinco algoritmos. Estos algoritmos pueden ser divididos en tres categorías en función de la referencia de alineamiento que utilizan:

1) Alineamiento contra el genoma de referencia: se usó como referencia el genoma humano versión GRCh37 (hg19) de Ensembl[374]:
    a) *Tophat2* (v2.1.0)[214]. Está basado en el programa de mapeo *Bowtie2*. En un primer paso alinea las lecturas procedentes del experimento de RNA-seq contra el genoma de referencia y después analiza los resultados del mapeo para identificar las uniones entre los exones. Este algoritmo se aplicó seleccionando como tipo de librería *fr-firststrand*, cuyo significado es que la lectura "1" procede de la cadena opuesta al transcrito y la lectura "2" procede de la cadena del transcrito (**Figura 2.3**), y determinando para cada muestra la distancia promedio interna entre los pares de lectura directa y reversa.
    b) *STAR* (v2.5.3a)[215]. El mapeo con *STAR* consiste en dos pasos: un primer paso de creación del índice del genoma y un segundo paso de mapeo de las lecturas contra el genoma de referencia.
    c) *Hisat2* (v2.0.0)[375]. Está basado en una extensión de la transformación de Burrows-Wheeler (BWT) llamada GCSA. *Hisat2* usa un índice global que representa la población global y además un gran número de pequeños índices que representan regiones más pequeñas del genoma que lo cubrirían completamente. *Hisat2* se ejecutó determinando como tipo de librería la opción RF que es análoga a la función *fr-firststrand* de *TopHat2*.

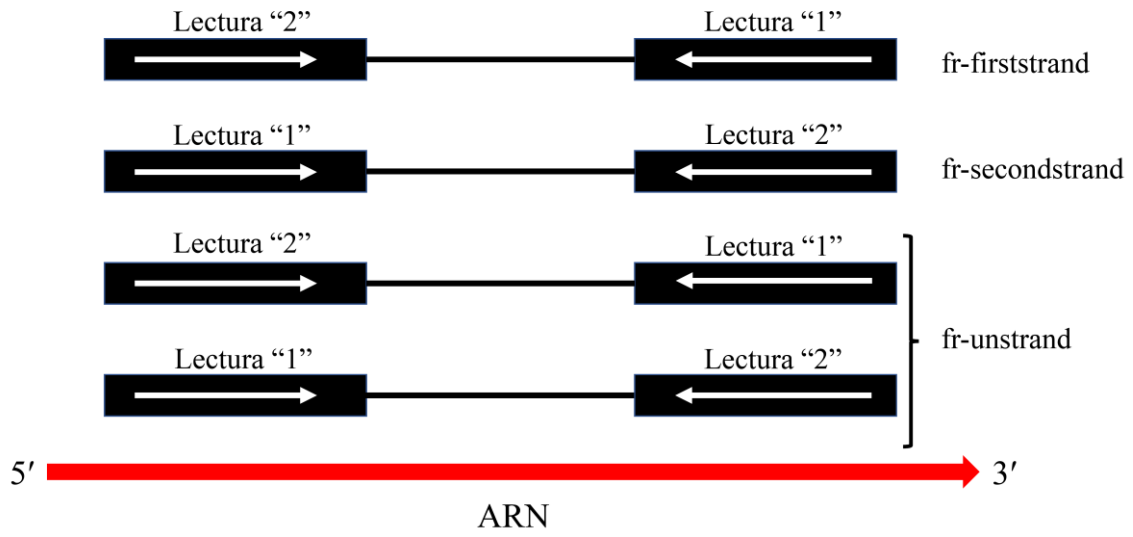



***Figura 3.3.*** *Tipos de librerías en TopHat2, adaptado del blog "One Tip Per Day" de Xianjun Dong (http://onetipperday.sterding.com/). "fr" significa forward-reverse y se refiere al sentido de la lectura de la cadena de ADN de referencia.*

2) Alineamiento híbrido contra el genoma y el transcriptoma:
    a) *RUM* (v2.0.5_06)[217]. Realiza el proceso de mapeo en tres pasos: primero alinea las lecturas contra el genoma utilizando *Bowtie*, seguido de un mapeo a transcriptoma también con *Bowtie* y por último vuelve a mapear contra el genoma usando el algoritmo de alineamiento *Blat*. *RUM* además requiere un paso de preprocesamiento de las lecturas mediante adicción de Ns a las lecturas con tamaño < 101 pb, ya que solamente admite FASTQ con los dos pares de lecturas del mismo tamaño. Además, para hacer concordar los archivos BAM (*Binary Alignment Map*) resultantes con las librerías de Ensembl fue necesario editar los encabezados de estos archivos.
3) Alineamiento contra el transcriptoma:
    a) *Bowtie2* (v.2.2.6)[376]. Está basado en BWT. Este algoritmo fue ejecutado indicando como tipo de librería la opción *--fr*, que considera como válidos los alineamientos en los que la lectura "1" aparece aguas arriba de la secuencia reversa complementaria de la lectura "2" o viceversa.
    b) *STAR*. Se ejecutó de manera similar a la versión genómica indicando como modo de cuantificación la cuantificación transcriptómica utilizando la opción *TranscriptomeSAM*.

En todos los casos, los alineamientos contenidos en los archivos BAM resultantes se ordenaron por posición y por nombre utilizando *samtools* para adaptarlos a los requisitos de los distintos algoritmos de contaje que se utilizaron posteriormente. Además, se procedió a un filtrado de los alineamientos, seleccionando solamente los alineamientos mapeados, de forma que los no mapeados fueron eliminados del archivo BAM.

Los contajes de las lecturas en función del tipo de alineamiento se llevaron a cabo sobre los archivos BAM utilizando *samtools*. Por su parte, los análisis estadísticos realizados para determinar las diferencias entre los algoritmos de alineamiento en la determinación de la expresión génica cruda se llevaron a cabo utilizando la prueba de Kruskal-Wallis seguida del test *post-hoc* de Dunn con el paquete *dunn.test* en R. Se consideraron como estadísticamente significativas todas las comparaciones cuyo FDR < 0,05.

### 3.2.3. Contaje y normalización

El proceso de contaje de lecturas mapeadas a nivel génico se llevó a cabo utilizando como referencia el archivo GTF de Ensembl versión 82. Se utilizaron seis métodos de contaje:

a) *Cufflinks* (v2.2.1)[222]. Estima las abundancias relativas de los transcritos basándose en cuántos alineamientos apoyan cada uno de estos transcritos, teniendo en cuenta los sesgos en la preparación de librerías.
b) *eXpress* (v1.5.1)[223]. Se basa en el algoritmo de esperanza-maximización (EM). La estimación de la abundancia génica se computa de manera continua mientras los alineamientos obtenidos son analizados.
c) *HTseq* (v0.6.1p1)[224]. Dado un archivo BAM con las lecturas mapeadas y un archivo GTF con los genes de referencia, *HTseq* cuenta para cada gen el número de lecturas alineadas que se ajustan a cada uno de sus exones. Para hacer este contaje se utilizaron sus variantes *Intersection-Strict* y *Union*, cuyas diferencias están gráficamente representada en la **Figura 3.4**.

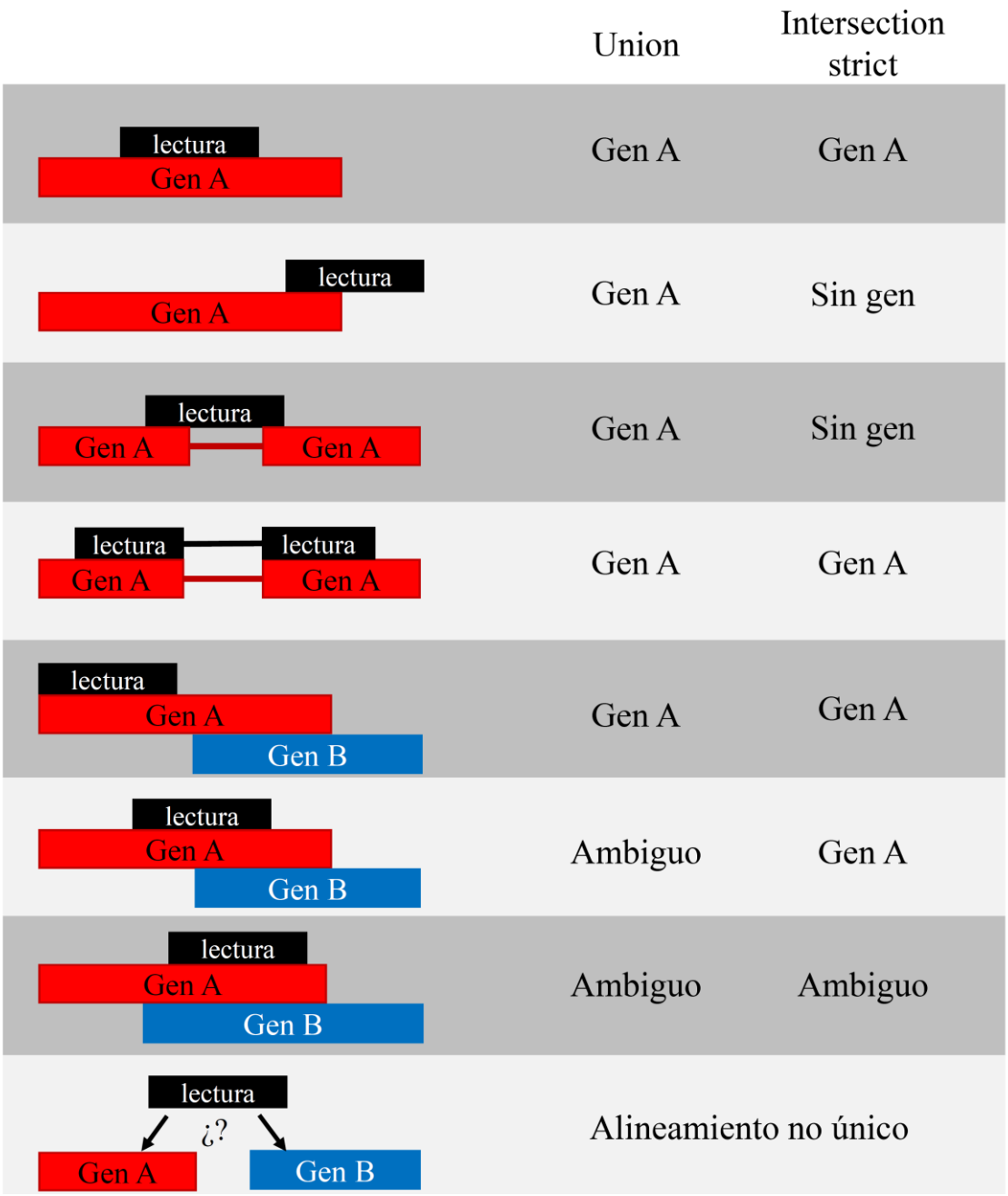



***Figura 3.4.*** *Esquema de la asignación por el algoritmo HTseq de las lecturas mapeadas a genes por los métodos de contaje Union e Intersection-Strict. Adaptado del manual de HTseq (https://htseq.readthedocs.io/en/master/count.html)*

d) *RSEM* (v1.2.31)[226]. Se trata de una herramienta para cuantificar las abundancias de transcritos procedentes de experimentos de RNA-seq, que utiliza un modelo generativo de lecturas de RNA-seq para tratar de que la probabilidad de seleccionar una lectura de un transcrito determinado sea igual a la abundancia de dicho transcrito, y estima a su vez la abundancia génica basándose en el algoritmo EM.

e) *Stringtie*.(v1.3.3b)[225]. Este método utiliza un algoritmo de red de flujo para alinear y cuantificar la abundancia de transcritos.

Los valores de expresión génica fueron cuantificados según los métodos de normalización disponibles asociados a cada algoritmo de contaje en sus respectivos paquetes a nivel génico. Los métodos empleados en este trabajo fueron: lecturas en crudo, FPKM, TPM, TMM (*edgeR*), RLE (*DESeq*), método del cuartil superior (UQ), cobertura o *coverage* (Cov), cuentas estimadas (Est_Counts) y cuentas efectivas (Eff_Counts)[234, 237, 377]. En resumen, estos métodos consisten en lo siguiente:

a) Lecturas en crudo (*Raw reads*). Este método se puede utilizar en los algoritmos *HTseq Intersection-Strict* y *HTseq Union*. Consiste en la determinación del número de lecturas totales mapeadas en cada gen.
b) FPKM (Fragmentos por kilobase de tránscrito por millón de lecturas mapeadas o *Fragments per Kilobase of transcript per Million mapped reads*). Esta técnica de normalización está disponible en los paquetes *Cufflinks*, *Stringtie*, *RSEM* y *eXpress*. Consiste en el cálculo de la expresión génica normalizada por la longitud del gen y el número de lecturas en cada muestra[378]:

$$FPKM_i = \frac{X_i}{(\frac{\tilde{l}}{10^3})(\frac{N}{10^6})} = \frac{X_i}{\tilde{l}_i\, N} * \ 10^9$$

donde *Xi* se trata del número de lecturas observadas sobre un gen de interés *i*, $\tilde{l}_i$ es su longitud efectiva o *Effective length*, definida por la ecuación:

$$\tilde{l}_i = \ l_i - \mu_{FLD} + 1$$

siendo $\mu_{FLD}$ la media de la distribución de la longitud del fragmento obtenida de la lectura alineada, $l_i$ la longitud del gen y *N* el número total de lecturas secuenciadas.

c) TPM (transcritos por millón o *Transcripts per Million*). Implementado en los paquetes *eXpress*, *Stringtie* y *RSEM*. Se trata del número de transcritos mapeados para cada gen en particular normalizado por la longitud de este gen y por la profundidad de la secuenciación en la muestra (expresada en millones de lecturas)[379]:

$$TPM_i = \frac{K_i}{\tilde{l}_i} * \left( \frac{1}{\Sigma_j \frac{K_j}{\tilde{l}_j}} \right) * 10^6$$

donde *Ki* se trata del número de lecturas observadas en un gen de interés *i*, $\tilde{l}_i$ es la longitud efectiva o *Effective length*, definida por la ecuación:

$$\tilde{l}_i = \ l_i - \mu_{FLD} + 1$$

donde $\mu_{FLD}$ es la media de la distribución de la longitud de fragmento obtenida de la lectura alineada, $l_i$ es la longitud del gen y *N* es el número total de lecturas secuenciadas.

d) TMM (media recortada de M-valores o *Trimmed Mean of M-values*)[380]. Este método está implementado en el paquete *edgeR*, se basa en la hipótesis de que la mayor parte de los genes no están diferencialmente expresados. Para cada muestra, TMM se computa como la media ponderada de los *log-ratios* entre la muestra y la referencia de alineamiento, tras la exclusión de los genes más expresados y de los genes con los mayores *log ratios*.

$$log_2\left(d_j^{TMM}\right) = \frac{\sum_{g\in G'} \omega_{gj} M_{gj}}{\sum_{g\in G'} \omega_{gj}}$$

donde:

$$M_{gj} = log_2((K_{gj}/N_j)/(K_{gr}/N_r))$$

$$\omega_{gj} = \left(N_j - K_{gj}\right)/N_j K_{gj} + \left(N_r - K_{gr}\right)/N_r K_{gr}$$

y $K_{gj}$y $K_{gr}$ son > 0. $N_j$ y $N_r$ son el número total de lecturas para la muestra *j* y la muestra referencia *r*, respectivamente. $K_{gj}$y $K_{gr}$ denotan las lecturas para el gen $g$ y las muestras *j* y *r*, respectivamente. *G'* representa el conjunto de genes no recortados $M_g$ y $A_g$ (niveles de expresión absolutos).

e) RLE (expresión relativa en escala logarítmica o *Relative Log Expression*)[381]. Implementado en el paquete *DESeq* y *DESeq2*. También basado en la hipótesis de que la mayoría de los genes no están diferencialmente expresados. En este caso los factores de escalado se calculan para cada muestra como la mediana de la ratio, para cada gen, de sus lecturas respecto a su media geométrica considerando todas las muestras[382].

$$d_j^{RLE} = median_g \frac{K_{gj}}{\left(\prod_{v=1}^{m} K_{gv}\right)^{1/m}}$$

Donde asumiendo *G* genes y *m* muestras, $d_j$ es el factor de escalado para la muestra *j* y $K_{gj}$ denota las lecturas para el gen *g* y la muestra *j*.

f) UQ (cuartil superior o *upper quartile*). Para llevar a cabo este método, el factor de escalado se calcula utilizando las lecturas del cuartil superior (percentil 75) de cada librería tras eliminar los genes con cero lecturas en todas las librerías.

$$d_j^{UQ} = UQ\left(\frac{K_{gj}}{\sum_{g=1}^{G} K_{gj}}\right)$$

Donde asumiendo *G* genes y *m* muestras, $d_j$ es el factor de escalado para la muestra *j*, $K_{gj}$ denota las lecturas para el gen *g* y la muestra *j*, y UQ es el cuartil superior de la muestra *j* con lecturas normalizadas y $K_{gj} > 0$.

g) Cov (cobertura o *coverage*). Implementado en *Stringtie*. Es la cobertura promedio por base del transcrito o exón.
h) Est_Counts (lecturas estimadas o *estimated counts*). Implementado en *eXpress*. Es el número estimado de fragmentos generados en el experimento mediante el algoritmo de esperanza-maximización (algoritmo EM).
i) Eff_Counts. (lecturas efectivas o *effective counts*). Es uno de los métodos de normalización de *eXpress*. Se trata del mismo concepto que las cuentas estimadas pero ajustadas por el sesgo experimental.

$$effCounts_i = X_i * \frac{l_i}{\tilde{l}_i}$$

Donde $X_i$ es el número de cuentas observadas en un gen de interés *i*, $\tilde{l}_i$ es la longitud efectiva o e*ffective length*, definida por la ecuación: $\tilde{l}_i = l_i - \mu_{FLD} + 1$, siendo $\mu_{FLD}$ la media de la distribución de la longitud de fragmento estimada de la lectura alineada y $l_i$ es la longitud del gen.

Todos los análisis estadísticos realizados para la comparación de los métodos de contaje y normalización a la hora de determinar la influencia de los métodos sobre la cuantificación de la expresión génica cruda se llevaron a cabo utilizando la prueba de Kruskal-Wallis seguida de la prueba *post-hoc* de Dunn con el paquete *dunn.test* en R. Se consideraron como estadísticamente significativas todas las comparaciones cuyo FDR < 0,05.

### 3.2.4. Pseudoalineamiento

Los algoritmos de pseudoalineamiento o *pseudoaligners* omiten el alineamiento tradicional que llevan a cabo los métodos de alineamiento y realizan el proceso de contaje de lecturas de una forma más rápida que los métodos tradicionales. Se utilizaron para el contaje de lecturas tres programas de pseudoalineamiento:

a) *Kallisto* (v0.43.1)[227]. Esencialmente, los pseudoalineamientos definen una relación entre una lectura y un conjunto de transcritos compatibles, relación que es computada basándose en mapear los k-meros en rutas en un gráfico de transcritos de De Bruijn. El gráfico se construye con los k-meros presentes en un transcriptoma inicial, en lugar de ser construidos con las lecturas como ocurre en los alineamientos a genoma o transcriptoma. Estos pseudoalineamientos dan más información que el conjunto de k-meros individuales ya que estos se mantienen acoplados en la lectura cuando esta lectura es asignada a un determinado

transcrito. *Kallisto* permite la normalización a través de los métodos TPM y Est_counts.

b) *Sailfish* (v0.9.2)[238]. Cuantifica la abundancia utilizando un sistema de indexado basado en la detección de k-meros. Además, cuantifica la abundancia de transcrito mediante el algoritmo EM. *Sailfish* permite la normalización de lecturas a través de TPM, y también proporciona lecturas en crudo.

c) *Salmon* (v 0.8.2)[383]. Uno de los conceptos claves en este algoritmo es el alineamiento *lightweight*, que es un alineamiento parcial de las lecturas, en el que algunas, pero no necesariamente todas las bases de una lectura son alineadas a bases específicas de las secuencias diana. Además, el índice que utiliza *Salmon* no es un índice de k-meros como los anteriores algoritmos, sino que utiliza una matriz de sufijos junto con un índice basado en BWT. En el caso de *Salmon* se utilizaron 2 técnicas de indexado: el modo basado en FMD[384] y el modo basado en quasi-mapeo (*Quasi-mapping-based mode* [QMB]). Además, dado que *Salmon* no funcionaba correctamente cuando se le suministró con el tipo de librería correspondiente al presente trabajo, se lanzó tanto en el modo de detección automática de librería, como en la opción en el que el usuario provee la librería (en este caso ISR: I = *Inward*, S = *Stranded*, R = la lectura 1 proviene de la cadera reversa). *Salmon*, al igual que *Sailfish*, permite la normalización de las lecturas por el método TPM y la obtención de las lecturas en crudo.

La comparación de los algoritmos de pseudoalineamiento a la hora de determinar la expresión génica cruda se llevaron a cabo utilizando la prueba de Kruskal-Wallis seguida de la prueba *post-hoc* de Dunn con el paquete *dunn.test* en R. Se consideraron como estadísticamente significativas todas las comparaciones cuyo FDR < 0,05.

### 3.2.5. Expresión génica diferencial

***Filtrado previo de los datos***

Para el análisis de expresión diferencial se seleccionaron los 10 *pipelines* que presentaban mejor ranking entre todos los analizados en el presente trabajo, sin considerar el método de normalización, ya que la mayoría de los métodos de expresión diferencial requieren como datos de partida las lecturas en crudo. Para su selección, se calculó la mediana del sumatorio de los rankings de precisión y exactitud de todos los *pipelines* que compartiesen los procesos de recortado, alineamiento y contaje. Los *pipelines* así elegidos para el análisis de expresión génica diferencial aparecen recogidos en la **Tabla 3.1**.

***Tabla 3.1**. Top 10 pipelines con mejores rankings en el estudio de la bondad de 192 pipelines utilizados en el estudio de la expresión génica diferencial.*

| Posición | Recortado | Alineamiento | Contaje |
|---|---|---|---|
| 1 | Trimomatic | RUM | HT-seq UNION |
| 2 | Trimomatic | RUM | HT-seq INTER |
| 3 | BBDuk | STAR | Stringtie |
| 4 | Cutadapt | STAR | Cufflinks |
| 5 | Cutadapt | RUM | Cufflinks |
| 6 | Trimomatic | STAR | Cufflinks |
| 7 | BBDuk | STAR | HT-seq UNION |
| 8 | Trimomatic | STAR | Stringtie |
| 9 | BBDuk | RUM | HT-seq UNION |
| 10 | BBDuk | HiSat2 | HT-seq UNION |

*El cálculo de los rankings de estos pipelines se llevó a cabo sin considerar el método de normalización utilizado en los 192 pipelines.*

En este análisis se utilizaron las 18 muestras descritas en la **Figura 3.1** resultando en un total de cinco cruces o escenarios de análisis. Estos cruces se caracterizaron por presentar variabilidad en cuanto a la cantidad de cambios de expresión génica esperados, de manera que la secuencia en orden descendente del número observado de genes estadísticamente significativos fue:

1) LCA-T0 *vs.* LCB-T0
2) LCA-T1 *vs.* LCA-T0
3) LCA-T2 *vs.* LCA-T0
4) LCB-T1 *vs.* LCB-T0
5) LCB-T2 *vs.* LCB-T0

A continuación, se procedió al filtrado de los genes con el objetivo de homogeneizar el número de genes analizados por todos los métodos de expresión diferencial. Los genes seleccionados para este análisis fueron aquellos con cuatro o más lecturas en las tres muestras de al menos una condición de tratamiento para cada cruce y cada *pipeline*. Aunque muchos de los algoritmos de expresión diferencial poseen un filtro intrínseco, este fue anulado para evitar que el número de genes variase con cada algoritmo. Sobre cada una de las tablas de lecturas resultantes de RNA-seq se aplicaron 17 variantes de métodos de expresión diferencial, algunas de ellas definidas en el trabajo de Seyednasrollah y colaboradores[253], y que se describen brevemente a continuación.

***Algoritmos y métodos para el análisis de la expresión génica diferencial***

a) *SAMseq* (*samr*, v2.0)[248]. Se trata de un método no paramétrico basado en le estadístico de Wilcoxon-Mann-Whitney, así como un método basado en permutaciones para calcular el FDR.
b) *edgeR* (v3.18.1)[243]. Implementa métodos empíricos bayesianos, tanto basados en estadísticos exactos (*exact test*) como en métodos basados en modelos generalizados lineales (GLM). En ambos casos, ajusta modelos basados en la

distribución binomial negativa. Este algoritmo se asoció en este trabajo a tres métodos de normalización de la expresión génica, TMM, RLE y UQ, que fueron comparados de forma independiente.

c) *DESeq2* (v1.16.1)[242]. Estima la dependencia entre la media y la varianza en los datos de contaje y lleva a cabo un análisis de expresión diferencial basado en la distribución binomial negativa usando modelos lineales generalizados.
d) *baySeq* (v2.10.0)[240]. Calcula la probabilidad estimada posterior de la expresión diferencial basado en métodos empíricos Bayesianos. Considera que los datos del experimento siguen una distribución binomial negativa.
e) *EBseq* (v1.16.0)[244]. Es una aproximación empírica bayesiana. Asume que las lecturas de las condiciones estudiadas siguen una distribución binomial negativa.
f) *limma* (v3.32.10)[246]. En un primer paso las lecturas son transformadas a *$log_2$ count-per-million* (logCPM), después se lleva a cabo la modelización de la relación media-varianza; si esta modelización se hace mediante la tendencia empírica de Bayes el método es conocido como *"limma-trend"*, por el contrario, si la modelización se hace con pesos de precisión el método se conoce como *"limma-voom"*[385].
g) *NOISeq* (v2.20.0)[247]. En el presente trabajo se utilizó la versión NOISeqBIO diseñada para réplicas biológicas. Lleva a cabo la expresión diferencial entre dos condiciones experimentales con suposiciones no paramétricas, no realiza ninguna suposición distribucional. Se aplicaron los tres métodos de normalización asociados a este algoritmo: FPKM, UQ y TMM. Todos ellos fueron comparados de forma independiente.
h) *Cuffdiff* (v2.2.1)[241]. Estima la incertidumbre calculando la confianza de que cada fragmento está correctamente asignado al transcrito que lo ha generado, esta incertidumbre es capturada como una distribución beta y la sobredispersión en los contajes la captura como una distribución binomial negativa. El programa mezcla las dos distribuciones, resultando en una distribución beta-binomial negativa, que reflejará las dos fuentes de variabilidad.
i) *Ballgown* (v2.8.4)[245]. Compara modelos lineales anidados mediante una prueba F paramétrica. Se ajustan dos modelos a cada variable, uno con la covariable de interés y el otro sin ella y entonces se calcula el estadístico F usando ambos modelos.

Estos algoritmos pueden clasificarse en función de la distribución estadística subyacente que siguen los contajes de lecturas de RNA-seq, de manera que *baySeq*, *Cuffdiff, DESeq2, EBSeq* y *edgeR* asumen una distribución binomial negativa, *limma* y *Ballgown* asumen una distribución log-normal y *NOISeq* y *SAMSeq* no asumen ningún tipo de distribución. Algunos trabajos, recomiendan, tal como hicieron los desarrolladores de los paquetes *edgeR* y *DESeq*[243, 386], el uso de herramientas basadas en la distribución binomial negativa de los contajes de lecturas génicas ya que, como demuestran con sus experimentos, la mayoría de genes es consistente con ese modelo[387]. Aseguran también

que el modelo log-normal también puede ser consistente salvo en los casos en que una o más réplicas contenga cero lecturas. En este trabajo, además de la evaluación de los algoritmos de expresión diferencial, se llevará a cabo una comparación de los tres tipos de distribución estadística mencionados y comprobaremos su aplicabilidad sobre datos reales de RNA-seq.

### *Comparación de la similitud de los algoritmos y métodos de análisis de la expresión génica diferencial*

La similitud entre los diferentes métodos empleados en el análisis de la expresión génica diferencial fue calculada mediante la distancia Euclídea de los FDR arrojados por cada método. Para homogeneizar el cálculo entre los 17 métodos de expresión diferencial, estos FDR fueron ordenados y transformados en función de su ranking en cada uno de los métodos, otorgando los rankings menores a los genes con un menor valor de FDR. El cálculo de la matriz de distancias se llevó a cabo utilizando la distancia Euclídea en el paquete estadístico SIMFIT (v.7.3.1)[337]. La representación gráfica de estas distancias se realizó en R mediante el uso de dendrogramas con el paquete *dendextend* (v.1.9.0)[388] utilizando la mediana de las distancias obtenidas entre los diferentes métodos en los 10 *pipelines* seleccionados para el análisis, en cada uno de los cinco escenarios de análisis considerados en este trabajo.

### *Comparación del rendimiento de los algoritmos y métodos de análisis de la expresión génica diferencial*

El rendimiento de cada algoritmo o método de análisis de la expresión diferencial fue evaluado mediante la siguiente batería de parámetros estadísticos, cuyo cálculo se apoyó en los datos de expresión génica validados mediante qRT-PCR, ya que los resultados obtenidos mediante esta técnica permiten la definición de los genes verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN):

a) *Sensibilidad o razón de verdaderos positivos* (RVP). Indica la capacidad del algoritmo de detectar como genes diferencialmente expresados aquellos genes que realmente lo están. Viene definido por la expresión:

$$RVP = \frac{VP}{VP + FN}$$

b) *Especificidad o razón de verdaderos negativos* (RVN). Indica la capacidad del algoritmo de detectar como genes que no están diferencialmente expresados a aquellos genes que realmente no lo están. Se caracteriza por la expresión:

$$RVN = \frac{VN}{VN + FP}$$

c) *Precisión o valor predictor positivo* (VPP). Indica la probabilidad de que el gen esté diferencialmente expresado cuando el algoritmo de análisis indica que sí lo está. Este parámetro viene dado por la expresión:

$$VPP = \frac{VP}{FP + VP}$$

d) *Exactitud* (ACC, del inglés *accuracy*). Es la capacidad del algoritmo de acercarse a la determinación real de genes tanto diferencialmente expresados como no diferencialmente expresados. La expresión que define la exactitud es:

$$ACC = \frac{(VP + VN)}{Total}$$

e) *Valor predictor negativo* (VPN). Indica la probabilidad de que el gen no esté diferencialmente expresado cuando el algoritmo de análisis indica que no lo está. Este parámetro viene dado por la expresión:

$$VPN = \frac{VN}{VN + FN}$$

f) *Área bajo la curva COR* (AUC). Es una medida de la exactitud del algoritmo. Para la construcción de la curva característica operativa del receptor (COR) se consideran dos parámetros: RVP y RVN. El AUC mide el área, en un espacio de dos dimensiones, que se localiza debajo de la curva COR. En este caso, el AUC podría interpretarse como la probabilidad de que el algoritmo clasifique un gen aleatorio diferencialmente expresado en mejor posición que un gen aleatorio que no esté diferencialmente expresado.

g) *Coeficiente de correlación de Matthews* (CCM). Mide la calidad de las clasificaciones binarias. Este parámetro toma valores comprendidos entre 1 y -1, siendo 1 cuando el algoritmo distingue perfectamente los genes que realmente están diferencialmente expresados de los que no lo están, y -1 cuando lo hace de forma completamente errónea.

$$CCM = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

# 3.3. Metodologías en el análisis de microarrays

## 3.3.1. Protocolos de hibridación y análisis bioinformático de las muestras estudiadas en este trabajo

***Hibridación y escaneado del microarray***

Para el estudio de los fármacos amilorida y TG003 sobre las líneas celulares KMS12-BM y JJN-3, se utilizó el array Human Transcriptome 2.0 (HTA2.0) de Affymetrix. Los protocolos de etiquetado e hibridación se llevaron a cabo en el horno de hibridación *GeneChip Hybridization Oven 640* siguiendo las recomendaciones del fabricante. El lavado y escaneado de los microarrays se llevaron a cabo utilizando el sistema *GeneChip* de Affymetrix mediante la estación de fluidos *GeneChip Fluidics Station 450* y el escáner *GeneChip Scaner 7G*, en ambos casos siguiendo las recomendaciones proporcionadas por el fabricante del microarray.

***Análisis bioinformático del microarray HTA2.0***

Para el análisis del HTA2.0 se utilizaron como punto de partida los archivos CEL procedentes del proceso de escaneado. Estos archivos CEL contienen los datos no procesados de intensidad de fluorescencia de cada uno de los *spots* del microarray. Con estos archivos se llevaron a cabo dos técnicas de preprocesamiento, diferentes en cuanto a la referencia empleada para la determinación de la expresión génica y a los algoritmos y software utilizados.

En lo que respecta a la primera técnica o técnica BrainArray (BA), los valores en bruto de la intensidad de fluorescencia fueron preprocesados eliminando el ruido de fondo, normalizados por cuantiles y transformados a $\log_2$ utilizando la función RMA[389, 390] existente en el paquete *oligo*(v.1.40.2)[391] en R. Los archivos CDF de referencia para este análisis fueron descargados desde la página web del proyecto BA[392] en su versión 19.0.0 para el nivel génico de Ensembl[374]. Este proceso devuelve una matriz con los valores de expresión génica en $\log_2$, en cuyas filas aparecen los identificadores génicos de Ensembl y en las columnas las muestras analizadas. En un siguiente paso se asoció a cada identificador de Ensembl los correspondientes identificadores de genes de Refseq y HGNC (*HUGO Gene Nomenclature Comitee*) provistos por Ensembl desde la aplicación Biomart (https://www.ensembl.org/biomart).

En cuanto a la segunda técnica o técnica de la consola Expression Console de Affymetrix (AEC), también se utilizaron como punto de partida los archivos CEL con los datos no procesados procedentes del escáner. El preprocesamiento de los datos se realizó mediante el método SST-RMA en la AEC (v.1.4.1.46). Este método añade algunos pasos adicionales de preprocesamiento que conducen al estiramiento de la distribución de los valores de intensidad de fluorescencia, ejecutando a continuación el método RMA. La

novedad que aporta este método al resultado final es la eliminación de la compresión de los FC propia de los microarrays, generando unos nuevos FC más fácilmente comparables con otras tecnologías como RNA-seq. Finalmente, se devuelve una matriz donde en las filas se indicarán los conjuntos de sondas o *probesets* y en las columnas las muestras. Los identificadores génicos de Refseq, HGNC y Ensembl fueron asociados a estos *probesets* tras haber sido descargados de Biomart.

### 3.3.2. Procedimientos bioinformáticos en el reanálisis de datos descargados de bases de datos

El reanálisis bioinformático de los datos en bruto se realizó siguiendo diferentes flujos de trabajo en función de la plataforma de microarray utilizada. En este trabajo podemos distinguir cinco variantes en cuanto a la plataforma de microarray empleada en los diferentes trabajos originales:

1) Microarrays 3′ de Affymetrix

Para el análisis de los microarrays 3′ de Affymetrix se partió también de los correspondientes archivos CEL. En este caso, los valores en bruto de la intensidad de fluorescencia fueron preprocesados eliminando el ruido de fondo, normalizados por cuantiles y transformados a $\log_2$ utilizando la función RMA implementada en el paquete affy (v.1.58.0)[393] en R. El resto del proceso de preprocesamiento se llevó a cabo de forma similar a la de la técnica BA para el HTA2.0. Este flujo de trabajo se utilizó para los microarray Human Genom U133 Plus 2.0, Affymetrix Human Genome U133 A Array, Affymetrix Human Genome U133 B Array y GeneChip® PrimeView™ Human Gene Expression Array.

2) Microarrays de gen, exón y transcriptoma de Affymetrix

De forma similar a los microarrays de Affymetrix citados anteriormente, se utilizaron como punto de partida los archivos CEL. La normalización se llevó a cabo mediante la función RMA presente en el paquete *oligo* en R, utilizando como referencia los archivos CDF a la carta de BA. Este flujo de trabajo fue utilizado con los microarrays Human Gen 1.0ST array, Human Exon 1.0 ST Array y Human Transcriptome Array 2.0 (HTA2.0).

3) Microarrays de Illumina

Se partió de los archivos de texto con los datos no normalizados escaneados directamente del microarray y sin procesar. El procesamiento de dichos datos se realizó con el paquete *limma* (v.3.28.14) en R. Para ello se utilizó la función *neqc*, la cual realiza la corrección del ruido de fondo utilizando los controles negativos del microarray, para posteriormente realizar una normalización por cuantiles, transformando finalmente los datos a $\log_2$. Finalmente devuelve una matriz con identificadores de sonda propios del

microarray en las filas y las muestras analizadas en las columnas. La asociación de los identificadores génicos de Ensembl, Refseq y HGNC se realizó tal y como se ha descrito en los apartados anteriores. Este flujo de trabajo se utilizó con los arrays HumanHT-12 V3.0 expression beadchip, HumanHT-12 V4.0 expression beadchip e Illumina Human Whole-Genome DASL HT array

4) Microarrays de un canal de Agilent

En el caso de los microarrays de un canal de Agilent se partió de los archivos de texto con los datos escaneados sin procesar. El preprocesamiento se llevó a cabo con el paquete *limma* (v.3.28.14)[246] en R y consta de un paso de corrección del ruido de fondo seguido de dos pasos de normalización. El primer paso de normalización es un proceso intraarray que fue realizado mediante el método *loess*, mientras que el segundo paso se trata de un proceso interarray en el que se realizó una normalización por cuantiles. El resultado final es una matriz con los datos procesados en la que en las filas se encuentran las sondas correspondientes del microarray junto con el identificador de HGNC y en las columnas las muestras analizadas. La asignación de los identificadores génicos de Ensembl y Refseq se realizó como se explica en apartados anteriores. Este flujo de trabajo se utilizó con los microarrays Agilent-014850 Whole Human Genome Microarray 4x44K G4112, Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray

5) Microarrays de dos canales de Agilent

El microarray de dos canales de Agilent fue preprocesado de un modo similar al microarray de un canal de Agilent, con la diferencia que al ser un array de dos canales se realizó el análisis por separado para cada canal. Este procedimiento se llevó a cabo de esta manera ya que los datos de las muestras tratadas estaban contenidos en el canal Cy5, mientras que el canal Cy3 contenía las muestras control. Este flujo de trabajo fue utilizado con el microarray Agilent-014850 Whole Human Genome Microarray 4x44K G4112F de dos canales.

### 3.3.3. Análisis de la expresión génica diferencial en microarrays

El análisis de la expresión génica diferencial del microarray se llevó a cabo mediante dos de los métodos más ampliamente utilizados en la literatura y con un mejor comportamiento a la hora de la determinación de la expresión génica diferencial[394]:

a) *SAMr*[395]. *SAMr* se basa en una estimación del FDR basada en un test *t* de Student con permutaciones. Este método compara el número de rechazos de la hipótesis nula en los subgrupos comparados frente a la mediana de hipótesis nulas rechazadas en una serie de subgrupos generados arbitrariamente. Este método se llevó a cabo en R a través del paquete *samr* (v.3.0)
b) *limma*[396]. *limma* ajusta un modelo lineal a los datos de expresión de cada gen. *limma* utiliza el test *t* de Student con métodos empíricos bayesianos para compartir

información del conjunto completo de genes haciendo así el análisis más estable, ya que modera las desviaciones estándar entre los genes. Este método se llevó a cabo en R en el paquete *limma* (v.3.28.14).

## 3.4. Metodologías en el análisis de qRT-PCR

El ARN utilizado para el análisis de qRT-PCR fue extraído usando el kit RNease Plus Mini de Qiagen. La integridad de este RNA fue medida con el Bioanalizador 2100 de Agilent. El ARN total fue sometido a transcripción inversa para obtener ADNc utilizando el kit *High-Capacity cDNA Reverse Transcription* de Applied Biosystems. Finalmente, la expresión génica fue cuantificada mediante el ensayo TaqMan qRT-PCR para mRNA de Applied Biosystems. La lista de oligonucleóticos utilizados en este ensayo aparece recogida en la **Tabla 3.2**.

***Tabla 3.2.*** *Sondas de oligonucleótidos utilizadas para el análisis por qRT-PCR, adquiridos a la compañía ThermoFisher.*

| Ensembl ID | HGNC ID | Cromosoma | TaqMan Gene Expression ID assay | Identificador |
|---|---|---|---|---|
| ENSG00000008018 | PSMB1 | 6 | Hs00427357_m1 | Cat#4448892 |
| ENSG00000044574 | HSPA5 | 9 | Hs00607129_gH | Cat#4448892 |
| ENSG00000065978 | YBX1 | 1 | Hs00358903_g1 | Cat#4448892 |
| ENSG00000068697 | LAPTM4A | 2 | Hs01092025_m1 | Cat#4448892 |
| ENSG0000007562 | ACTB | 7 | Hs01060665_g1 | Cat#4448892 |
| ENSG00000100138 | NHP2L1 (SNU13) | 22 | Hs03025442_s1 | Cat#4448892 |
| ENSG00000101361 | NOP56 | 20 | Hs00197340_m1 | Cat#4448892 |
| ENSG00000105193 | RPS16 | 19 | Hs01598518_gH | Cat#4448892 |
| ENSG00000105887 | MTPN | 7 | Hs00377581_m1 | Cat#4448892 |
| ENSG00000105968 | H2AFV | 7 | Hs00606542_mH | Cat#4448892 |
| ENSG00000107223 | EDF1 | 9 | Hs00610152_m1 | Cat#4448892 |
| ENSG00000107581 | EIF3A | 10 | Hs01025769_m1 | Cat#4448892 |
| ENSG00000111640 | GAPDH | 12 | Hs02786624_g1 | Cat#4453320 |
| ENSG00000115758 | ODC1 | 2 | Hs00159739_m1 | Cat#4453320 |
| ENSG00000119421 | NDUFA8 | 9 | Hs00204417_m1 | Cat#4448892 |
| ENSG00000127884 | ECHS1 | 10 | Hs00187943_m1 | Cat#4448892 |
| ENSG00000131495 | NDUFA2 | 5 | Hs04187282_g1 | Cat#4448892 |
| ENSG00000138279 | ANXA7 | 10 | Hs00559410_m1 | Cat#4448892 |
| ENSG00000142541 | RPL13A | 19 | Hs04194366_g1 | Cat#4448892 |
| ENSG00000142937 | RPS8 | 1 | Hs01374307_g1 | Cat#4448892 |
| ENSG00000143158 | MPC2 | 1 | Hs00967250_m1 | Cat#4448892 |
| ENSG00000147677 | EIF3H | 8 | Hs00186779_m1 | Cat#4448892 |
| ENSG00000154723 | ATP5J | 21 | Hs01081389_g1 | Cat#4448892 |
| ENSG00000161016 | RPL8 | 8 | Hs00361285_g1 | Cat#4453320 |
| ENSG00000163956 | LRPAP1 | 4 | Hs00936301_m1 | Cat#4448892 |
| ENSG00000165502 | RPL36AL | 14 | Hs00733231_m1 | Cat#4448892 |
| ENSG00000166337 | TAF10 | 11 | Hs00359540_g1 | Cat#4448892 |
| ENSG00000166710 | B2M | 15 | Hs00187842_m1 | Cat#4453320 |
| ENSG00000168264 | IRF2BP2 | 1 | Hs02930738_m1 | Cat#4448892 |
| ENSG00000175130 | MARCKSL1 | 1 | Hs00702769_s1 | Cat#4448892 |
| ENSG00000178741 | COX5A | 15 | Hs00362067_m1 | Cat#4448892 |
| ENSG00000182117 | NOP10 | 15 | Hs00430282_m1 | Cat#4448892 |

Los valores de Ct obtenidos fueron posteriormente normalizados mediante el método del ΔCt, calculado como:

$$\Delta Ct = Ct_{gen\ control} - Ct_{gen\ diana}$$

Con este cálculo se consigue mantener la relación de proporcionalidad directa entre el valor del ΔCt y la expresión real del gen. Además, el valor del ΔCt fue calculado mediante tres métodos diferentes:

a) *Normalización mediante control endógeno*: se utilizó como factor de normalización la media de los valores de Ct de los genes *ACTB* y *GAPDH* en cada una de las muestras.
b) *Normalización mediante la mediana global*: el factor de normalización fue calculado usando en cada muestra la mediana de los genes con un Ct < 35.
c) *Normalización mediante el gen más estable*: el gen más estable fue determinado mediante los cuatro algoritmos disponibles en la herramienta online RefFinder. Los algoritmos utilizados fueron *BestKeeper*[397], *NormFinder*[398], *Genorm*[399], y el *método comparativo del delta Ct*[400]. En el presente trabajo, el gen *ECHS1* fue el que alcanzó del mejor ranking considerando los cuatro algoritmos y fue el seleccionado para llevar a cabo el proceso de normalización.

El método utilizado finalmente para llevar a cabo aguas abajo los diferentes análisis de qRT-PCR fue la *mediana global*. El método *de normalización mediante control endógeno* fue descartado debido a la presencia de variabilidad en los genes *ACTB* y *GAPDH* asociada a los tratamientos empleados (**Figura 3.5**), mientras que el método del *gen más estable* fue descartado al ser menos robusto que la *mediana global*, ya que la *mediana global* captura mejor la dispersión de los valores de Ct en cada muestra al considerar un mayor número de genes.

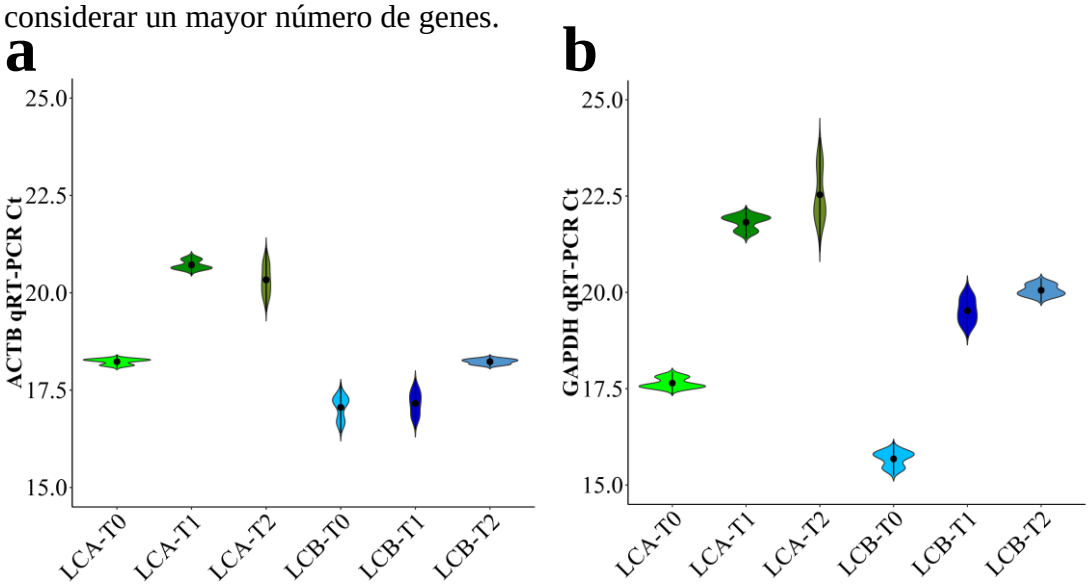



***Figura 3.5.*** *Expresión mediante qRT-PCR de los genes* ***a)*** *ACTB y* ***b)*** *GAPDH en las dos líneas celulares (LCA y LCB) y las tres condiciones de tratamiento (T0, T1 y T2). Los valores del eje de ordenadas representan el Ct obtenido a través de qRT-PCR.*

# 3.5. Determinación de la precisión y exactitud de los *pipelines* de RNA-seq

### 3.5.1. Determinación de la precisión en estudios de RNA-seq

Se eligieron 1.181 genes expresados de forma constitutiva en 32 tejidos sanos procedentes del *Human Protein Atlas* obtenidos de los datos de RNA-seq publicados previamente en el trabajo de Uhlén y colaboradores[401]. De estos genes fueron seleccionados para la determinación de la precisión los de mayor expresión en las seis muestras control de nuestro trabajo considerando los 192 *pipelines* analizados. El proceso de filtrado consistió en eliminar todos aquellos genes con menos de cuatro eventos (lecturas crudas, FPKM, TPM, etc.) en las tres muestras control de cada *pipeline* en cada una de las líneas celulares estudiadas. El resultado fue la selección de 107 genes óptimos que fueron utilizados para los estudios de precisión y exactitud.

Para el cálculo de la precisión se utilizó el "coeficiente de variación no paramétrico" (CoV) como índice de dispersión. El valor de este coeficiente fue calculado como el cociente entre la desviación absoluta de la mediana (MAD) y el valor absoluto de la mediana.

De esta manera, la MAD se calculó según la siguiente expresión:

$$MAD = mediana(|X_i - mediana(X)|)$$

donde $X_i$ es el valor de expresión del gen $X$ en la muestra $i$, expresado en las unidades que corresponda a cada *pipeline*: lecturas en crudo, FPKM, TPM, etc. Una vez determinado el valor de la MAD, se procedió al cálculo del CoV:

$$CoV = \frac{MAD}{|Mediana(X)|}$$

Este estadístico CoV define en tanto por uno, y de manera adimensional, la dispersión de cada gen en cada *pipeline*. Para definir qué *pipeline* fue el más preciso se otorgó un ranking a cada uno de los *pipelines* para cada gen en base al CoV obtenido. Finalmente, se calculó la mediana de todos los rankings considerando los 107 genes en cada *pipeline*, analizando las dos líneas celulares por separado. El *pipeline* con un ranking mediano menor sería la más precisa (ver **Figura 3.6**).

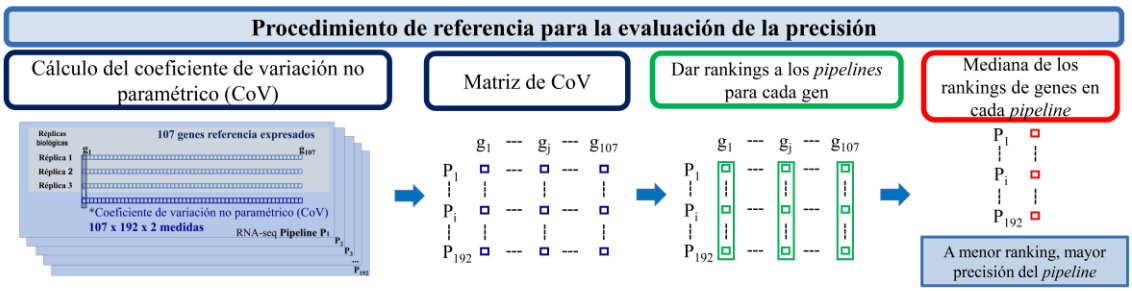



***Figura 3.6.*** *Descripción del procedimiento de evaluación de la precisión de los 192 pipelines de RNA-seq. $P_i$ hace referencia a los i pipelines analizados en este trabajo, mientras $g_j$ hace referencia a los j genes seleccionados para la evaluación de la precisión (107 genes).*

### 3.5.2. Determinación de la exactitud en RNA-seq mediante qRT-PCR

Para evaluar la exactitud en RNA-seq se seleccionaron 30 de los 107 genes óptimos previamente elegidos. De estos 107 genes, se seleccionaron los 10 genes con mayor y con menor CoV mediano (considerando los 192 *pipelines*), además de otros 10 genes al azar con CoV intermedio, de modo que quedase cubierto todo el rango dinámico del conjunto de datos de expresión génica. También se decidió añadir los genes *GAPDH* y *ACTB*, comúnmente utilizados en estudios de qRT-PCR. Con estos 32 genes se hizo el estudio mediante qRT-PCR utilizando cebadores Taqman prediseñados. La hipótesis es que ha de existir gran concordancia entre la expresión génica determinada por RNA-seq y qRT-PCR, como ya ha sido demostrado previamente[402]. En base a esta hipótesis, se calculó el coeficiente de correlación de Pearson ($r$) entre los Ct obtenidos por qRT-PCR y los diferentes valores de expresión génica para los 192 *pipelines* en $\log_2$. Del mismo modo que para la precisión, se calculó el ranking que ocupó cada gen considerando los 192 *pipelines* y se calculó un ranking mediano para cada *pipeline* sobre la base de los 32 genes (**Figura 3.7**). Este proceso se hizo de forma independiente en las dos líneas celulares.

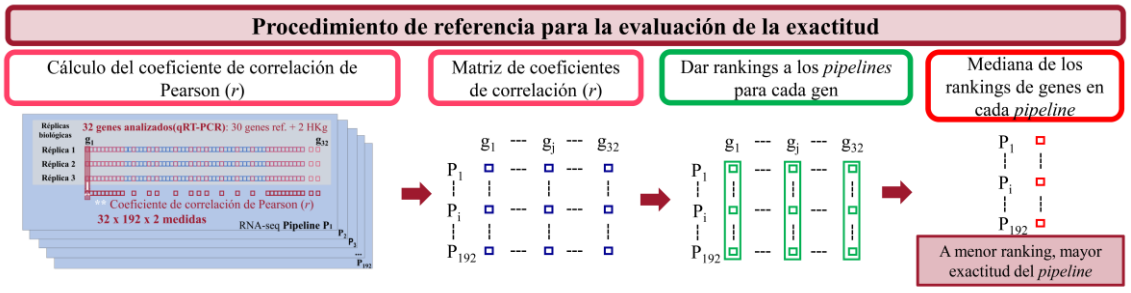



***Figura 3.7.*** *Descripción del procedimiento de evaluación de la exactitud de los 192 pipelines de RNA-seq. $P_i$ hace referencia a los i pipelines analizados en este trabajo, mientras $g_j$ hace referencia a los j genes seleccionados para la evaluación de la exactitud (32 genes).*

## 3.6. Determinación del ranking de los *pipelines* estudiados

Tras la obtención de los rankings de precisión y exactitud para cada uno de los 192 *pipelines* en las dos líneas celulares, se procedió a la determinación del ranking global mediante la ponderación de estos dos parámetros. La determinación del valor del ranking global que ocuparía cada *pipeline* dentro del conjunto de los 192 *pipelines* se realizó calculando el sumatorio de los rankings obtenidos por cada uno de los 192 *pipelines* para la exactitud y para la precisión, para lo que se utilizó la siguiente fórmula:

$$Perf_j = Med(RANK\left(CoV(X_{ij})\right))_{JJN-3} + Med(RANK\left(CoV(X_{ij})\right))_{KMS12-BM} + Med(RANK\left(r(X_{ij})\right))_{JJN-3} + Med(RANK\left(r(X_{ij})\right))_{KMS12-BM}$$

donde $Perf_i$ es el ranking global calculado en cada *pipeline j* para cada gen $X_i$, *CoV* es el coeficiente de variación no paramétrico, *r* es el coeficiente de correlación de Pearson, *RANK* es el ranking expresado en valores ordinales y *Med* es el valor mediano de los rankings. A menor valor de $Perf_j$, mejor funcionamiento del *pipeline*.

## 3.7. Comparación de los valores de expresión génica de RNA-seq frente a los medidos con microarrays

La comparación de las dos tecnologías de análisis de expresión génica utilizadas en este trabajo, RNA-seq y el microarray HTA2.0, se llevó a cabo a nivel de cuantificación de la expresión génica cruda y a nivel de expresión génica diferencial. El procedimiento comparativo a ambos niveles se detalla a continuación.

***Comparación a nivel de cuantificación de la expresión génica cruda***

Para hacer la comparación de la cuantificación de la expresión génica cruda se utilizaron las seis muestras control de las líneas celulares KMS12-BM y JJN-3. Se compararon los datos obtenidos de los 10 mejores *pipelines* (**Tabla 3.1** del **Apartado 3.2.5**) de RNA-seq con los datos resultantes de la normalización del microarray HTA2.0 mediante los procedimientos AEC y BA, mencionados en el **Apartado 3.3**. La asociación entre los datos se calculó mediante el coeficiente de correlación de Pearson y se llevó a cabo utilizando dos aproximaciones: una primera en la que se correlacionaron los genes comunes a ambas plataformas y una segunda en la que se correlacionaron los valores de ambas plataformas con los 32 genes seleccionados para el análisis con qRT-PCR. En todos los casos el análisis de correlación fue realizado con el paquete estadístico SIMFIT (versión 7.3.1).

Sobre los datos de correlación con los 32 genes se llevó a cabo también un análisis de búsqueda de genes con valores atípicos o *outliers* utilizando el paquete *mvoutlier* (v.2.0.9)[403] en R. Mediante el uso de este paquete se consideraron como *outliers* todos los genes situados más allá del percentil 97,5 de acuerdo con la diferencia entre la función empírica de la distribución de la distancia robusta de Mahalanobis y una función teórica simulada. Tras la detección de los genes *outliers* se procedió a un nuevo análisis de correlación sin considerar estos genes, conocido como correlación robusta.

***Comparación a nivel de expresión génica diferencial***

La comparación a nivel de expresión génica diferencial se llevó a cabo utilizando las 18 muestras que son el resultado del estudio por triplicado de las seis condiciones experimentales diferentes estudiadas en el presente trabajo. Se realizaron un total de cinco comparaciones o escenarios de análisis, resultantes de los cruces entre las seis condiciones experimentales, siendo estas comparaciones las siguientes:

a) LCA-T0 vs. LCB-T0
b) LCA-T1 vs. LCA-T0
c) LCA-T2 vs. LCA-T0
d) LCB-T1 vs. LCB-T0
e) LCB-T2 vs. LCB-T0

Se procedió al análisis comparativo cruzando los resultados obtenidos mediante *SAMr* y *limma* para los métodos BA y AEC del microarray contra el método que obtuvo un mejor rendimiento del estudio de RNA-seq, *limma trend* (ver **Figura 4.18** en el **Apartado 4.1.2.2** de la **Sección de Resultados y discusión**). Para realizar el correspondiente cruce, en un primer paso se seleccionaron los genes comunes entre el microarray y la RNA-seq, de manera que todos los análisis se llevaron a cabo con un mismo número de genes de entrada. Una vez seleccionados los genes a enfrentar, se procedió a la comparación de su expresión génica diferencial a dos niveles:

a) A nivel de detección: para este análisis se consideró el número total de genes comunes interrogados por las dos tecnologías. Se calculó el porcentaje de solapamiento de las listas de genes diferencialmente expresados detectados de manera global por el microarray, es decir, considerando todos los genes analizados por comparación, con las listas obtenidas a partir del método *limma trend* correspondientes a los 10 mejores *pipelines* sin considerar el método de normalización.
b) A nivel de rendimiento: se consideraron los 32 genes validados por qRT-PCR. Para realizar este análisis se calcularon 7 parámetros, como son la sensibilidad, la FVP, la FVN, el VPP, el VPN, la exactitud, el AUC y el CCM, de cada uno de los métodos de detección de expresión génica diferencial con el apoyo de los

datos de 32 genes validados por qRT-PCR. Estos cálculos se hicieron de forma independiente en cada uno de los cinco escenarios de análisis.

## 3.8. Metaanálisis del efecto farmacológico en la expresión génica en líneas celulares de mieloma múltiple

***Criterios de búsqueda de estudios***

Se llevó a cabo la búsqueda y obtención de estudios de expresión génica mediante microarrays de expresión génica y secuenciación de ARN en HMCLs, en los que se hubiese empleado algún tratamiento de uso habitual en el MM, considerando a tales efectos aquellos tratamientos recogidos por *Sociedad Americana de Cáncer* como terapias para el MM (https://www.cancer.org/). Para ello se realizaron búsquedas sistemáticas en los principales repositorios de almacenamiento masivo de datos: GEO (https://www.ncbi.nlm.nih.gov/geo/), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) y SRA (https://www.ncbi.nlm.nih.gov/sra). Del mismo modo, también se procedió a la búsqueda de este tipo de datos a través de portales especializados en análisis de datos de MM, como el portal CoMMpass de la *Multiple Myeloma Research Foundation* (MMRF) (https://themmrf.org/) o el portal MMGP del *Broad Institute* (http://portals.broadinstitute.org/mmgp/home).

Para cada uno de los fármacos candidatos para el metaanálisis se hicieron dos búsquedas, de manera que en la primera de ellas la búsqueda se basó en: (“cell line” OR “cell-line”), mientras que en la segunda búsqueda la secuencia fue (“myeloma” OR “plasma cell leukemia” OR “plasma cell leukaemia”). En ambos casos se incluyeron además las distintas denominaciones que pueden encontrarse para cada uno de los fármacos estudiados. Así, se emplearon las siguientes palabras clave específicas de cada fármaco, acompañadas del operador booleano “OR”.

1) *Melfalán*: “Melphalan”, “Alkeran”, “Sarcolysin”.
2) *Vincristina*: “Vincristine”, "Oncovin".
3) *Ciclofosfamida*: “Cyclophosphamide”, "Cytosan".
4) *Doxorrubicina*: “Doxorubicin”, "Adriamycin".
5) *Etopósido*: “etoposide”.
6) *Bendamustina*: “bendamustine” “Treanda” “Ribomustin”.
7) *Dexametasona*: “Dexamethasone”, “Decadron”, “Dexasone”, “Diodex”, “Hexadrol”, “Maxidex”.
8) *Prednisona*: “Prednisone”, “Deltasone”, “Liquid Pred”, “Orasone”, “Adasone”, “Deltacortisone”, “Prednisonum”.
9) *Talidomida*: “Thalidomide”, “Thalomid”, “Immunoprin”, “Talidex”, “Talizer”, “Neurosedyn”.
10) *Lenalidomida*: “Lenalidomide”, “Revlimid”.

11) *Pomalidomida*: “Pomalidomide”, “Imnovid”.
12) *Bortezomib*: “Bortezomib”, “Velcade”, “Neomib”, “Bortecad”.
13) *Carfilzomib*: “Carfilzomib”, “Kyprolis”.
14) *Ixazomib*: “Ixazomib”, “Ninlaro”.
15) *Panobinostat*: “Panobinostat”, “LBH-589”, “LBH589”, “Farydak”.
16) *Daratumumab*: “Daratumumab”, “Darzalex”.
17) *Elotuzumab*: “Elotuzumab”, “Empliciti”, “HuLuc63”.
18) *Interferon*: “Interferon”.
19) *Decitabina*: “Decitabine”, “5-aza-2′deoxycytidine”, "5-aza-2'-deoxycytidine", “5-aza-2-deoxycytidine”, “Dacogen”.
20) *Azacitidina*: “Azacitidine”, “Azacytidine”, “Ladakamycin”, “5-azacytidine”, “320-67-2”, “U-18496”, “4-Amino-1-beta-D-ribofuranosyl-s-triazin-2(1H)-one”, “Vidaza”.
21) *JQ1*: “JQ1”, “JQ-1”, “JQ 1”.

***Criterios de inclusión y exclusión de estudios***

Se establecieron 7 criterios de inclusión para las series encontradas en los distintos repositorios. Estos criterios serían aplicados de manera secuencial según aparecen mencionados a continuación:

a) *Especie*: solamente se admitirán datos humanos.
b) *Patología a estudio*: MM o leucemia de células plasmáticas.
c) *Diseño experimental*: los estudios deben presentar una condición de estudio control, sin tratar o a tiempo cero, y una condición donde las células hayan sido tratadas en monoterapia con el fármaco considerado.
d) *Tecnología*: estudios con tecnologías de alta resolución como microarray o RNA-seq, donde el objeto de estudio sea la expresión génica.
e) *Muestras*: las muestras deben proceder de líneas celulares y deben ser líneas celulares sensibles al fármaco.
f) *Datos*: los datos que presentan los estudios deben ser reanalizables, a fin de garantizar la homogeneidad en el tratamiento de los datos de expresión génica.
g) *Datos propios (Servicio de Hematología de Salamanca)*: se permite la introducción de datos de líneas celulares sometidas a tratamiento procedentes del servicio de Hematología de Salamanca no depositados en repositorios online, siempre que cumplan los criterios de inclusión anteriores.

No hubo criterios específicos de exclusión

***Extracción y análisis de los datos en bruto***

Preferentemente fueron extraídos los datos en bruto de expresión génica, es decir, los datos sin ningún tipo de preprocesamiento aplicado. En los estudios que presentaron ausencia de estos datos, se extrajo la matriz proporcionada por los autores en los correspondientes repositorios. La ausencia de ambos formatos en todos los repositorios estudiados implicaría que el estudio sería descartado del metaanálisis.

Los datos en bruto extraídos de los repositorios online fueron analizados de manera acorde a la plataforma de expresión génica utilizada por los autores. Todos los estudios fueron homogeneizados en cuanto a la nomenclatura génica utilizando como identificador génico la versión de Ensembl 77. La metodología empleada para cada microarray en función de la plataforma utilizado se detalla en el **Apartado 3.3.2**.

En el caso del análisis de los datos procedentes de estudios de RNA-seq se llevó a cabo utilizando los algoritmos del **Apartado 3.2** que obtuvieron un mejor ranking acorde al procedimiento explicado en los **Apartados 3.5 y 3.6**.

***Extracción de genes comunes entre los estudios seleccionados para metaanálisis***

Se procedió a homogenizar el número de genes candidatos a ser analizados en cada uno de los metaanálisis. Así, en el caso de los metaanálisis del efecto farmacológico sobre la expresión génica en líneas celulares, cada uno de los estudios seleccionados para cada fármaco contaría con un mismo número de genes iniciales. La homogenización en el número de genes consistió en determinar los genes comunes a los estudios analizados en cada metaanálisis mediante el cruce de las listas de genes interrogados por las plataformas de dichos estudios. Los genes comunes serían los utilizados como lista de genes inicial en cada metaanálisis.

## 3.9. Metaanálisis de la respuesta a tratamiento en pacientes con mieloma múltiple

***Criterios de búsqueda de estudios***

De manera similar a lo expuesto en el **Apartado 3.9**, se realizó una búsqueda de estudios de expresión génica mediante microarrays y RNA-seq en los repositorios de almacenamiento masivo de datos como son GEO, ArrayExpress y SRA (https://www.ncbi.nlm.nih.gov/sra), así como en portales especializados en análisis de datos de MM, como el portal CoMMpass o el portal MMGP. En este caso el objetivo de la búsqueda fue la obtención de datos de expresión génica de pacientes de MM que tuvieran datos disponibles de la respuesta a tratamiento, para ello la única palabra clave empleada fue la palabra “myeloma”. La inclusión y exclusión de estudios se realizó

revisando uno a uno todos los estudios candidatos, considerando los criterios de inclusión y exclusión que se exponen en el siguiente apartado.

***Criterios de inclusión y exclusión de estudios***

Los criterios de inclusión específicos para el estudio de la respuesta a tratamiento en muestras de pacientes con MM fueron:

a) *Especie*: muestras de tejido humano.
b) *Población*: médula ósea de pacientes con MM.
c) *Estadio de la patología*: MM en el momento del diagnóstico.
d) *Tecnología*: estudios de tecnologías de alta resolución como microarray y RNA-seq donde el objeto de estudio sea la expresión génica.
e) *Tratamiento*: el estudio debe disponer de datos suficientes para determinar el tratamiento al que ha sido sometido cada paciente tras el diagnóstico y no debe ser un tratamiento *in vitro* de células extraídas del paciente.
f) *Respuesta*: el estudio debe facilitar los datos de la respuesta a la primera línea de tratamiento para cada paciente estudiado.
g) *Datos propios (Servicio de Hematología de Salamanca)*: se permite la inclusión de datos de respuesta sobre series publicadas cuya procedencia sea el servicio de Hematología de Salamanca, siempre que la serie publicada cumpla con los criterios de inclusión anteriores. Todos los estudios seleccionados del servicio de Hematología de Salamanca fueron aprobados por un comité ético de investigación y se obtuvo un consentimiento informado por escrito de todos los pacientes de acuerdo con la Declaración de Helsinki.

No se establecieron criterios específicos de exclusión.

***Extracción y análisis de los datos en bruto***

Los procesos de extracción y análisis de los datos en bruto obtenidos para cada uno de los estudios seleccionados en el análisis de la respuesta a tratamiento en pacientes con MM se realizaron de forma similar a lo expuesto en el **Apartado 3.9**.

***Extracción de genes comunes entre los estudios seleccionados para metaanálisis***

La homogenización del número de genes analizados en el caso del metaanálisis de la respuesta a tratamiento se llevó a cabo de forma similar al procedimiento expuesto en el **Apartado 3.9**, con la singularidad de que en este caso esta homogenización se llevó a cabo por régimen de tratamiento. Los genes comunes serían los utilizados como lista de genes inicial en cada metaanálisis.

# 3.10. Procedimientos estadísticos utilizados en los metaanálisis

## 3.10.1. Métodos estadísticos de aplicación general

Los metaanálisis, en todos los casos, se realizaron empleando el modelo de efectos aleatorios del paquete *metafor* (v.2.0-0)[336] en R. Se decidió proceder de manera general con el modelo de efectos aleatorios debido a que los resultados obtenidos por este modelo son más conservadores que los obtenidos por el modelo alternativo de efecto fijo, ya que incorpora la medida de la variabilidad *entre* los estudios, además de la variabilidad *dentro* de cada estudio. Lo que nos motivó a realizar el análisis de este modo conservador fue la previsión de la existencia de una cierta heterogeneidad entre los estudios a combinar, ya que no solamente encontraríamos diferencias en cuanto al tiempo de tratamiento y la concentración aplicada del fármaco, sino también en cuanto a las líneas celulares y pacientes seleccionados en cada estudio. Por estas razones, además de usar siempre el modelo de efectos aleatorios, hubo que realizar en la mayor parte de los casos un análisis por subgrupos, según los tiempos de tratamiento y las concentraciones empleadas de los fármacos estudiados.

***Análisis de la heterogeneidad***

La heterogeneidad en cada uno de los metaanálisis fue identificada y cuantificada mediante los estadísticos Q de Cochran y el $I^2$.

La Q de Cochran[342] fue calculada como el sumatorio ponderado de las diferencias al cuadrado entre los efectos individuales de cada estudio y el efecto combinado con todos los estudios. Su valor viene definido por la siguiente expresión matemática:

$$Q = \sum_{i=1}^{k} W_i(Y_i - M)^2$$

donde $W_i$ es el peso del estudio *i* calculado como el inverso de la varianza ($1/V_i$), $Y_i$ es el tamaño del efecto del estudio *i*, *M* es el efecto combinado y *k* es el número de estudios.

El estadístico Q así calculado sigue una distribución $\chi^2$ y contrasta la hipótesis nula de que los estudios son homogéneos, el rechazo de esta hipótesis implica presencia de heterogeneidad. Para más detalles consultar la página 109 del trabajo de Borenstein y colaboradores[334].

Sin embargo, este estadístico Q, en ocasiones no muestra suficiente potencia en el cálculo de la heterogeneidad, principalmente cuando se dispone de un número bajo de estudios en el metaanálisis[404]. Por este motivo, suele utilizarse también el estadístico $I^2$ de medida de la heterogeneidad[345], ya que este test no tiene una dependencia exclusiva del número de estudios considerados. El estadístico $I^2$ cuantifica el porcentaje de

variación entre los estudios que es debido a la heterogeneidad y no al azar, reportando una estimación cuantitativa, en porcentaje, del valor de la heterogeneidad. Se calcula como:

$$I^2 = \left(\frac{Q - gl}{Q}\right) x\ 100\%$$

donde *Q* es el valor del estadístico Q de Cochran y *gl* son los grados de libertad calculados como:

$$gl = k - 1$$

siendo *k* el número de estudios considerados en el metaanálisis.

Se suele considerar que, cuando el valor del $I^2$ es superior al 50%, la heterogeneidad de los estudios recogidos en el metaanálisis es importante.

***Metaanálisis de efectos aleatorios***

En contraposición al metaanálisis que sigue un modelo de efectos fijos, que parte de la asunción de que únicamente existe un verdadero tamaño del efecto poblacional ($\mu$) y es el mismo en todos los estudios analizados, el metaanálisis de efectos aleatorios asume que el tamaño del efecto puede variar entre los diferentes estudios, de manera que cada tamaño del efecto individual puede estimar un diferente tamaño del efecto poblacional. Así, la medida del efecto ($Y_i$) de cada estudio en un modelo de efectos aleatorios se asume que contiene tres componentes:

$$Y_i = \mu + \delta_i + \varepsilon_i$$

donde $\mu$ es el promedio de todos los tamaños del efecto verdaderos (efecto que se quiere estimar), $\delta_i$ es la variación real del estudio *i* en el tamaño del efecto respecto al verdadero tamaño del efecto poblacional ($\mu$) y $\varepsilon_i$ es el error debido al muestreo. La representación de este modelo de efectos aleatorios se recoge en la **Figura 3.8**.

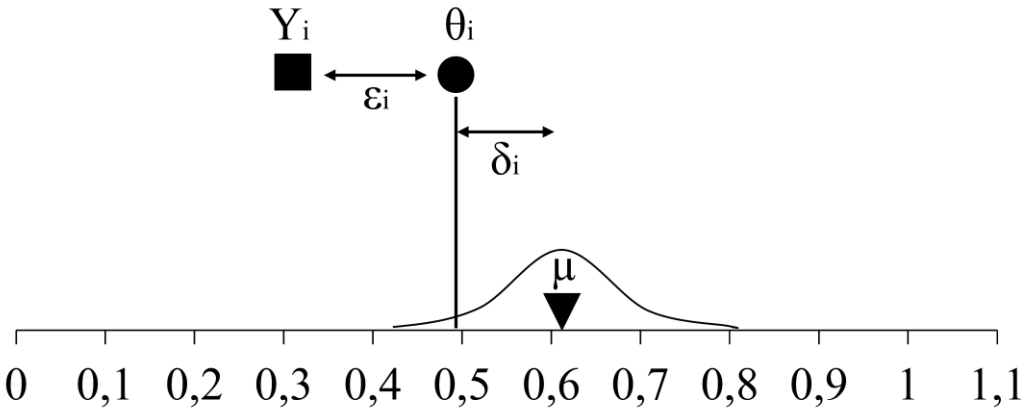



***Figura 3.8.*** *Representación del modelo de efectos aleatorios. Adaptado del libro "Introduction to meta-analysis" de Borenstein (2009).* $\theta_i$ *corresponde al verdadero tamaño del efecto del estudio i,* $Y_i$ *es el tamaño del efecto observado para el estudio i,* $\mu$ *es el promedio de todos los tamaños del efecto verdaderos,* $\varepsilon_i$ *es el error debido al muestreo y* $\delta_i$ *es la variación real en el tamaño del efecto del estudio i respecto al valor verdadero.*

Por tanto, de aquí se deducen dos posibles fuentes de variabilidad en el metaanálisis, una variabilidad *dentro* de los estudios ($\varepsilon_i$) y una variabilidad *entre* los estudios ($\delta_i$). Lo que se consigue a través de la aplicación del modelo de efectos aleatorios es minimizar tanto la variabilidad *entre* los estudios, considerando la varianza *entre* estudios o tau cuadrado ($\tau^2$), como la variabilidad *dentro* de los estudios a través de la consideración de la varianza *intra*estudio ($S_i^2$).

En lo que respecta a la $\tau^2$, en este trabajo fue calculada por el método de DerSimonian-Laird[343], a través de la ecuación:

$$\tau^2 = \frac{Q - gl}{C}$$

donde

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}$$

y

$$gl = k - 1$$

siendo *k* el número de estudios, y

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

A continuación, se llevó a cabo el cálculo del efecto promedio o combinado. La medida del tamaño del efecto que se utilizó en todos los metaanálisis fue la ratio de respuesta de los valores promedio de la expresión génica de los dos grupos contrastados, "grupo problema" (1) y "grupo control" (2):

$$R_i = \frac{\bar{X}_{i,1}}{\bar{X}_{i,2}}$$

donde $R_i$ es la ratio de respuesta en el estudio *i*, $\bar{X}_{i,1}$ es el promedio del "grupo problema" y $\bar{X}_{i,2}$es el promedio del "grupo control". Esta ratio fue expresada en escala logarítmica, realizando este cálculo de la siguiente manera:

$$lnR_i = \ln\left(\frac{\bar{X}_{i,1}}{\bar{X}_{i,2}}\right) = \ln\left(\bar{X}_{i,1}\right) - \ln\left(\bar{X}_{i,2}\right)$$

La varianza de este *lnR* se calcula de acuerdo con la expresión:

$$V_{lnR,i} = S^2_{combinada,i}\left(\frac{1}{n_{i,1}\left(\bar{X}_{i,1}\right)^2} + \frac{1}{n_{i,2}\left(\bar{X}_{i,2}\right)^2}\right)$$

donde $n_{i,1}$ y $n_{i,2}$son los tamaños muestrales de los grupos problema y control del estudio *i*, respectivamente, y $S^2_{combinada,i}$ la varianza combinada, cuyo cálculo viene dado por la fórmula:

$$S^2_{combinada,i} = \sqrt{\frac{\left(n_{i,1}-1\right)S^2_{i,1} + \left(n_{i,2}-1\right)S^2_{i,2}}{n_{i,1}+n_{i,2}-2}}$$

siendo $S_{i,1}$ y $S_{i,2}$ las desviaciones estándar de los grupos problema y control, respectivamente.

Ahora se procede al cálculo del efecto combinado propiamente dicho de todos los estudios analizados. El efecto promedio ($Y_p$) viene dado por la expresión:

$$Y^*_p = \frac{\sum W^*_i lnR_i}{\sum W^*_i}$$

donde el asterisco denota el análisis bajo el modelo de efectos aleatorios y $W^*_i$ corresponde a los pesos estadísticos de cada estudio *i*, calculados como:

$$W^*_i = \frac{1}{V_{lnR,i} + \tau^2}$$

En un siguiente paso, se calcula el error estándar (SE) del promedio:

$$SE_p^* = \sqrt{\frac{1}{\sum W_i^*}}$$

determinándose finalmente el efecto medido (Y) como:

$$Y = Y_p^* \pm 1{,}96 * SE_p^*$$

Tras estos cálculos, se aplicó una prueba estadística "z" a estos valores para comprobar la significancia del dato promedio. Para más información acerca del modelo de efecto aleatorios, consultar el Capítulo 12 de Borenstein[334].

***Metaanálisis en presencia de subgrupos y metaanálisis global***

Los estudios seleccionados para cada metaanálisis fueron estratificados en función del tiempo de tratamiento y de la concentración aplicada del fármaco. En ambos casos, se estableció un máximo de tres subgrupos determinados por la mediana, o bien del tiempo de tratamiento, o bien de la concentración de fármaco y la MAD en sentido positivo o negativo, indicando este valor los dos puntos de corte para el establecimiento de los tres subgrupos. En cada uno de los subgrupos se llevó a cabo un metaanálisis de efectos aleatorios para cada uno de los genes seleccionados. En cada subgrupo fue calculada también su $\tau^2$ correspondiente. Finalmente, los resultados del efecto combinado de la expresión génica en cada subgrupo fueron comparados entre sí mediante una prueba estadística tipo Wald disponible en el paquete *metafor*, para la comparación de dos subgrupos cualesquiera entre sí. La expresión matemática para esta prueba viene dada por la fórmula:

$$z = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{SE[\hat{\mu}_1]^2 + SE[\hat{\mu}_2]^2}}$$

Donde $\hat{\mu}_1$ y $\hat{\mu}_2$ son los valores combinados estimados de los dos subgrupos y $SE[\hat{\mu}_1]$ y $SE[\hat{\mu}_2]$ son sus correspondientes errores estándar.

Aunque se hagan metaanálisis por subgrupos y se calculen los efectos combinados dentro de cada subgrupo, a veces puede resultar de interés el calcular un efecto combinado global que incluya conjuntamente los estudios de todos los subgrupos de cada fármaco. En este caso se hizo un cálculo siguiendo el método de efectos aleatorios y utilizando un $\tau^2$ conjunto considerando todos los estudios a la vez sin tener en cuenta su pertenencia a los subgrupos. Como medida del tamaño del efecto se empleó la ratio de los valores medios de los grupos contrastados transformada de manera logarítmica, tal y como ha sido explicado anteriormente.

***Detección de sesgo de publicación***

El análisis de sesgo de publicación se llevó a cabo en todos los genes analizados mediante la prueba de asimetría o regresión de Egger[344]. Este test se encarga de cuantificar el grado de asimetría del gráfico en embudo o *Funnel plot* mediante el ajuste de una recta de regresión entre la precisión de los estudios y el efecto estandarizado.

### 3.10.2. Particularidades del metaanálisis con líneas celulares

***Selección de los genes candidatos***

La selección de los genes candidatos para el metaanálisis con líneas celulares se llevó a cabo seleccionando los genes que, considerando el valor absoluto del FC, presentaron un cambio de expresión génica mayor a 1,5 veces entre los grupos de tratamiento y control en todos los estudios seleccionados para cada uno de los fármacos estudiados. Adicionalmente, fueron añadidos a esta lista los genes que tuvieron un valor absoluto del FC mayor a 1,5 en todos los estudios de alguno de los subgrupos de tiempo de tratamiento o concentración de fármaco, excluyendo en este cálculo los subgrupos que solamente constaron de un único estudio. La distribución de los FC de los genes seleccionados fue chequeada a través del uso de diagramas de caja realizados en R con los paquetes *ggplot2* (v.3.1.0)[405] y *ggbeeswarm* (v.0.6.0)[406].

***Desviación estándar de los estudios con una única muestra por grupo de tratamiento***

Los estudios con una única muestra por grupo de tratamiento carecieron de valores de desviación estándar (SD), por lo que fue necesario su imputación para su inclusión en el metaanálisis. Para proceder con la imputación, en primer lugar, se seleccionó la referencia a imputar aplicando los siguientes criterios ordenados por prioridad, de manera que ante la ausencia del criterio que ocupa el ranking superior se aplicó el siguiente:

1º Estudio que utilice la misma tecnología (RNA-seq o microarray), de la misma casa comercial, hibridado en la misma plataforma que el estudio a imputar y el estudio esté realizado en el mismo fármaco.

2º Estudio que utilice la misma tecnología (RNA-seq o microarray), de la misma casa comercial e hibridado en la misma plataforma que el estudio a imputar, independiente del fármaco estudiado.

3º Estudio que utilice la misma tecnología (RNA-seq o microarray), de la misma casa comercial y el estudio esté realizado en el mismo fármaco.

4º Estudio que utilice la misma tecnología (RNA-seq o microarray) y de la misma casa comercial, independientemente del fármaco estudiado.

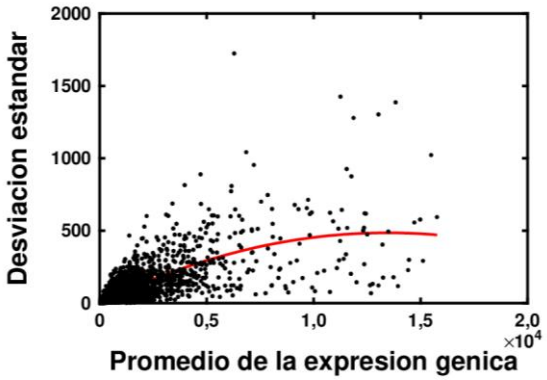
5º Estudio que utilice la misma tecnología (RNA-seq o microarray) y el estudio esté realizado en el mismo fármaco.

6º Estudio que utilice la misma tecnología (RNA-seq o microarray), independientemente del fármaco estudiado.

Si dos o más estudios referencia empataron en alguno de los criterios, se optó por el que presentó un mayor tamaño muestral.

Nótese que en todos los casos se dio prioridad a la técnica de análisis frente al compuesto analizado, ya que para realizar la imputación no se consideró la posición concreta del gen a imputar dentro de la referencia, sino su expresión dentro del conjunto de la expresión génica global, lo que depende en mayor medida de la técnica de análisis (tecnología y plataforma), que de las condiciones experimentales analizadas.

Tras la determinación de la referencia necesaria para la imputación, se procedió al ajuste de un modelo polinómico utilizando para ello los valores promedio de expresión génica y la SD de cada uno de los genes en cada grupo de tratamiento de este estudio de referencia. El ajuste se llevó a cabo utilizando un polinomio de segundo grado utilizando la función *polnom* en SIMFIT, obteniendo un resultado como el que se muestra en la **Figura 3.9**.




$$y = 8{,}7 * 10^{-1} + 7{,}2 * 10^{-2}x - 2{,}7 * 10^{-6}x^2$$

***Figura 3.9.*** *Ejemplo gráfico y ecuación del ajuste del modelo polinómico de grado 2 para imputación de la desviación estándar en estudios con una única muestra por grupo de tratamiento. En la Figura se muestra en rojo el modelo de referencia ajustado para el grupo de tratamiento con bortezomib en el metaanálisis en líneas celulares tratadas con ese mismo compuesto.*

Una vez el modelo estuvo ajustado, se procedió a la imputación de las SD en el estudio con una única muestra por grupo de tratamiento ($y$ en la ecuación), considerando

para ello la ecuación obtenida del modelo y los valores de expresión génica del estudio a imputar (*x* en la ecuación).

***Determinación de los subgrupos de tiempo de tratamiento y concentración de fármaco***

En todos los metaanálisis de fármacos con líneas celulares se procedió a la determinación de tres subgrupos de tiempo de tratamiento y concentración de fármaco aplicada, establecidos en función de la mediana ± MAD de estos dos parámetros, de la siguiente manera:

a) *Subgrupo 1*: tiempo o concentración ≤ mediana – MAD
b) *Subgrupo 2*: tiempo o concentración > mediana – MAD y ≤ mediana + MAD
c) *Subgrupo 3*: tiempo o concentración > mediana + MAD

### 3.10.3. Particularidades del metaanálisis con pacientes con mieloma múltiple

***Selección de los genes candidatos***

Se determinaron los genes comúnmente interrogados por las plataformas de análisis de expresión génica en los estudios recogidos para cada régimen de tratamiento. Con estos genes comunes se llevó a cabo un análisis de la expresión génica diferencial en cada estudio mediante los algoritmos *limma* o *edgeR* según correspondiese a muestras de *microarray* o RNA-seq, respectivamente, seleccionando los genes que obtuvieron un *p*-valor < 0,05. A continuación se procedió al cruce de las listas de genes de todos los estudios por régimen de tratamiento mediante diagramas de Venn, seleccionando finalmente como genes candidatos al metaanálisis aquellos que fueron comunes a al menos dos de los estudios.

## 3.11. Modelos de predicción de respuesta al tratamiento en pacientes con mieloma múltiple

### 3.11.1. Grupos de pacientes utilizados

La predicción de la respuesta a tratamiento en pacientes con MM se llevó a cabo en tres grupos de tratamiento:

a) Pacientes tratados con bortezomib en monoterapia.
b) Pacientes tratados con regímenes basados en bortezomib, pero que no incluían agentes inmunomoduladores (IMiDs).
c) Pacientes tratados con bortezomib en terapia combinada con IMiDs.

En los tres grupos, se procedió a la estratificación de los pacientes en función de la codificación de la respuesta, la cual, fue codificada de cuatro formas diferentes con el fin de llevar a cabo los análisis predictivos:

a) Pacientes respondedores *vs.* no respondedores (OR *vs.* NR): Se crearon dos grupos de pacientes en todos los estudios. El grupo de pacientes respondedores comprendió los pacientes que alcanzaron como mínimo una respuesta parcial (RP) al tratamiento, mientras que el segundo grupo recogió los pacientes en los que la respuesta fue inferior a RP, generalmente aquellos pacientes con enfermedad estable o progresiva.
b) Pacientes que alcanzan respuesta completa *vs.* resto de pacientes (RC *vs.* Resto): La respuesta en este caso fue de nuevo estratificada en dos grupos: los que alcanzaron RC, que incluía los pacientes que como mínimo tenían una inmunofijación (IFE) negativa, y los que no llegaron a alcanzar RC.
c) Codificación en tres grupos: Se codificó la respuesta en tres grupos basados en la publicación de Zhang y colaboradores[198].
d) Multirrespuesta: En cada estudio se utilizaron todos los niveles de respuesta disponibles para realizar la predicción de la respuesta.

Una vez estratificados los grupos, se procedió a la selección de las muestras que formaron las matrices de entrenamiento y de validación, de manera que para cada grupo de respuesta dos terceras partes de los pacientes fueron seleccionados para formar parte de la matriz de entrenamiento, mientras que una tercera parte de los pacientes formó parte de la serie de validación.

### 3.11.2. Selección de los genes a utilizar en la predicción

La selección de los genes de partida para los estudios de predicción se realizó de manera diferente en función de la codificación de la respuesta. De esta manera, en el caso de las predicciones para las respuestas codificadas como OR *vs.* NR y RC *vs.* Resto, se partió de los genes estadísticamente significativos ($p$-valor < 0,05) de los correspondientes metaanálisis de la respuesta a tratamiento en pacientes de MM. En las predicciones en las que la respuesta fue codificada en tres grupos, se llevó a cabo en un primer paso un análisis mediante ANOVA de un factor para grupos independientes en cada uno de los estudios seleccionados con el paquete *limma* en R. En un siguiente paso, se seleccionaron en cada uno de los estudios los genes que en el ANOVA tuvieron un $p$-valor < 0,05. Finalmente, se procedió al cruce de las listas de genes de todos los estudios, seleccionando los genes comunes para realizar los estudios de predicción. En el caso de la predicción considerando multirrespuesta, el análisis se realizó de manera independiente en cada uno de los estudios seleccionados, ya que la codificación de la respuesta a tratamiento por cada uno de los autores fue diferente. Así, en cada estudio se realizó un estudio mediante técnicas ANOVA considerando todos los grupos de respuesta. Los

genes que alcanzaron un *p*-valor < 0,05 en estos análisis fueron seleccionados para hacer los análisis de predicción de respuesta.

Con estas listas iniciales de genes se llevó a cabo una primera predicción de la respuesta. Además se realizaron otras dos predicciones: la que sería la segunda predicción, realizada tras aplicar a estas listas un proceso de filtrado para reducir el número de genes analizado, y la tercera predicción, considerando la suma de los genes de todos los estudios de la segunda predicción, como se explicará más adelante.

El proceso de filtrado fue aplicado en primer lugar para reducir la dimensionalidad del estudio ya que en muchos casos se dispuso de un gran número de variables, lo que podría producir cierta ralentización e ineficiencia computacional debido al efecto Hughes[407]. Este filtrado también estuvo justificado en segundo lugar para reducir el ruido de fondo que existe de manera implícita en un conjunto tan amplio de variables predictoras, y tratar así de separarlo del patrón predictor subyacente. En un tercer lugar, con el filtrado se trató de evitar el sobreajuste del modelo de predicción, ya que podría ocurrir que, al contar con un número tan elevado de variables, los resultados sobre la serie de entrenamiento podrían no ser reproducibles en la serie de validación debido a que se ha sobreajustado esa serie de entrenamiento.

El proceso de filtrado se llevó a cabo con el paquete *Boruta* (v.6.0.0)[408] en R. *Boruta* trata de capturar las variables predictoras de interés respecto a una variable respuesta basándose en un algoritmo de clasificación de *Random Forest*. Este algoritmo asigna un parámetro de "importancia" a cada una de las variables predictoras, de modo que finalmente determina qué variables son "importantes" para explicar la variable respuesta. El filtrado con *Boruta* se realizó de manera independiente en cada uno de los estudios de cada grupo de respuesta, y con las listas de genes resultantes se procedió al segundo de los análisis de la predicción de la respuesta a tratamiento.

Por último, se realizó, como se indicó anteriormente, un tercer análisis de la predicción de respuesta que consistió en crear una lista de genes que sería la suma de los genes de las listas de genes obtenidas mediante el paquete *Boruta* en todos los estudios de un mismo grupo de respuesta: OR *vs.* NR, RC *vs.* Resto o para respuesta codificada en tres grupos. Esto no se pudo hacer con el grupo multirrespuesta ya que, como se indicó anteriormente, cada estudio dentro de esta aproximación tiene su propia codificación de la respuesta.

### 3.11.3. Algoritmos para la predicción de la respuesta a tratamiento

Para realizar los análisis de predicción de respuesta se utilizó una batería de cinco algoritmos sobre los cuatro regímenes de tratamiento estudiados, en los cuatro grupos de codificación de respuesta, con las tres listas de genes. Estos cinco algoritmos son:

a) Máquinas de vectores de soporte o *Support Vector Machines* (SVM)[363]: SVM es un método de clasificación de aprendizaje supervisado. Este método funciona mediante la búsqueda de uno o múltiples hiperplanos, que son "líneas" que separan y clasifican un conjunto de datos en un espacio multidimensional que permitan la división del conjunto de datos en dos clases. En este método tiene gran relevancia el concepto de vectores de soporte, que son los puntos de datos que se encuentran más cercanos a la superficie de decisión (hiperplano). La distancia entre el hiperplano y el vector de soporte es lo que se conoce como margen. Por tanto, el objetivo de este método sería encontrar el hiperplano con el mayor margen posible. La **Figura 3.10** muestra un resumen gráfico de estos parámetros.

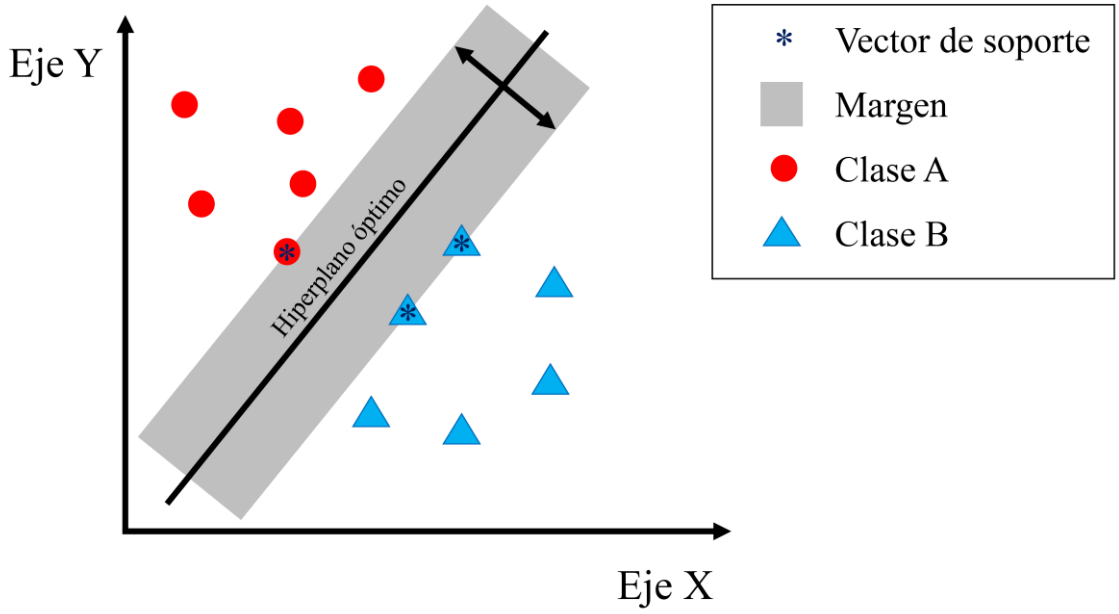



***Figura 3.10.*** *Representación de los parámetros considerados por el método SVM. Adaptado del tutorial de Avinash Navlani "**Support Vector Machines with Scikit-learn**" de la comunidad DataCamp (https://www.datacamp.com/community/tutorials).*

SVM fue ejecutado en R a través del paquete *e1071* (v.1.7-0)[409], utilizando para la predicción de resultados sobre la serie de entrenamiento validación cruzada dejando uno fuera o Leave-One-Out Cross-Validation (LOOCV). Sin embargo, previamente a la ejecución de SVM hubo que determinar los valores de la función kernel, y de los parámetros coste y gamma, ya que todos ellos son clave a la hora de realizar una predicción con SVM. La función kernel es una solución al problema de clasificación de dos grupos cuando estos no pueden ser separados por una línea recta en un espacio bidimensional, esto se conoce como *kernel trick*,

y consiste en llevar los datos a un espacio de una dimensionalidad mayor donde los datos sí puedan ser linealmente separados mediante un plano (**Figura 3.11**). La función kernel óptima para cada análisis de SVM fue determinada mediante el paquete *OptimClassifier* (v.0.1.4)[410], seleccionando entre las funciones kernel lineal, radial, polinomial o sigmoidal.

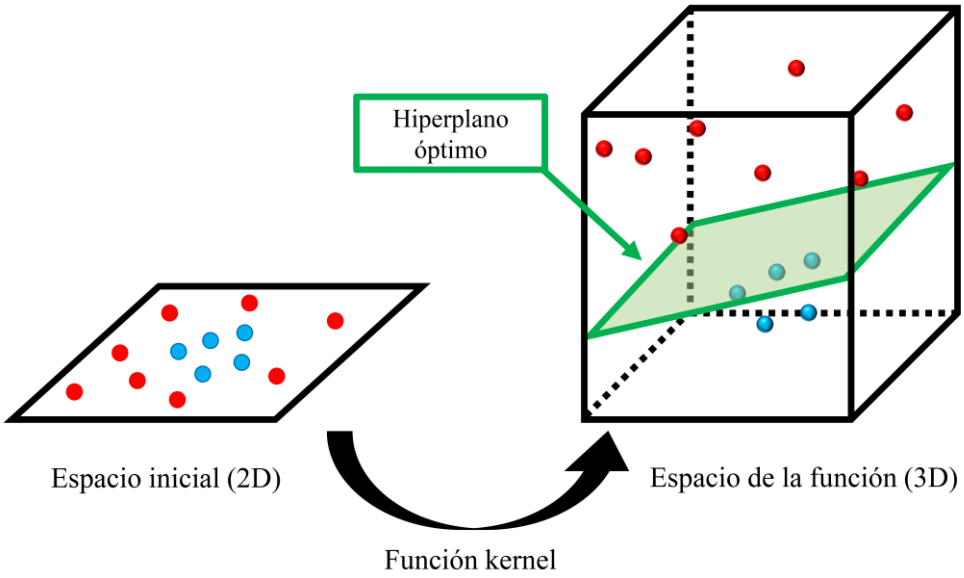



***Figura 3.11.*** *Representación gráfica del método kernel trick. Los círculos corresponden a las muestras a clasificar mediante SVM. En rojo las muestras correspondientes a la clase A y en azul a la clase B.*

El parámetro coste, del inglés *cost* (C), por su parte, controla la influencia de cada vector de soporte. Si este parámetro toma valores altos, la optimización del SVM elegirá un hiperplano de menor margen. Sin embargo, si los valores que toma son bajos, la optimización tomará hiperplanos de mayor tamaño, lo que aumentará el error en la serie de entrenamiento, pero también la robustez del modelo. Este parámetro supone por tanto una penalización de los puntos que se encuentran en el margen.

El parámetro gamma ($\gamma$) es un parámetro de la función kernel. Este parámetro ajusta la similitud que deben tener dos puntos de datos para ser considerados iguales. Mayores valores de gamma implican una mayor desviación estándar entorno a cada punto.

Ambos parámetros fueron ajustados en la serie de entrenamiento para encontrar sus valores óptimos mediante la opción *tune* del paquete *e1071*. La representación de su influencia sobre la clasificación de las muestras se recoge en la **Figura 3.12**.

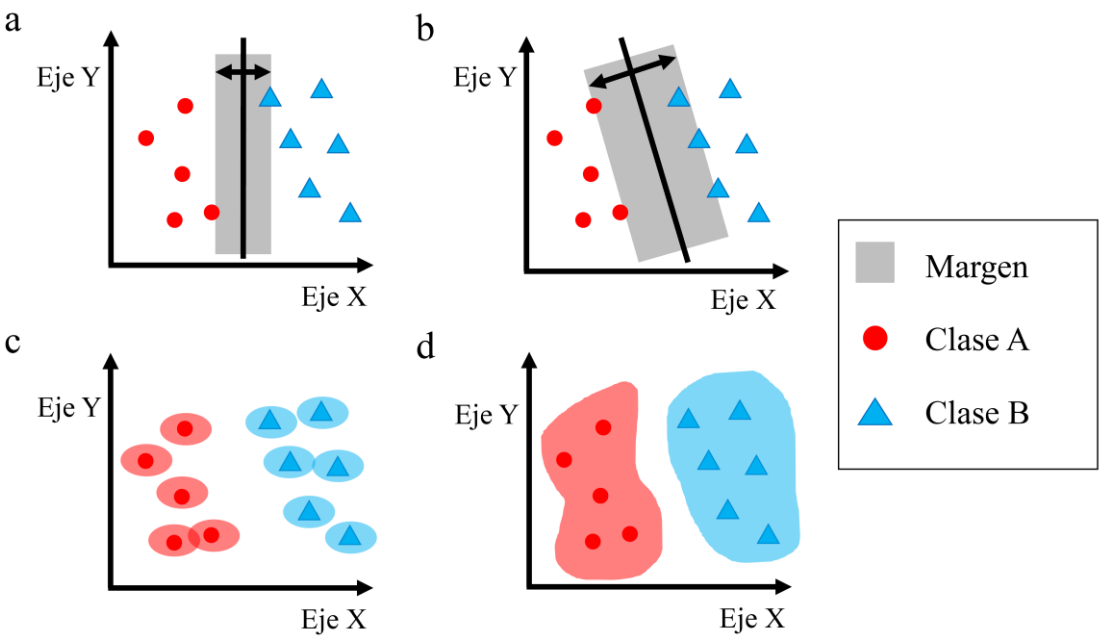



***Figura 3.12.*** *Representación de la influencia de los parámetros coste y gamma en la clasificación mediante SVM.* ***a)*** *clasificación con coste alto,* ***b)*** *clasificación con coste bajo,* ***c)*** *clasificación con gamma alto y* ***d)*** *clasificación con gamma bajo.*

b) Máquinas de vectores de soporte ponderada o *weighted Support Vector Machines* (wSVM). Las características y la ejecución del wSVM son similares a las expuestas para el SVM, con la diferencia de que para este método se llevó a cabo la asignación de pesos diferentes a los grupos a clasificar en función del número de muestras de que disponían. Los pesos se asignaron en base a la siguiente fórmula:

$$Peso_i = \frac{N}{k * n_i}$$

Donde, *N* es el número total de muestras analizadas en la matriz de entrenamiento, *k* es el número de grupos de respuesta estudiados y $n_i$ es el número de muestras en la matriz de entrenamiento para el grupo *i*.

c) Mínimos cuadrados parciales o *Partial Least Squares* (PLS)[411]: el método PLS funciona extrayendo un conjunto de factores latentes que son capaces de explicar la mayor parte de la covarianza existente entre las variables predictoras y la variable respuesta de tipo dicotómica en nuestras predicciones. Así, PLS está basado en la utilización de dos matrices: una primera matriz **X** de *n* muestras por *m* variables predictoras (***n*m***), y una segunda matriz **Y** de *n* muestras por *r* variables respuesta (***n*r***). De acuerdo con el procedimiento expuesto por Bardsley[337], si $\mathbf{X_1}$ es la matriz centrada obtenida de **X**, e $\mathbf{Y_1}$ es la matriz obtenida de **Y**, el primer factor se obtiene mediante regresión en un vector de *n* puntuaciones normalizadas $t_1$:

$$t_1 = X_1 w_1$$

donde $w_1$ son los pesos para las variables predictoras. Los cálculos para las estimaciones de $X_1$ e $Y_1$ se realizaron como:

$$\hat{X}_1 = t_1 p_1^T$$

$$\hat{Y}_1 = t_1 c_1^T$$

$$t_1^T t_1 = 1$$

donde *T* denota la transposición del vector, y los vectores columna $p_1$ de *m* cargas *x* y $c_1$ de *r* cargas *y* son calculados mediante mínimos cuadrados:

$$p_1^T = t_1^T X_1$$

$$c_1^T = t_1^T Y_1$$

Todas las predicciones con PLS se realizaron considerando un máximo de 12 factores y un mínimo de un factor. Se utilizaron dos criterios para determinación de la correcta clasificación de las muestras sobre la serie de validación:

1) Clasificación de respuestas binarias: se determinó como clasificación correcta aquellas muestras cuyo $\hat{Y}$ fuese superior o igual a 0,5 en las muestras del grupo 1, o inferior a 0,5 en las muestras del grupo 0.
2) Clasificación de respuestas en tres o más grupos: se determinó como clasificación correcta aquella en la que el grupo de respuesta predicho con el mayor valor del $\hat{Y}$ coincidiese con el grupo de respuesta real.

Este método se ejecutó en el paquete estadístico SIMFIT utilizando los módulos PLS, en el caso de que la respuesta estuviese codificada en tres o más grupos, y PLSVIP[366], cuando se analizaron respuestas dicotómicas. Adicionalmente, las variables predictoras fueron clasificadas en función de su influencia en la variable respuesta a través de la puntuación VIP (del inglés *Variance Influence in Projection*). El valor de la puntuación VIP para cada variable predictora $\ell$, viene dado por la ecuación:

$$VIP_\ell = \sqrt{m * \frac{\sum_{j=1}^{r}\sum_{i=1}^{k} \frac{w_{\ell,i}^2 \left(YCV_{i,j} - YCV_{i-1,j}\right)}{YCV_{k,j}}}{r}}$$

donde *i* representa el índice del factor latente, *k* es el número de factores, *j* es el índice que hace referencia a las variables respuesta "y", *m* es el número de variables predictoras $\ell$, $w_{\ell,i}$ es el peso de las variable predictora $\ell$ para cada factor *i*, $YCV_{i,j}$ es la variable respuesta *j* para el factor *i* y *r* es el número de variables "y".

d) K vecinos más cercanos o *K-Nearest Neighbors* (KNN)[364]: KNN es un método de clasificación de aprendizaje supervisado. Este algoritmo busca, para cada uno de

los puntos que se quieren clasificar, los k puntos más cercanos, y toma las clases con una ocurrencia más frecuente entre estos, asignando esta clase al punto que se quiere clasificar. Este algoritmo fue ejecutado mediante el paquete *caret* (v.6.0-80)[412] en R, realizando el preprocesamiento de la matriz de entrenamiento mediante la centralización y el escalado de los datos. La determinación del número de k-vecinos óptimo fue determinada utilizando como métrica la curva COR de los distintos modelos ajustados utilizando distintos valores de k.

e) Bosques aleatorios o *Random Forest* (RF)[365]: RF es un algoritmo de aprendizaje supervisado. Este algoritmo se basa en la generación y combinación de un conjunto de árboles de decisión para crear un modelo predictivo, basado en el promedio de las predicciones de cada árbol de decisión (**Figura 3.13**). Este algoritmo fue ejecutado a través del paquete *caret* en R de forma similar a lo indicado en KNN.

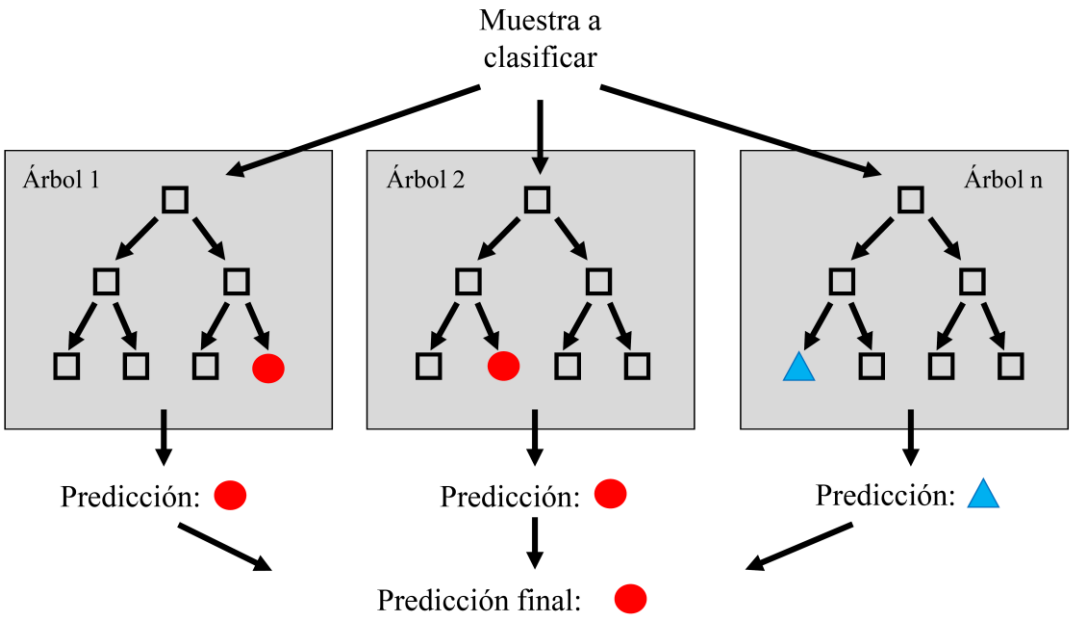



***Figura 3.13.*** *Ejemplo de predicción de clases utilizando el algoritmo Random Forest.*

### 3.11.4. Determinación de la bondad de la predicción

La medida de la bondad en los análisis de predicción se llevó a cabo considerando los resultados obtenidos en las matrices de validación. La función de la matriz de entrenamiento en este trabajo fue exclusivamente la del establecimiento del modelo de predicción que fue aplicado sobre la matriz de validación. La determinación de la bondad de los diferentes modelos ajustados se evaluó mediante cinco parámetros en las predicciones con variable respuesta dicotómica

1) Tasa de acierto global (TAC): el valor de la TAC corresponde al número de respuestas acertadas frente al total de muestras estudiadas en la matriz de validación.
2) Razón de verdaderos positivos (RVP)
3) Razón de verdaderos negativos (RVN)
4) Valor predictor positivo (VPP)
5) Valor predictor negativo (VPN)

Estos cinco parámetros han sido previamente descritos en el Apartado 3.2.5, siendo el TAC equivalente al parámetro ACC allí recogido.

Finalmente, en el caso de las predicciones en las que la variable respuesta constó de tres o más grupos, se evaluaron tanto la TAC como la tasa de acierto en cada uno de los grupos de respuesta estudiados.

## 3.12. Análisis de sobrerrepresentación de genes en rutas biológicas

El análisis de sobrerrepresentación de genes (ORA) fue ejecutado a través de la herramienta online Webgestalt (http://www.webgestalt.org/option.php)[413-415]. Para el análisis de rutas biológicas se utilizó la base de datos KEGG (versión del 1 de octubre de 2016)[416-418], mientras que para los análisis de procesos biológicos (PB), funciones moleculares (FM) y componente celular (CC) se utilizó la base de datos Gene Ontology (versión del 24 de octubre de 2016)[419]. Se utilizó como identificador génico el identificador ENSG de Ensembl y como grupo de genes de referencia el conjunto de genes del genoma completo que son codificantes de proteínas. El número mínimo de genes permitido por cada categoría funcional fue de dos, mientras que el máximo fue de 2000. En todos los análisis se reportaron las 10 categorías funcionales (rutas biológicas, procesos biológicos, funciones moleculares o componentes de la célula) con menor FDR. Los gráficos de las rutas biológicas KEGG fueron representados con la herramienta Pathview (https://pathview.uncc.edu/home)[420, 421].

## 3.13. Disponibilidad de datos

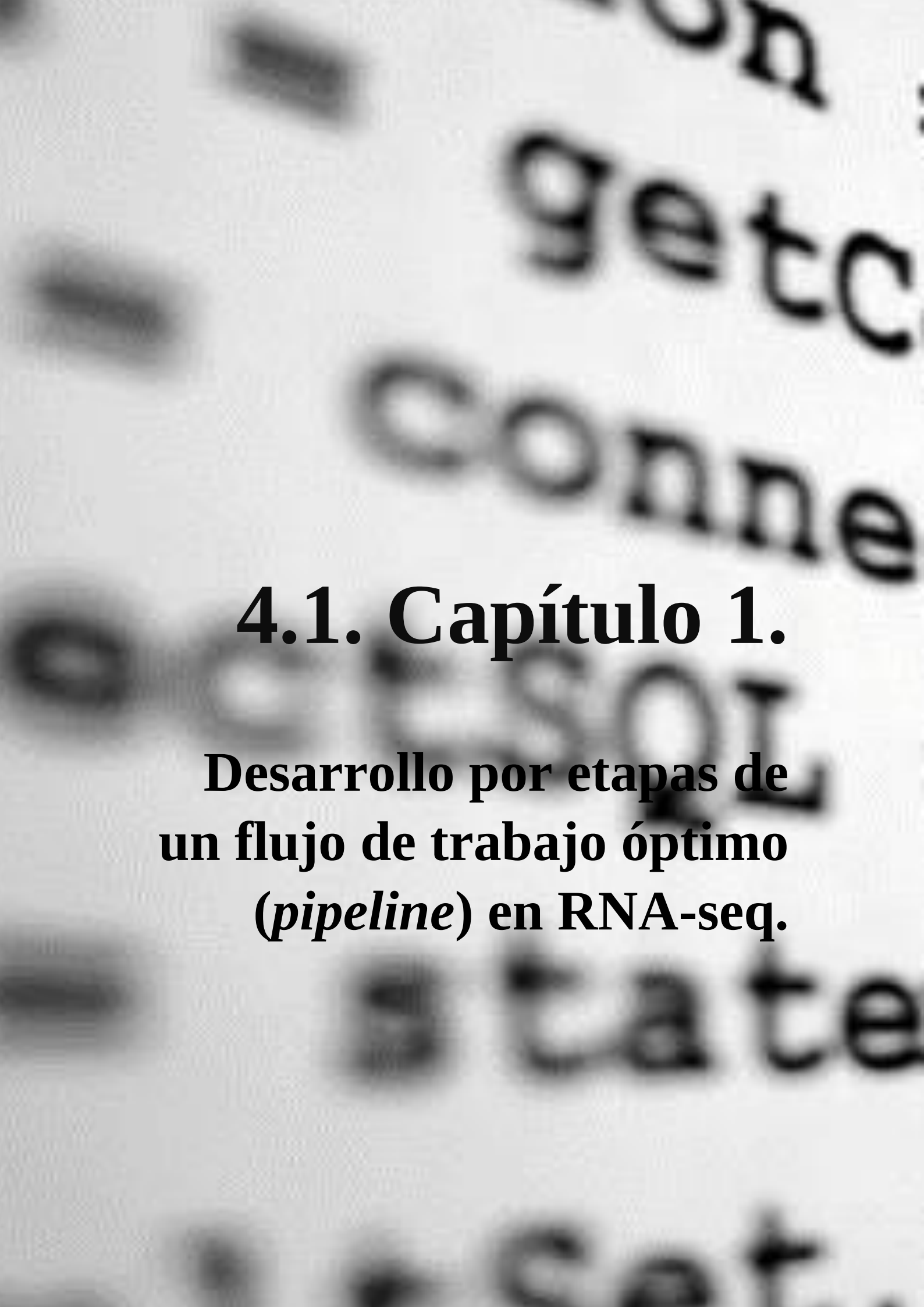
Los datos de RNA-seq y microarray exclusivos de este trabajo y del laboratorio de Hematología de Salamanca fueron depositados en la plataforma Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gob/geo/) con los identificadores de acceso GSE95077 y GSE116291.

Los *scripts* necesarios para ejecutar los programas utilizados en este trabajo pueden ser consultados en el **Anexo 1** y online en Github en la dirección https://github.com/lacorsan/.

Todos los archivos anexos al presente trabajo pueden ser descargados de Dropbox en:

https://www.dropbox.com/sh/0fcenlc786wwca7/AADXXHuAmijYrhEEhrOUdeNVa?dl=0

# 4. Resultados y discusión



# 4.1. Capítulo 1.

## Desarrollo por etapas de un flujo de trabajo óptimo (*pipeline*) en RNA-seq.

Actualmente, la RNA-seq ha emergido como un método alternativo a los microarrays para el análisis de la expresión génica. Se considera una metodología más potente y adaptable que los microarrays. Esta técnica además ofrece ventajas añadidas, permitiendo la detección de nuevos transcritos, fusión de genes y detección de polimorfismos. Sin embargo, el análisis de la RNA-seq es sumamente complejo y como consecuencia la investigación en algoritmos y flujos de trabajo para su análisis ha sido muy extensa. Esto ha resultado en un incremento incesante en el número de opciones de análisis conduciendo a la falta de consenso en cuanto a la determinación de los métodos y algoritmos más adecuados. Por este motivo, se propone a continuación una guía para el análisis de datos procedentes de RNA-seq basada en el estudio de 192 flujos de trabajo o *pipelines* para la determinación de la expresión génica en bruto y 17 métodos correspondientes a 11 algoritmos para el análisis de la expresión génica diferencial. El resumen de los métodos evaluados se recoge en la **Figura 4.1**.

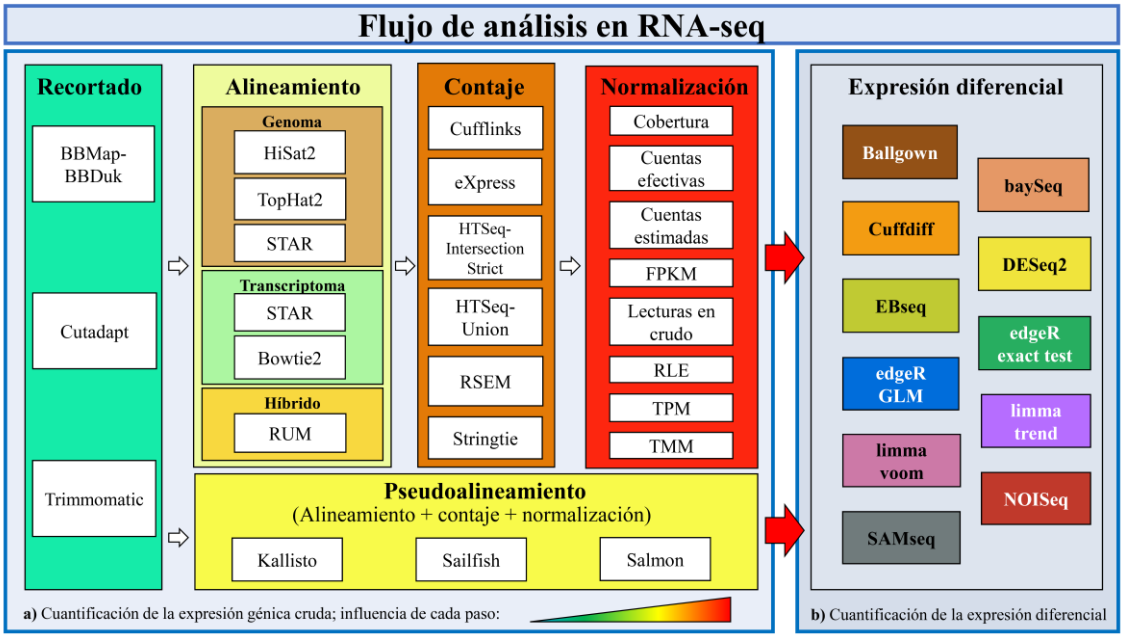



***Figura 4.1**. Flujo de análisis de RNA-seq. El panel **a)** representa el flujo de trabajo seguido para la determinación de la expresión génica cruda. Cada caja contiene los algoritmos utilizados en este trabajo a los niveles de recortado de lecturas, alineamiento, contaje, normalización y pseudoalineamiento. La coloración de cada caja representa la influencia de cada etapa del flujo de análisis sobre la expresión génica cruda. El panel **b)** contiene los algoritmos utilizados en este trabajo para la determinación de la expresión génica diferencial.*

### 4.1.1. Cuantificación de la expresión génica cruda

El primer objetivo de este trabajo fue el desarrollo de un *pipeline* para la determinación de la expresión génica cruda a partir de datos en bruto de RNA-seq. La expresión génica cruda se define en este trabajo como la cantidad de ARNm de cada gen en cada una de las muestras estudiadas estimada a partir del número de lecturas de RNA-seq asociadas a dicho gen. Este parámetro puede ser expresado en diferentes unidades en

función de la metodología empleada para su obtención, de modo que es posible utilizar diferentes unidades, como lecturas crudas, FPKM, TPM, etc. (**Apartado 3.2.3 en la Sección de Material y métodos**) para referirse a la expresión de un mismo gen.

#### 4.1.1.1. Evaluación de los métodos y algoritmos de análisis de expresión génica cruda

En este Apartado se determinó el desempeño de los métodos y algoritmos empleados en cada una de las etapas del análisis de expresión génica cruda de RNA-seq. Esta evaluación se realizó considerando los resultados de precisión y exactitud obtenidos *a posteriori* por cada uno de los *pipelines* en los que participó cada uno de los métodos empleados. Además, se evaluaron en cada etapa algunos parámetros adicionales que son considerados de interés crucial para determinar el rendimiento de cada uno de los métodos.

***Algoritmos de recortado de lecturas***

El rendimiento de los algoritmos de recortado fue valorado a través del análisis del número de lecturas supervivientes, la ratio de alineamiento, la precisión y exactitud de los pipelines en que cada método estuvo implicado y el tiempo de ejecución.

El primer parámetro evaluado fue el efecto de los algoritmos de recortado sobre el número de lecturas supervivientes, donde *Cutadapt* fue el algoritmo que obtuvo un mayor número de lecturas pareadas supervivientes con un 95,5%, frente al 89,8% y el 83,8% de *BBDuk* y *Trimmomatic*, respectivamente (**Tabla 4.1**). Las diferencias entre los tres algoritmos fueron sancionadas mediante una prueba de Kruskal-Wallis ($p$-valor < 0,001) seguida de la prueba *post-hoc* de Dunn, mostrando este último test diferencias estadísticamente significativas en todas las comparaciones pareadas (FDR < 0,05) (**Anexo 2**).

***Tabla 4.1**. Porcentaje de lecturas supervivientes en cada una de las seis muestras control (T0) analizadas de las líneas celulares KMS12-BM (LCA) y JJN-3 (LCB).*

| Muestra | Número de pares de lecturas de inicio | Lecturas supervivientes (%) | | |
|---|---|---|---|---|
| | | Trimmomatic | Cutadapt | BBDuk |
| LCA-T0-M1 | 50.946.237 | 83,5 | 95,9 | 92,7 |
| LCA-T0-M2 | 62.571.514 | 84,4 | 95,0 | 88,8 |
| LCA-T0-M3 | 77.906.369 | 84,2 | 95,4 | 89,6 |
| LCB-T0-M1 | 46.810.430 | 88,9 | 96,8 | 93,9 |
| LCB-T0-M2 | 60.376.359 | 79,4 | 94,5 | 85,6 |
| LCB-T0-M3 | 40.840.185 | 83,2 | 95,6 | 89,9 |
| **Mediana LCA** | **62.571.514** | **84,2** | **95,4** | **89,6** |
| **Mediana LCB** | **46.810.430** | **83,2** | **95,6** | **89,9** |
| **Mediana Global** | **55.661.298** | **83,8** | **95,5** | **89,8** |

En un segundo paso se procedió a determinar la influencia del recortado de lecturas sobre la siguiente etapa de análisis de la RNA-seq: el proceso de alineamiento. Para ello se calculó la ratio de alineamiento mediana de los tres algoritmos de recortado, considerando las ratios de los cinco algoritmos de alineamiento empleados en este trabajo. El algoritmo que obtuvo una mejor ratio mediana de alineamiento fue *BBDuk*, con un 97,5% (MAD = 1,2) de lecturas alineadas, seguido de *Trimmomatic* con un 96,1% (MAD = 1,2) (**Tabla 4.2**). Las ratios obtenidas por ambos algoritmos no presentaron diferencias estadísticamente significativas (FDR = 0,305), sin embargo, sí se detectaron diferencias al compararse ambas ratios contra el algoritmo con menor ratio de alineamiento, *Cutadapt* (mediana = 93,4%, MAD = 2,5), obteniendo un FDR = 0,001 en el caso de *BBDuk* y FDR = 0,002 en el caso de *Trimmomatic* (**Anexo 3**).

***Tabla 4.2**. Ratio de lecturas mapeadas en función del algoritmo de recortado considerando las seis muestras control (T0) analizadas de las líneas celulares KMS12-BM (LCA) y JJN-3 (LCB).*

| Muestra | Ratio de alineamiento (%) | | |
|---|---|---|---|
| | Trimmomatic | Cutadapt | BBDuk |
| LCA-T0-M1 | 94,9 ± 0,4 | 90,5 ± 3,1 | 97,0 ± 1,2 |
| LCA-T0-M2 | 96,9 ± 0,8 | 95,0 ± 1,9 | 98,0 ± 0,7 |
| LCA-T0-M3 | 96,8 ± 1,6 | 94,0 ± 2,1 | 97,9 ± 1,1 |
| LCB-T0-M1 | 96,1 ± 0,5 | 94,0 ± 2,4 | 97,6 ± 0,8 |
| LCB-T0-M2 | 95,8 ± 1,2 | 93,8 ± 2,5 | 97,3 ± 1,4 |
| LCB-T0-M3 | 96,0 ± 1,1 | 93,1 ± 2,6 | 97,8 ± 0,8 |
| **Mediana LCA** | **96,1 ± 1,4** | **93,1 ± 2,6** | **97,4 ± 1,3** |
| **Mediana LCB** | **96,0 ± 1,0** | **93,8 ± 2,5** | **97,6 ± 1,0** |
| **Mediana Global** | **96,1 ± 1,2** | **93,4 ± 2,5** | **97,5 ± 1,2** |

*Los porcentajes recogidos en esta tabla aparecen representados como la mediana ± desviación absoluta de la mediana (MAD)*

A continuación, se realizó la exploración del efecto de los algoritmos de recortado sobre la cuantificación de la expresión génica cruda. Para realizar este análisis se consideraron los rankings de la precisión y de la exactitud obtenidos por los *pipelines* en los que participa cada uno de los algoritmos de recortado. Como se muestra en la **Figura 4.2a**, tras realizar una prueba de Kruskal-Wallis seguida de la prueba *post-hoc* de Dunn, no se encontraron diferencias estadísticamente significativas (FDR > 0,05) entre los rankings de los *pipelines* en los que intervenía cada uno de los tres algoritmos de recortado evaluados. En este mismo gráfico, se observa que en cada una de las cajas aparece una distribución bimodal de los puntos que representan los diferentes *pipelines*. De forma interesante, los valores que se agrupan en torno al final del tercer cuartil en estas cajas, es decir, los valores que tienen un peor ranking son los correspondientes a los *pipelines* que utilizan como metodologías de normalización técnicas como las cuentas estimadas, las cuentas efectivas, la cobertura o no utilizan ninguna metodología de normalización (lecturas crudas). La presencia de estos métodos no debería traducirse en un sesgo del resultado final sobre la evaluación de los algoritmos de recortado, ya que estuvieron homogéneamente repartidos entre estos tres algoritmos. No obstante, se procedió a la reevaluación de la etapa del recortado descartando los métodos de normalización mencionados, y nuevamente no se encontraron diferencias estadísticamente significativas entre los *pipelines* en función del método de recortado empleado (**Figura 4.2b**).

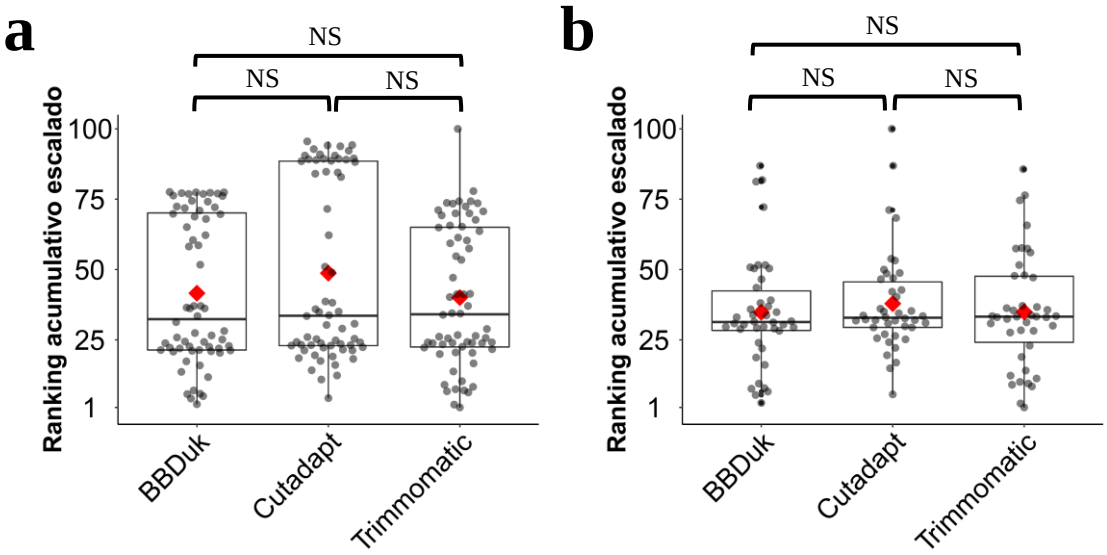



***Figura 4.2**. Influencia de los tres algoritmos de recortado sobre el ranking final en el estudio de la expresión génica cruda mediante RNA-seq. **a)** Diagrama de cajas considerando los 192 pipelines. **b)** Diagrama de cajas descartando los pipelines cuyo método de normalización fuesen cuentas estimadas, cuentas efectivas, cobertura o lecturas en crudo. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline. NS = No significativo (FDR en la prueba post-hoc de Dunn > 0,05)*

Finalmente, se procedió a la comparación de los tiempos de ejecución de cada uno de los algoritmos de recortado que se han utilizado en el presente trabajo. Este análisis demostró una mayor eficiencia de *BBDuk* frente a los otros dos métodos, ya que fue 11,8 y 5,3 veces más rápido que *Cutadapt* y *Trimmomatic*, respectivamente (**Figura 4.3**).

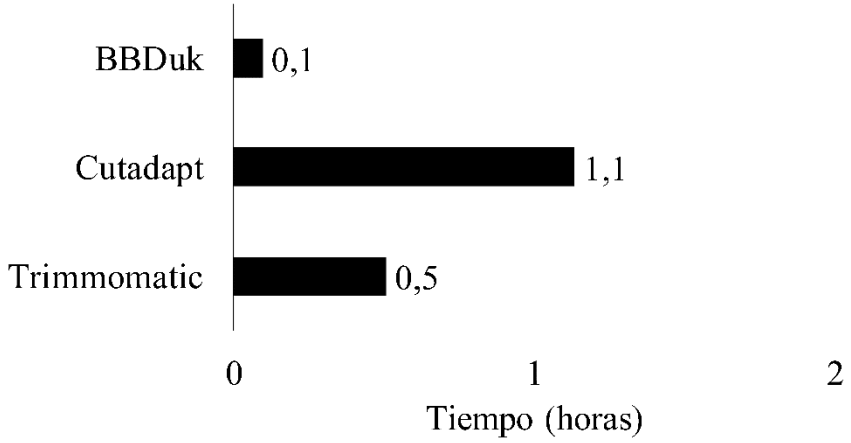



***Figura 4.3.*** *Tiempo de ejecución, en horas, de los tres algoritmos empleados en este trabajo para el proceso de recortado de lecturas.*

De este modo, pese al diferente resultado de cada uno de los algoritmos en cuanto a número de lecturas supervivientes y ratio de lecturas alineadas, se ha podido constatar que la elección del método no juega un papel relevante en la cuantificación de la expresión génica cruda. La falta de influencia del algoritmo elegido para el recortado sobre la cuantificación de la expresión génica cruda puede ser debida al modo en que este proceso ha sido efectuado en este trabajo, ya que se procuró que la ejecución se realizase de una manera poco agresiva y cuidadosa. De no haber sido así, los resultados obtenidos podrían haber experimentado sesgos en los valores finales de expresión génica, tal y como se ha apuntado previamente en el estudio de Williams y colaboradores[209].

***Algoritmos de alineamiento***

La evaluación inicial de estos algoritmos se realizó a través de la determinación del número de lecturas mapeadas de forma concordante en una única región del genoma de referencia y el número de lecturas no mapeadas. Como nota aclaratoria, se define como alineamiento concordante al par de lecturas que son mapeadas en el mismo *contig* o cromosoma con la orientación adecuada y con la distancia apropiada entre sus extremos. Este proceso se evaluó sobre 156 *pipelines* ya que los 36 *pipelines* restantes utilizaron métodos de pseudoalineamiento y su evaluación se realizará de forma separada. En lo que respecta a estas lecturas únicas concordantes, *STAR* fue el algoritmo con un mayor porcentaje de lecturas únicas alineadas de forma concordante (mediana = 93,3%, MAD = 1,0), seguido muy de cerca por *TopHat2* (mediana = 90,1%, MAD = 2,0), y de hecho, no se registraron diferencias estadísticamente significativas entre ambos métodos (FDR = 0,067). En el extremo opuesto se situó el algoritmo *Bowtie2*, con una mediana de 41,5%

(MAD = 1,7) de lecturas únicas concordantes mapeadas, presentando además diferencias estadísticamente significativas con los cuatro algoritmos restantes (FDR < 0,05 en todos los casos).

En cuanto a las lecturas no mapeadas, los algoritmos *HiSat2* y *STAR* tuvieron un comportamiento notablemente superior al resto de algoritmos ya que solamente fallaron en el mapeo del 1,6% (MAD = 0,3) y del 3,0% (MAD = 0,8) de las lecturas, respectivamente. A pesar del buen comportamiento de los dos métodos, la prueba *post-hoc* de Dunn reveló diferencias estadísticamente significativas entre ambos (FDR = 0,035). Por otro lado, *Bowtie2* fue el algoritmo con un mayor porcentaje de lecturas no mapeadas, ya que no consiguió alinear el 15,8% de las lecturas (MAD = 0,6) contra la correspondiente referencia genómica (**Figura 4.4**). Este último algoritmo presentó además diferencias estadísticamente significativas (FDR < 0,05) en el porcentaje de lecturas no mapeadas cuando se comparó con los cuatro algoritmos restantes empleados en este trabajo.

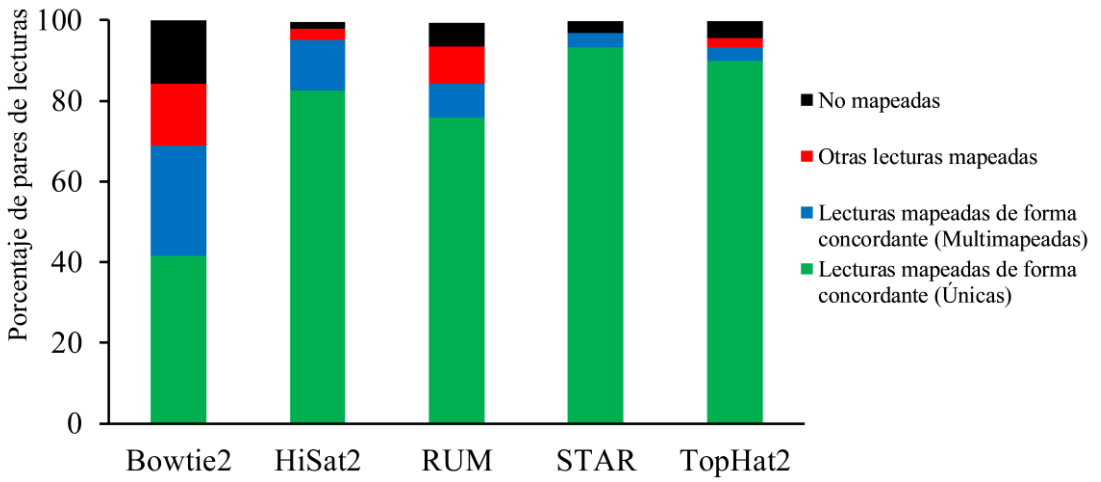



***Figura 4.4.*** *Mediana del porcentaje de pares de lecturas mapeadas por los cinco algoritmos de alineamiento analizados en este trabajo. El número de pares de lecturas alineadas se representa como porcentaje.*

En un segundo paso se investigó la influencia de los algoritmos de alineamiento sobre la cuantificación de la expresión génica cruda. Este análisis se llevó a cabo considerando la mediana de los rankings de precisión y exactitud que ocuparon los *pipelines* en los que estuvieron implicados cada uno de los cinco algoritmos de alineamiento. Como se muestra en la **Figura 4.5a,** no se observaron diferencias estadísticamente significativas en la mediana de los rankings ocupados por estos algoritmos. Sin embargo, cuando se procedió al agrupamiento de estos cinco algoritmos en función de la referencia frente a la que se realizó el alineamiento, se encontraron diferencias estadísticamente significativas entre los *pipelines* en los que se utilizó una referencia transcriptómica respecto a los que utilizaron una referencia genómica o una

referencia híbrida de genoma y transcriptoma (*p*-valor de la prueba de Kruskal-Wallis = 0,01; FDR de la prueba de Dunn = 0,008 and 0,014, respectivamente) (**Figura 4.5c**). No obstante, como ya se apuntó en el caso de los algoritmos de recortado, examinando detenidamente los gráficos nos volvemos a encontrar con una distribución bimodal, que se asocia al proceso de normalización aplicado en los 156 *pipelines*, con lo que las diferencias detectadas entre la referencia de alineamiento utilizada podrían deberse a la mayor prevalencia de *pipelines* que utilizan métodos de normalización como cuentas efectivas, cuentas estimadas o lecturas en crudo en el grupo que utilizó referencia transcriptómica. Para descartar o confirmar esta hipótesis se procedió a la reevaluación de los algoritmos y referencias de mapeo sin considerar los *pipelines* cuyo método de normalización fueron cuentas estimadas, cuentas efectivas, cobertura o lecturas en crudo. Este reanálisis reveló en el caso de los algoritmos de alineamiento (**Figura 4.5b**) diferencias estadísticamente significativas entre *Bowtie2* y *STAR* (FDR = 0,039) a favor de este último. No se detectaron diferencias estadísticamente significativas entre otras parejas de algoritmos. En lo que respecta al reanálisis de la referencia de alineamiento, confirmó la tendencia observada con los 156 *pipelines*, de manera que los *pipelines* cuyo alineamiento fue realizado contra el transcriptoma obtuvieron peores rankings que los realizados contra genoma o contra una referencia híbrida (**Figura 4.5d**).

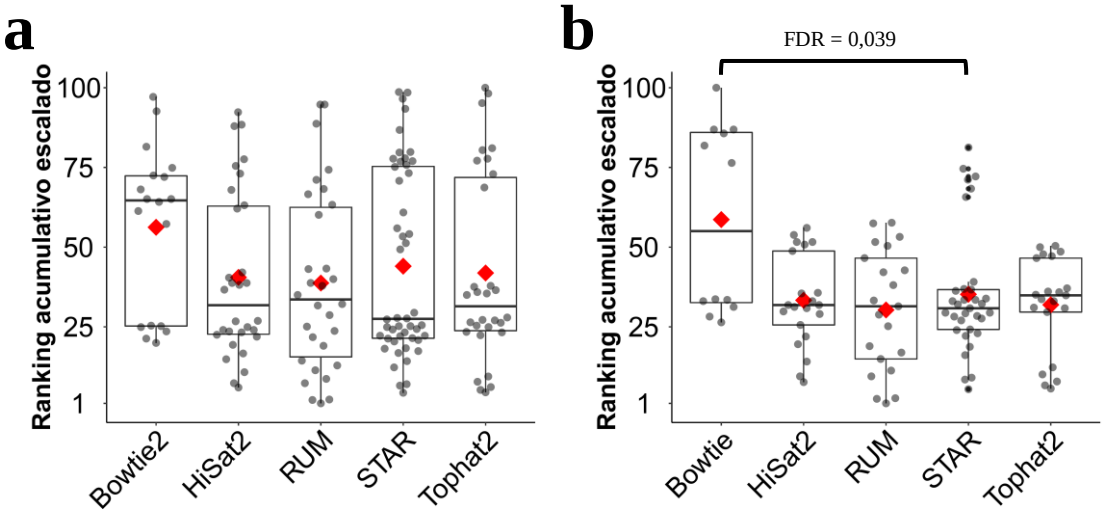



***Figura 4.5.*** *Influencia de los algoritmos de alineamiento sobre los rankings finales de la cuantificación de la expresión génica cruda.* ***a)*** *Influencia de los cinco algoritmos de alineamiento considerando 156 pipelines (excluidos los que utilizaron métodos de pseudoalineamiento).* ***b)*** *Influencia de los cinco algoritmos de alineamiento descartando los pipelines cuyo método de normalización fue cuentas estimadas, cuentas efectivas, cobertura o lecturas en crudo. Solamente se indican las comparaciones estadísticamente significativas a FDR < 0,05 en el test de Dunn. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline.*

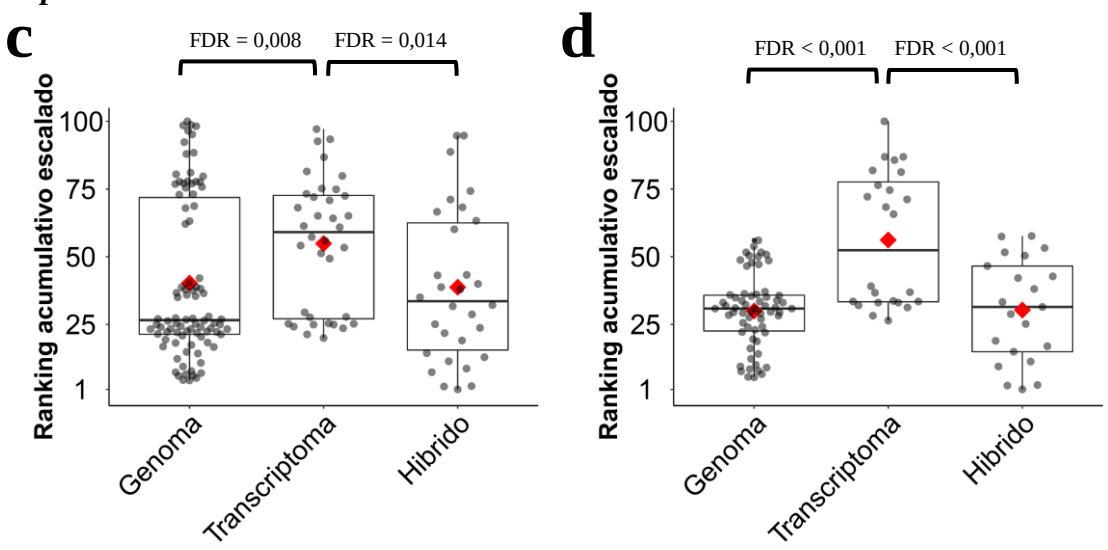



***Figura 4.5 (continuación).*** *Influencia de los algoritmos de alineamiento sobre los rankings finales de la cuantificación de la expresión génica cruda.* ***c)*** *Influencia en función de la referencia contra la que se ha realizado el alineamiento considerando los 156 pipelines.* ***d)*** *Influencia en función de la referencia contra la que se ha realizado el alineamiento descartando los pipelines cuyo método de normalización fue cuentas estimadas, cuentas efectivas, cobertura o lecturas en crudo. Solamente se indican las comparaciones estadísticamente significativas a FDR < 0,05 en el test de Dunn. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline.*

Teniendo en cuenta estos resultados, podría concluirse que existe una mayor dependencia sobre la bondad del alineamiento de la referencia que haya sido elegida para realizar el mapeo, que del propio algoritmo de alineamiento. Sin embargo, en una revisión detallada de los resultados se observó que los *pipelines* que usaron una referencia transcriptómica localizados en el tercer cuartil correspondieron a un único método de contaje, *eXpress*, que como se discutirá en el apartado de los métodos de contaje tuvo un rendimiento pobre en comparación con el resto de los métodos empleados, mientras que el resto de *pipelines* con referencia transcriptómica obtuvieron rankings similares a los que utilizaron referencias genómica o híbrida.

Finalmente, en lo concerniente a los tiempos de ejecución, *STAR* fue el algoritmo que completó el proceso de alineamiento en un tiempo menor, de modo que fue 1,4 y 50 veces más rápido que los algoritmos que quedaron en segunda (*Hisat2*) y última posición (*TopHat2*), respectivamente. (**Figura 4.6**).

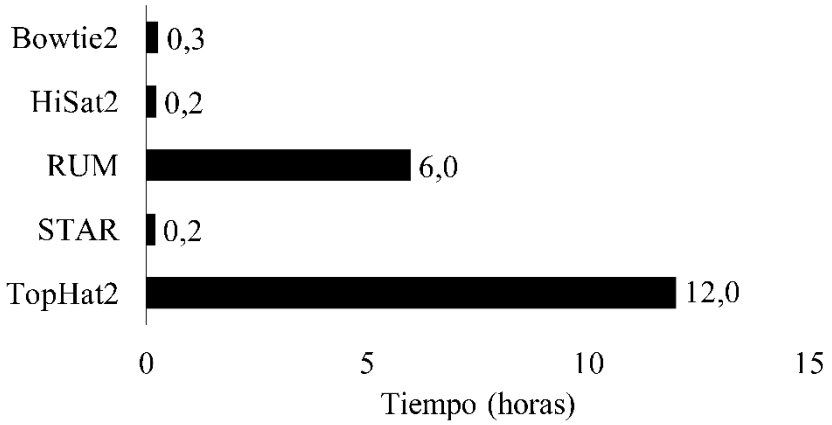



***Figura 4.6.*** *Tiempo de ejecución, en horas, de los cinco algoritmos empleados en este trabajo para el proceso de alineamiento o mapeado de lecturas.*

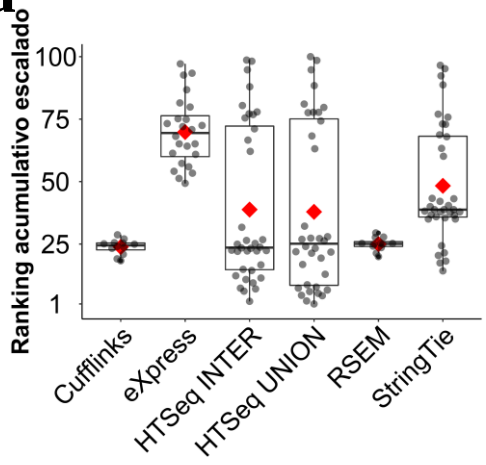
***Métodos de contaje***

Tras la evaluación del proceso de alineamiento de lecturas, se llevó a cabo el análisis de la influencia de los métodos de contaje sobre la cuantificación de la expresión génica en crudo. Este análisis se realizó mediante la evaluación de la mediana de los rankings que ocuparon los *pipelines* en los que estuvo implicado cada uno de los métodos de contaje estudiados, y únicamente sobre los 156 *pipelines* en los que no intervinieron métodos de pseudoalineamiento. De esta manera, se observó que los *pipelines* basados en *Cufflinks* y *RSEM* fueron los que alcanzaron mejores rankings en promedio, seguidos por los *pipelines* basados en *HTseq* y *Stringtie* (**Figura 4.7a y 4.7b**). Sin embargo, esta clasificación podría estar sesgada debido al hecho de que *HTSeq* y *Stringtie*, a diferencia de *Cufflinks* y *RSEM*, contienen *pipelines* basados en métodos de normalización como son las cuentas estimadas, cuentas efectivas, cobertura y lecturas en crudo, que se sitúan en el tercer y cuarto cuartil generando como ya se indicó en apartados anteriores una distribución bimodal. Para comprobar la existencia o no de este sesgo, se procedió a la reevaluación de los métodos de contaje descartando los *pipelines* indicados. Como se puede observar en la **Figura 4.7c**, el sesgo generado por los *pipelines* en los que intervinieron los mencionados métodos de normalización fue evidente. En este nuevo análisis los pipelines que utilizaron *HTSeq* tanto en su versión *Intersection-Strict* (INTER) como en su versión *Union*, consiguieron los mejores promedios del ranking, no presentando diferencias estadísticamente significativas con *Cufflinks* (FDR = 0,211 y FDR = 0,209, respectivamente), pero sí con el resto de los métodos de contaje.

Merece una mención aparte el resultado obtenido por los *pipelines* basados en el algoritmo *eXpress*, ya que quedaron significativamente alejados del resto de algoritmos. Este comportamiento difiere del planteado por Teng y colaboradores[260], quienes proponen un comportamiento similar a algoritmos como *Cufflinks* o *RSEM*, aunque este último obtuvo mejores resultados en escenarios con datos simulados. Una de las posibles

causas de esta diferencia podría ser que *eXpress* únicamente es capaz de utilizar datos procedentes de un alineamiento contra referencia transcriptómica. Sin embargo, esta causa queda descartada ya que el caso de *eXpress* es similar al de *RSEM*, que también realiza únicamente el contaje de lecturas alineadas contra transcriptoma, y sin embargo *RSEM* tiene un rendimiento muy cercano a los métodos que ocupan las primeras posiciones del ranking. Otra posible explicación es que el método de cálculo de la expresión génica a partir de la expresión de transcritos, descrito en la **Sección de Material y métodos**, no haya funcionado debidamente produciendo resultados inapropiados, ya que para poder trabajar a nivel génico con el algoritmo *eXpress* es necesario colapsar los resultados obtenidos a nivel transcrito en un único identificador génico.

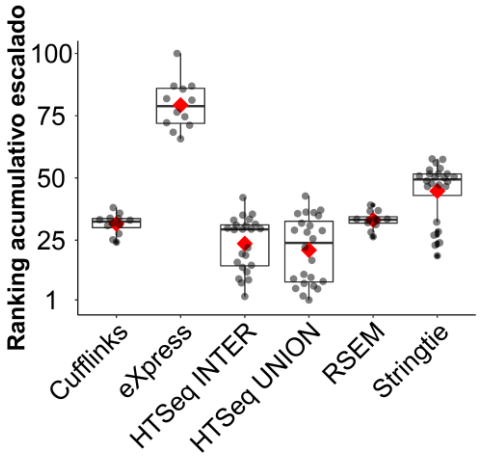
**a**




**b**

| FDR (test de Dunn) | Cufflinks | eXpress | HTSeq INTER | HTSeq UNION | RSEM | Stringtie |
|---|---|---|---|---|---|---|
| Cufflinks | | < 0,001 | 0,211 | 0,209 | 0,413 | 0,012 |
| eXpress | < 0,001 | | < 0,001 | < 0,001 | < 0,001 | 0,020 |
| HTSeq INTER | 0,211 | < 0,001 | | 0,470 | 0,0295 | 0,021 |
| HTSeq UNION | 0,209 | < 0,001 | 0, 470 | | 0,291 | 0,023 |
| RSEM | 0,413 | < 0,001 | 0,295 | 0,291 | | 0,021 |
| Stringtie | 0,012 | 0,020 | 0,021 | 0,023 | 0,021 | |

**c**




**d**

| FDR (test de Dunn) | Cufflinks | eXpress | HTSeq INTER | HTSeq UNION | RSEM | Stringtie |
|---|---|---|---|---|---|---|
| Cufflinks | | < 0,001 | 0,077 | 0,060 | 0,361 | 0,035 |
| eXpress | < 0,001 | | < 0,001 | < 0,001 | < 0,001 | < 0,001 |
| HTSeq INTER | 0,077 | < 0,001 | | 0,399 | 0,036 | < 0,001 |
| HTSeq UNION | 0,060 | < 0,001 | 0,399 | | 0,027 | < 0,001 |
| RSEM | 0,361 | < 0,001 | 0,036 | 0,027 | | 0,077 |
| Stringtie | 0,035 | < 0,001 | < 0,001 | < 0,001 | 0,077 | |

***Figura 4.7.*** *Influencia de los algoritmos de contaje sobre los rankings finales de la cuantificación de la expresión génica cruda.* ***a)*** *Considerando 156 pipelines (excluidos los que utilizaron métodos de pseudoalineamiento).* ***b)*** *Significancia estadística del test de Dunn correspondiente a las comparaciones por parejas del* ***panel a****.* ***c)*** *Sin considerar los pipelines cuyos métodos de normalización fueron cuentas estimadas, cuentas efectivas, cobertura o lecturas en crudo.* ***d)*** *Significancia estadística del test de Dunn correspondiente a las comparaciones por parejas del* ***panel c****. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline.*

Finalmente, en lo que respecta a los tiempos de ejecución, encontramos una alta variabilidad del tiempo empleado por cada método, ya que estos tiempos oscilaron entre los 10 minutos que llevó a *Stringtie* completar el análisis, y los 180 minutos que tardan los algoritmos *HTseq* y *RSEM*. (**Figura 4.8**)

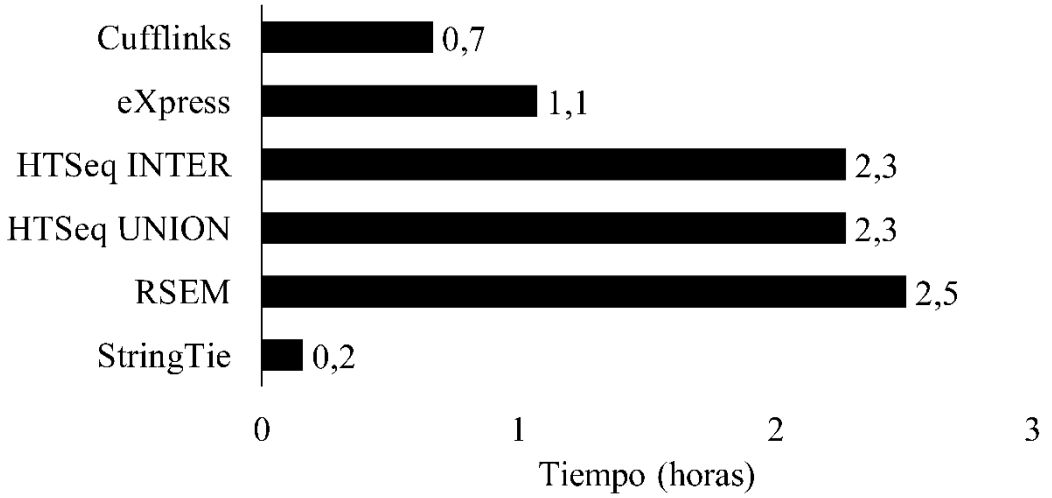



***Figura 4.8.*** *Tiempo de ejecución, en horas, de los seis métodos empleados en este trabajo para el proceso de contaje de lecturas.*

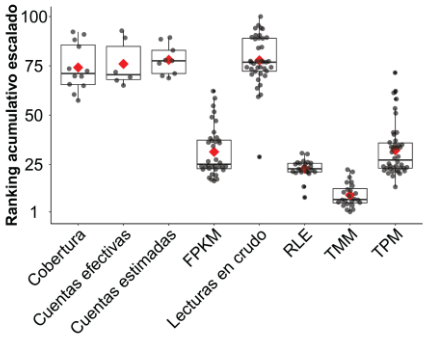
Los datos arriba expuestos apoyan la afirmación de que este paso del proceso de análisis de RNA-seq es uno de los más determinantes a la hora del cálculo de la expresión génica absoluta. Esto ha sido previamente puesto de manifiesto en el trabajo de Robert y colaboradores[422], quienes además demostraron que un mal procesamiento de los datos en este paso podría sobre o infraestimar la expresión génica, lo que convierte a esta etapa en un paso absolutamente decisivo en el resto del análisis posterior.

***Métodos de normalización***

El proceso de normalización en muchos casos suele estar asociado con el de contaje en cuanto a que el mismo paquete de análisis puede llevar a cabo ambas funciones. En el presente trabajo, además de utilizar las opciones provistas por los diferentes paquetes de contaje, se implementaron algunos de los métodos de normalización más populares en el análisis de RNA-seq, siempre que el acoplamiento de ambas técnicas fuese posible. Así, encontramos que los *pipelines* que utilizaron como método de normalización la media recortada de M valores o TMM mostraron un rendimiento superior de manera estadísticamente significativa (FDR < 0,05) al resto de métodos de normalización. Por otro lado, también observamos que los métodos que requirieron un tratamiento menos exhaustivo a nivel de normalización, como el método de las lecturas en crudo o *raw counts*, se comportaron significativamente peor que los métodos anteriormente mencionados (**Figura 4.9**). De este análisis puede extrapolarse al resto de las etapas del desarrollo del *pipeline* de RNA-seq la aparición de una distribución bimodal, ya que se

diferencian claramente dos grupos de *pipelines* en función de los métodos de normalización empleados. Así, se observa que, tal como se viene indicando en las anteriores etapas, los métodos de normalización cobertura, cuentas efectivas, cuentas estimadas y lecturas en crudo, presentan peores rankings que los métodos FPKM, RLE, TMM y TPM, con diferencias entre los métodos de ambos grupos estadísticamente significativas en todos los casos (**Figura 4.9b**).

**a**




**b**

| FDR (test de Dunn) | Cobertura | Cuentas efectivas | Cuentas estimadas | FPKM | Lecturas en crudo | RLE | TMM | TPM |
|---|---|---|---|---|---|---|---|---|
| Cobertura | | 0,487 | 0,426 | < 0,001 | 0,394 | < 0,001 | < 0,001 | < 0,001 |
| Cuentas efectivas | 0,487 | | 0,449 | 0,002 | 0,451 | < 0,001 | < 0,001 | 0,002 |
| Cuentas estimadas | 0,426 | 0,449 | | < 0,001 | 0,494 | < 0,001 | < 0,001 | < 0,001 |
| FPKM | < 0,001 | 0,002 | < 0,001 | | < 0,001 | 0,096 | < 0,001 | 0,480 |
| Lecturas en crudo | 0,394 | 0,451 | 0,494 | < 0,001 | | < 0,001 | < 0,001 | < 0,001 |
| RLE | < 0,001 | < 0,001 | < 0,001 | 0,096 | < 0,001 | | 0,005 | 0,073 |
| TMM | < 0,001 | < 0,001 | < 0,001 | < 0,001 | < 0,001 | 0,005 | | < 0,001 |
| TPM | < 0,001 | 0,002 | < 0,001 | 0,480 | < 0,001 | 0,073 | < 0,001 | |

***Figura 4.9**. Influencia de los métodos de normalización sobre los rankings finales de la cuantificación de la expresión génica cruda **a)** Diagrama de cajas de los 8 métodos de normalización. **b)** Resultados estadísticos correspondientes al FDR obtenido en el test de Dunn para los métodos recogidos en el **panel a**. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline.*

Por todo esto, el proceso de normalización se presenta, junto al de contaje, como uno de los pasos críticos en el análisis de RNA-seq. Observamos que los *pipelines* basados en TMM obtuvieron un comportamiento mejor que el resto de los métodos de normalización. Estos hallazgos están de acuerdo con lo previamente postulado por Wu y colaboradores y Maza y colaboradores[232, 233]. Por el contrario, estudios como el de Li y colaboradores[234] propusieron que los métodos de normalización no tendrían efecto sobre la determinación de la expresión génica, y que un método de normalización no mejoraría los resultados obtenidos a partir de las lecturas en crudo.

***Algoritmos de pseudoalineamiento***

En este trabajo además se decidió evaluar tres algoritmos de pseudoalineamiento, así como algunas de sus variantes. Estos algoritmos tienen la ventaja sobre los métodos de alineamiento tradicionales de que pueden llevar a cabo los pasos de contaje y normalización en una única etapa. Además, tienen una segunda ventaja que es su rápido tiempo de ejecución. En nuestro caso, los métodos de pseudoalineamiento se ejecutaron en un rango de tres a 9 veces más rápido que el algoritmo de alineamiento más rápido (*STAR*). En cuanto a la comparación de los tiempos de ejecución entre los propios métodos de pseudoalineamiento, pudimos comprobar que el tiempo empleado por el

algoritmo *Sailfish* fue 1,6 y 2,7 veces más rápido que los métodos *Kallisto* y *Salmon*, respectivamente **Figura 4.10**.

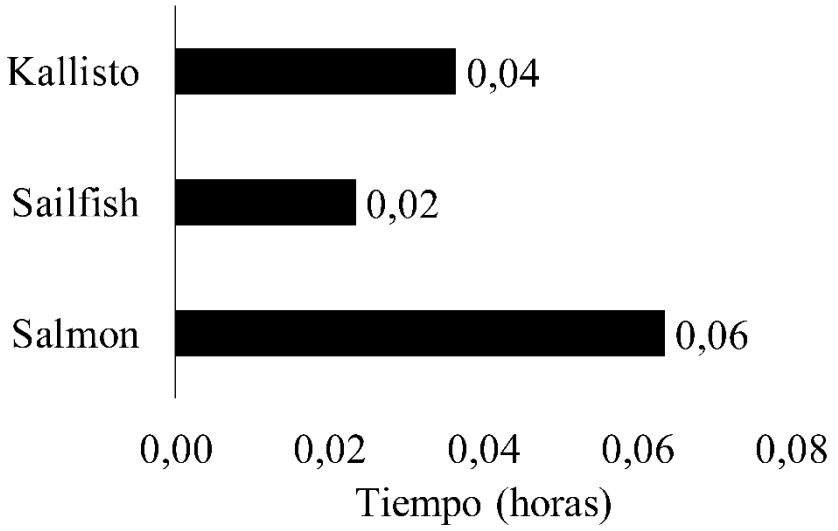



***Figura 4.10.*** *Tiempo de ejecución, en horas, de los tres algoritmos empleados en este trabajo para el pseudoalineamiento.*

En cuanto a la influencia que tienen estos algoritmos sobre la cuantificación de la expresión génica en crudo, no encontramos diferencias significativas entre ellos (**Figura 4.11**), sin embargo, observamos que los *pipelines* en los que se utilizó *Salmon* seguido de la normalización TPM, tuvieron un rendimiento mejor que el resto de los algoritmos y combinaciones. Esto aparece representado como una distribución bimodal, donde los pipelines de *Salmon* que utilizaron TPM están recogidos en el primer cuartil, mientras que el resto de los *pipelines* se encuentran en la zona de intersección del tercer y cuarto cuartil. Al comprobar el efecto sobre la cuantificación en comparación con los métodos de alineamiento tradicionales, no se encontraron diferencias estadísticamente significativas.

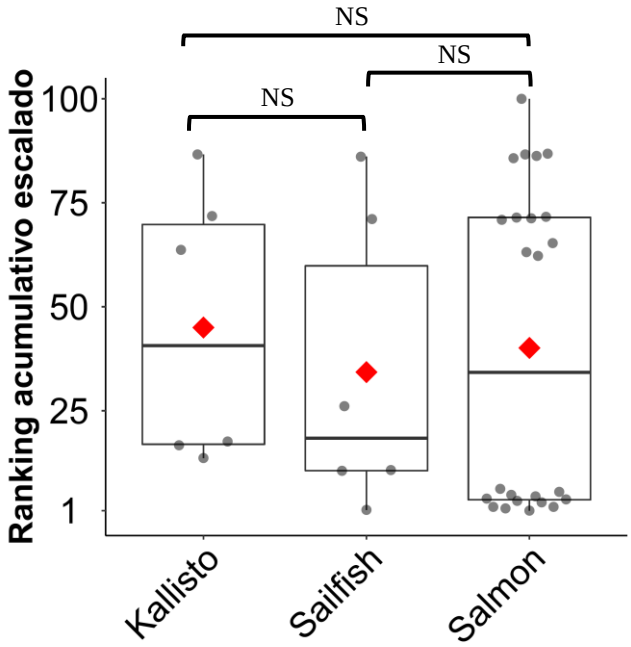



***Figura 4.11***. *Influencia de los métodos de pseudoalineamiento sobre los rankings finales de la cuantificación de la expresión génica cruda. El rombo rojo representa la media escalada de los rankings alcanzados por todos los pipelines en que intervienen cada uno de los algoritmos. Menores valores del ranking acumulativo escalado implican mayor rendimiento del pipeline. NS = resultado no significativo para el test de Dunn a FDR < 0,05.*

Como consecuencia de estos resultados, observamos que una de las principales ventajas de estos métodos es que condensan los pasos de alineamiento, contaje y normalización en un único proceso, simplificando de este modo todo el proceso de análisis. Además, estos métodos mostraron una gran precisión a la hora de estimar la expresión génica en crudo (**Anexo 4**), sin embargo, su exactitud aún se mantiene alejada de los métodos de alineamiento convencionales (**Anexo 5**). En cualquier caso, los algoritmos de pseudoalineamiento seguirían siendo una excelente alternativa como herramienta de exploración en estudios de RNA-seq, gracias también a su elevada velocidad de ejecución.

#### 4.1.1.2. Precisión del *pipeline*

El estudio de la precisión se llevó a cabo sobre los 192 *pipelines* a nivel de la cuantificación génica cruda. Se utilizaron las seis muestras control procedentes de dos líneas celulares de MM bien establecidas, KMS12-BM y JJN3. El primer paso para el estudio de la precisión fue la selección de 107 genes expresados en ambas líneas celulares. Estos genes se seleccionaron en función de su dispersión en los 192 *pipelines* a partir de una lista de 1.181 genes codificantes expresados en 32 tejidos normales extraída del estudio de Uhlen y colaboradores[401]. El proceso de selección de estos 107 genes, así como el cálculo de la puntuación de la precisión global para cada *pipeline* se detalla en el **Apartado 3.5 de la Sección de Material y métodos**. El resultado de este análisis mostró una sobrerrepresentación en las primeras posiciones del ranking global de precisión de

los *pipelines* que usaron el algoritmo de pseudoalineamiento *Salmon* unido al método de normalización TPM. Es interesante destacar que, del mismo modo, los algoritmos de pseudoalineamiento *Kallisto* y *Sailfish*, asociados también al método TPM, lograron muy buenas posiciones en este ranking de precisión, superando a la mayor parte de los algoritmos de alineamiento tradicional, lo que convierte a la unión de algoritmos de pseudoalineamiento y de normalización por TPM en una combinación ganadora en lo que a precisión se refiere (**Anexo 4**).

En contraposición a lo anterior, los *pipelines* menos precisos fueron aquellos en los que el proceso de normalización estuvo ausente como en el caso de las lecturas en crudo (*Raw reads*), o consistió en los métodos de cuentas efectivas, cuentas estimadas o cobertura. No se apreció ningún otro patrón en cuanto a los algoritmos de recortado, alineamiento o contaje empleados (**Anexo 4**).

#### 4.1.1.3. Exactitud del *pipeline* mediante qRT-PCR

La exactitud de los 192 *pipelines* fue estimada de un modo similar a la precisión. Se seleccionaron 30 genes de la lista de 107 genes utilizados en la precisión para llevar a cabo el análisis además de los genes *GAPDH* y *ACTB*. Se procedió al análisis mediante qRT-PCR obteniendo los valores de Ct de cada uno de estos 32 genes por duplicado en las seis muestras de las dos líneas celulares empleadas. El cálculo de la puntuación asociada a la exactitud aparece detallado en la **Sección de Material y métodos**. El resultado de dicho análisis mostró que las primeras posiciones del ranking estaban ocupadas por *pipelines* en las que se utilizó el método *HTSeq* como método de contaje, sin distinciones entre las aproximaciones *Union* e *Intersection-Strict*, unido a la aproximación para la normalización TMM. Observamos también que las primeras posiciones estaban ocupadas por *pipelines* que utilizaban métodos de alineamiento tradicionales, como *RUM*, *STAR* y *TopHat2*. La mayoría de los *pipelines* que utilizaron un método de pseudoalineamiento se situaron en las últimas posiciones del ranking, independientemente del método de normalización empleado, aunque curiosamente, un *pipeline* que utilizó el algoritmo *Sailfish*, fue el que obtuvo la primera posición en el ranking de exactitud (**Anexo 5**).

#### 4.1.1.4. Análisis de la bondad de los *pipelines*

En un último paso se procedió a la elaboración de un ranking global para los 192 *pipelines* considerando los resultados de precisión y exactitud en las dos líneas celulares. Se decidió dar tanto a precisión como a exactitud el mismo peso a la hora de calcular la bondad global de cada *pipeline*. Como resultado, observamos que los *pipelines* que utilizaron el algoritmo *HTseq* como método de contaje, con predominancia de la opción *Union* entre las 10 primeras posiciones (**Tabla 4.3**), además del método de normalización TMM, fueron los que estuvieron más representados en las primeras posiciones del

ranking. En cuanto a los algoritmos de alineamiento, comprobamos que los *pipelines* que utilizaron *RUM*, *TopHat2* y *STAR* aparecieron más frecuentemente en estas primeras posiciones. No encontramos ninguna tendencia en cuanto al método de recortado de lecturas utilizado. Particularmente, el pipeline que obtuvo una mejor puntuación considerando la precisión y la exactitud global fue en el que intervino *Trimmomatic* como algoritmo de recortado, *RUM* como algoritmo de alineamiento, *HTSeq Union* como método de contaje y TMM como método de normalización. En el extremo opuesto, los *pipelines* que tuvieron peores posiciones en el ranking, y, por ende, peor rendimiento, fueron aquellos en los que el método de normalización empleado fue cuentas estimadas, cuentas efectivas, cobertura o no se aplicó método alguno de normalización (lecturas crudas), con especial predominancia en las últimas posiciones de los *pipelines* que utilizaron como método de recortado *Cutadapt* (**Anexo 6**).

***Tabla 4.3.*** *Top 10 pipelines con mejores rankings en el estudio de la bondad de 192 pipelines.*

| Ranking | Algoritmo de recortado | Algoritmo de alineamiento | Método de contaje | Método de normalización | Precisión (mediana) | Exactitud (mediana) | Precisión y exactitud global |
|---|---|---|---|---|---|---|---|
| 1 | Trimmomatic | RUM | HTSeq Union | TMM | 68,5 | 56 | 249 |
| 2 | Trimmomatic | RUM | HTSeq INTER | TMM | 68,5 | 57,75 | 252,5 |
| 3 | BBDuk | RUM | HTSeq Union | TMM | 70 | 56.5 | 253 |
| 4 | BBDuk | STAR | HTSeq Union | TMM | 68 | 62 | 260 |
| 5 | Cutadapt | TopHat2 | HTSeq Union | TMM | 62,5 | 67,75 | 260,5 |
| 6 | BBDuk | TopHat2 | HTSeq Union | TMM | 63,5 | 68 | 263 |
| 7 | BBDuk | HiSat2 | HTSeq Union | TMM | 63,5 | 69,25 | 265,5 |
| 8 | BBDuk | TopHat2 | HTSeq INTER | TMM | 62,5 | 70,5 | 266 |
| 9 | Trimmomatic | STAR | HTSeq Union | TMM | 69 | 64,75 | 267,5 |
| 10 | Trimmomatic | STAR | HTSeq INTER | TMM | 63,5 | 71 | 269 |

*La posición en el ranking viene determinada por los valores globales de precisión y exactitud calculados como el sumatorio de ambos valores en dos líneas celulares independientes.*

### 4.1.2. Cuantificación de la expresión génica diferencial

El término expresión génica diferencial hace referencia a la cantidad relativa de mRNA que es observado en una condición problema respecto a una condición control. Para realizar este estudio se dispuso de 18 muestras de RNA-seq correspondientes a seis condiciones experimentales. Estas seis condiciones experimentales responden a la utilización de tres regímenes de tratamiento (amilorida [T1], TG003 [T2] y vehículo [T0]) sobre dos HMCLs (KMS12-BM [LCA] y JJN3 [LCB]). Cada condición contaba con tres réplicas biológicas.

#### 4.1.2.1. Evaluación de los métodos de expresión génica diferencial

Para realizar el análisis de la expresión génica diferencial se recalcularon los rankings del estudio de la expresión génica cruda de modo que no se consideró el proceso de normalización. Esto fue necesario debido a que la mayor parte de los métodos de expresión diferencial utilizan como datos de partida el valor crudo del número de lecturas de cada experimento de RNA-seq. Se seleccionaron de este modo los 10 mejores *pipelines*, sobre los cuales se procedió a la evaluación de 17 métodos de expresión diferencial. Estos 17 métodos son el resultado de la combinación de 11 algoritmos de expresión diferencial y las diferentes opciones de normalización que acompañan a cada uno de estos algoritmos en sus respectivos paquetes. Los 17 métodos se evaluaron a través del contraste estadístico de las seis condiciones experimentales disponibles en este trabajo, de forma que se procedió a la evaluación de cinco escenarios correspondientes a cinco comparaciones estadísticas que fueron variables en cuanto al número de cambios de expresión génica. Así, la secuencia de estas comparaciones, ordenadas de mayor a menor número de cambios de la expresión génica fue:

1) LCA-T0 *vs.* LCB-T0
2) LCA-T1 *vs.* LCA-T0
3) LCA-T2 *vs.* LCA-T0
4) LCB-T1 *vs.* LCB-T0
5) LCB-T2 *vs.* LCB-T0

Todas estas comparaciones estadísticas fueron evaluadas a tres niveles de significancia: FDR < 0,05, FDR < 0,01 y FDR < 0,001.

#### 4.1.2.2. Bondad de los métodos de expresión génica diferencial

Considerando todo lo anterior, en un primer paso se llevó a cabo un estudio comparativo de los 17 métodos a nivel de detección de la expresión génica diferencial. Para ello se realizó el contaje del número de genes diferencialmente expresados (GDE) detectados por cada uno de los 17 métodos a los tres niveles de significancia estadística propuestos. Se detectó una gran homogeneidad entre los distintos métodos, principalmente en las comparaciones estadísticas donde el número de GDE fue alto. Sin embargo, esta homogeneidad se vio reducida a medida que las diferencias entre las condiciones experimentales contrastadas se redujeron, de forma que métodos como *Ballgown*, *baySeq* o *Cuffdiff* presentaron unos niveles de detección mucho menores que el resto de los métodos estudiados (**Figura 4.12**). Este hallazgo está de acuerdo con estudios anteriores como el de Seyednasrollah y colaboradores[253], donde indican que el algoritmo *Cuffdiff* se comporta de una forma conservadora en cuanto al número de genes detectados.

Asimismo, también observamos que métodos como *SAMseq* o *Ballgown* perdieron su poder de detección a FDR < 0,001 en los contrastes con un menor número de GDE. Esta pérdida fue especialmente acusada en el caso de *Ballgown*, ya que tampoco consigue detectar GDE a FDR < 0,01. Otros métodos como *BaySeq*, *EBSeq* y *Cuffdiff* también perdieron poder de detección, aunque en este caso no llegó a ser tan acusada como para no detectar GDE a un determinado nivel de significancia estadística. En cualquier caso, estos hallazgos deben ser tenidos en cuenta a la hora de considerar estos algoritmos en futuros estudios, ya que esa falta de poder de detección podría conducir a un alto número de falsos negativos en el análisis de expresión génica diferencial.

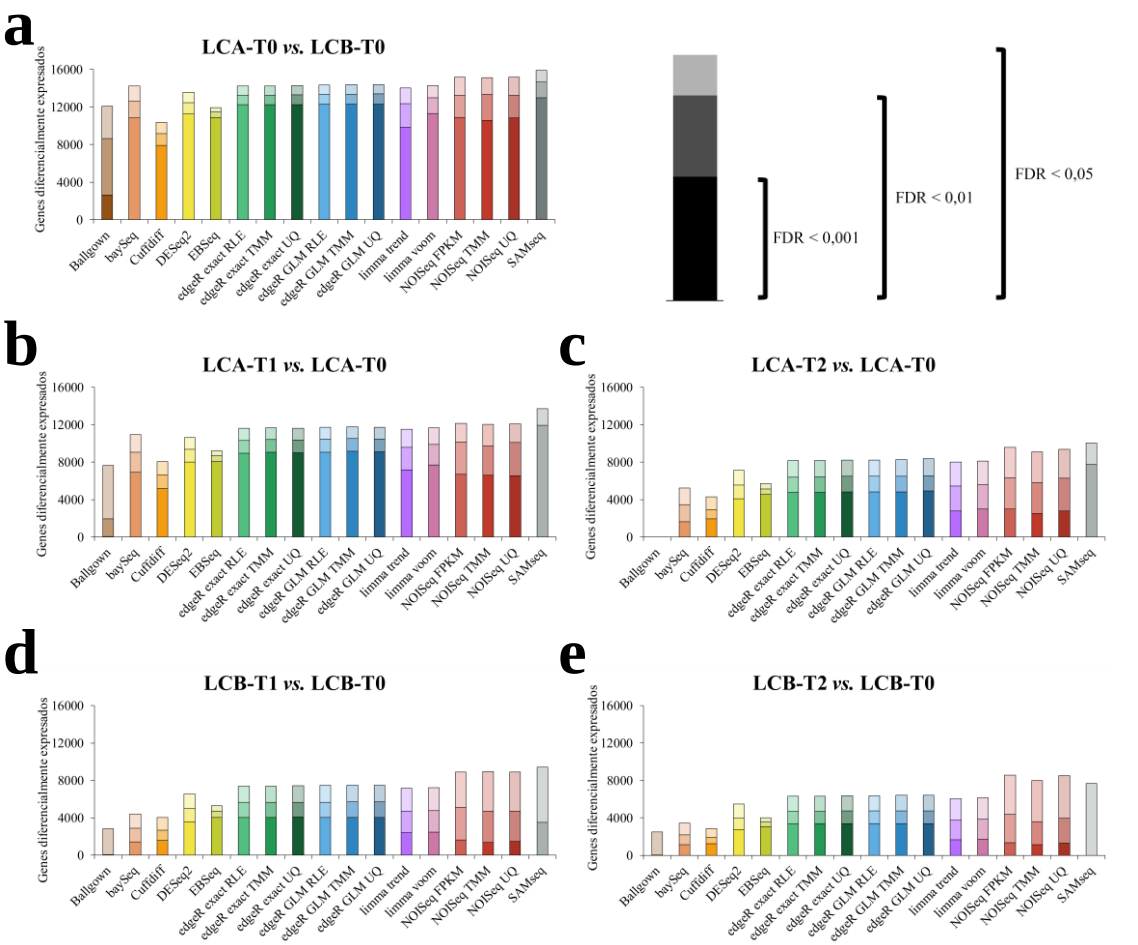



***Figura 4.12.*** *Detección de la expresión génica diferencial. Número de genes diferencialmente expresados (GDE) detectados por los 17 métodos de expresión génica diferencial a tres puntos de corte del FDR: 0,05, 0,01 y 0,001. Los paneles representan los cinco escenarios de expresión génica diferencial en orden descendente del número de GDE desde el panel* ***a)*** *al* ***e)****. Para este análisis se utilizaron 18 muestras correspondientes a las líneas celulares KMS12-BM (LCA) y JJN-3 (LCB) tratadas con amilorida (T1) o TG003 (T2), y las respectivas muestras control (T0).*

La comparación de la similitud entre los 17 algoritmos a través del cálculo de la distancia Euclidea reveló una gran igualdad entre los métodos basados en el algoritmo *edgeR*, los basados en *limma*, *DESeq2* y *baySeq* en todos los escenarios de análisis (**Figura 4.13**). Sin embargo, *SAMseq*, *Cuffdiff* y *Ballgown* fueron los métodos que presentaron unas mayores distancias frente a la mayor parte de los métodos de análisis empleados. En el caso de los métodos basados en *NOIseq*, fueron los que exhibieron una mayor variabilidad, con amplias distancias frente a la mayoría de métodos de expresión diferencial en los escenarios con mayor número de cambios (**Figura 4.13a y 4.13b**), pero obteniendo una gran similitud a métodos como los basados en *edgeR* o *limma* en los escenarios con menor número de cambios de expresión génica (**Figura 4.13e**).

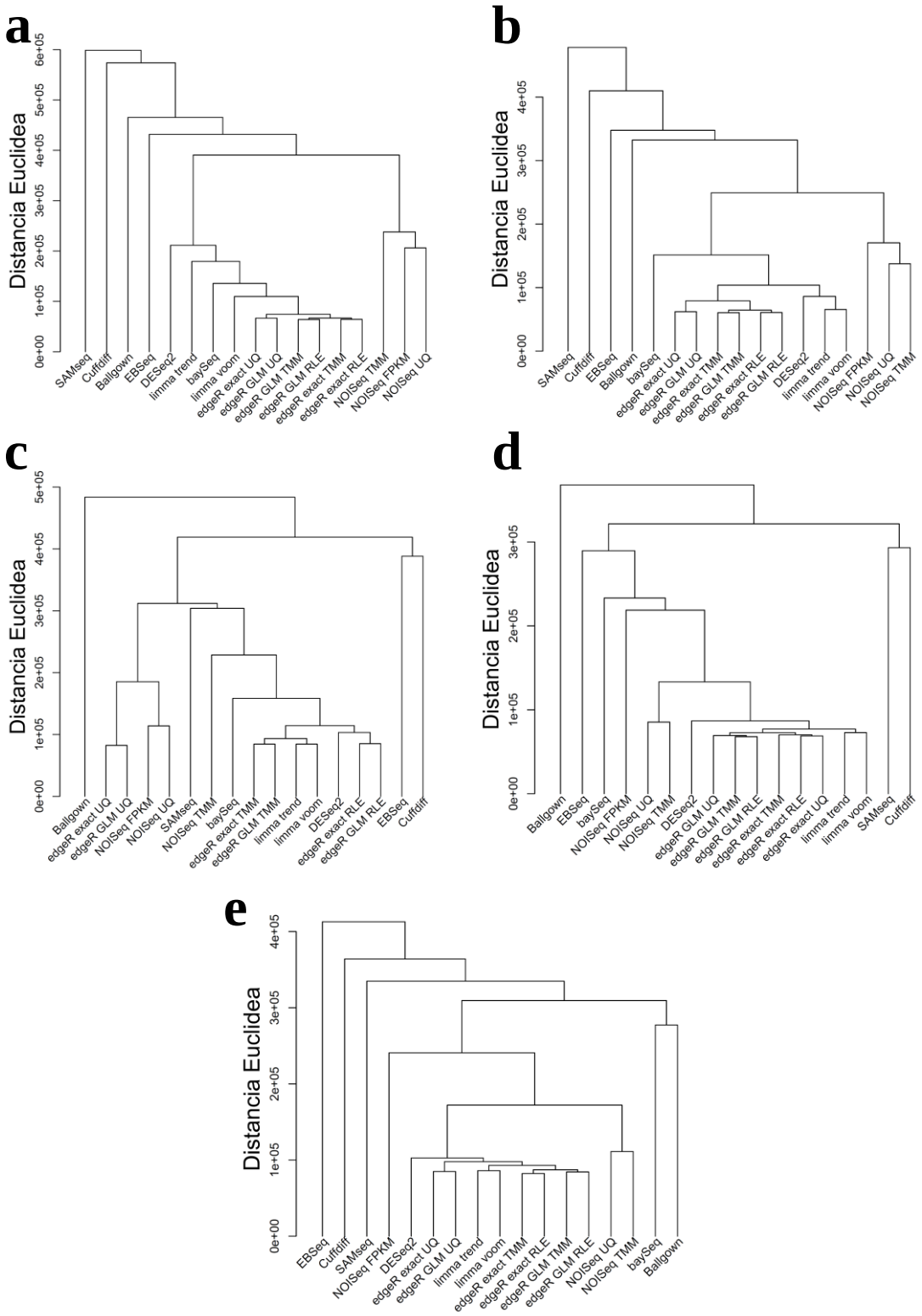



***Figura 4.13.*** *Similitud de los métodos de detección de la expresión génica diferencial en RNA-seq. En cada panel se representa la distancia entre los 17 métodos mediante dendrogramas a través de la distancia Euclídea.* ***a)*** *LCA-T0 vs. LCB-T0,* ***b)*** *LCA-T1 vs. LCA-T0,* ***c)*** *LCA-T2 vs. LCA-T0,* ***d)*** *LCB-T1 vs. LCB-T0 y* ***e)*** *LCB-T2 vs. LCB-T0.*

En un segundo paso se procedió al análisis de 7 parámetros para la evaluación del desempeño de los 17 métodos de expresión diferencial. Los parámetros analizados en este trabajo fueron la precisión (valor predictor positivo [VPP]), la exactitud (del inglés *accuracy* [ACC]), la sensibilidad (fracción de verdaderos positivos [FVP]), la especificidad (fracción de verdaderos negativos [FVN]), el valor predictor negativo (VPN), el área bajo la curva (AUC) y el coeficiente de correlación de Matthews (CCM). Para ello se midió por duplicado en las 18 muestras la expresión de los 32 genes seleccionados para el análisis de la exactitud de los *pipelines* (**Apartado 4.1.1.3**) utilizando la técnica qRT-PCR. Los resultados de estos análisis aparecen recogidos en la **Figura 4.14**.

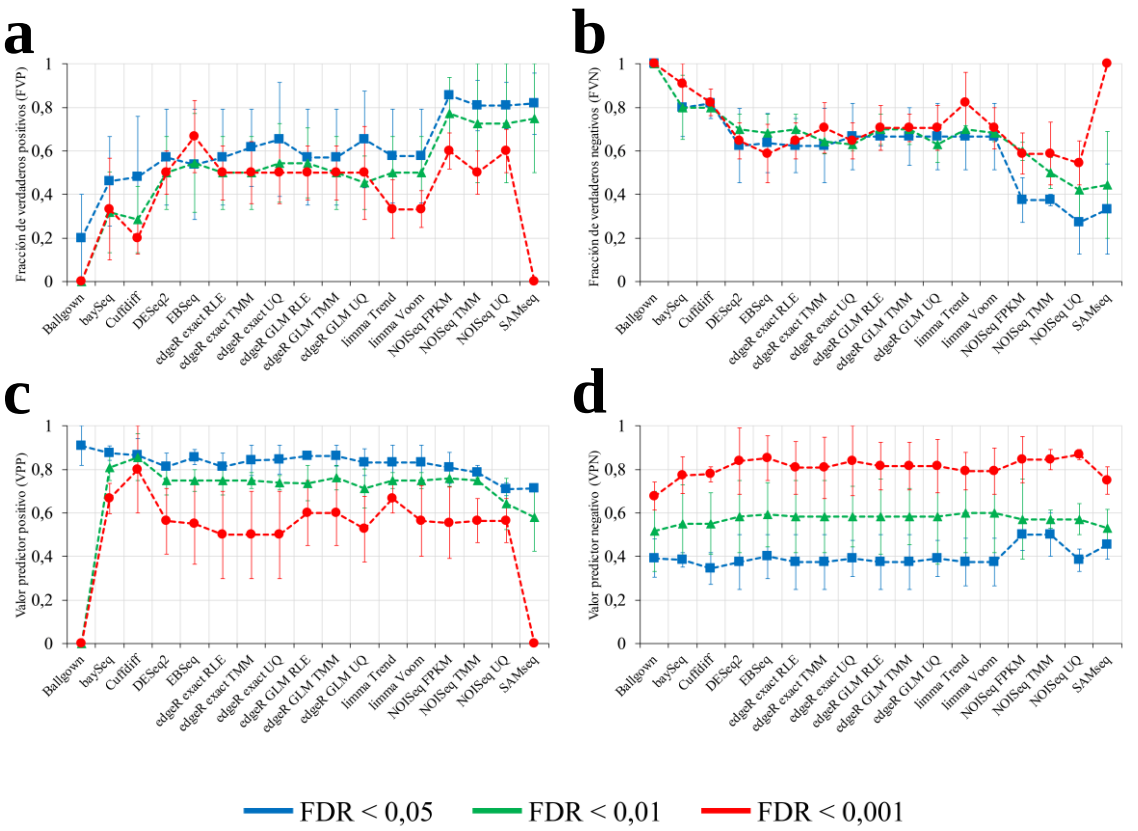



***Figura 4.14**. Análisis del desempeño de los métodos de expresión génica diferencial a través de la medición de 7 parámetros. **a)** Fracción de verdaderos positivos (FVP), **b)** fracción de verdaderos negativos (FVN), **c)** valor predictor positivo (VPP), **d)** valor predictor negativo (VPN). Los 7 parámetros fueron evaluados a tres puntos de corte del FDR: 0,05, 0,01 y 0,001. Para cada parámetro se representa la mediana ± MAD de los cinco escenarios de expresión génica diferencial: LCA-T0 vs. LCB-T0, LCA-T1 vs. LCA-T0, LCA-T2 vs. LCA-T0, LCB-T1 vs. LCB-T0 y LCB-T2 vs. LCB-T0.*

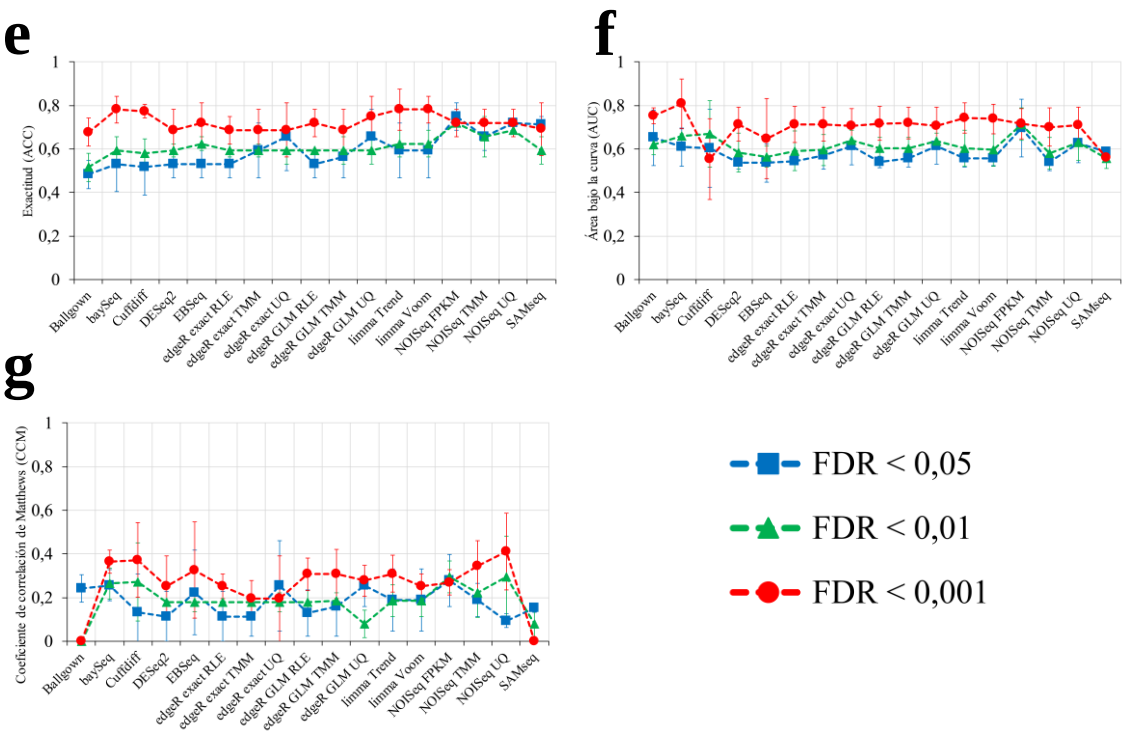



***Figura 4.14 (continuación)**. Análisis del desempeño de los métodos de expresión génica diferencial a través de la medición de 7 parámetros. **e)** Exactitud (ACC), **f)** área bajo la curva (AUC) y **g)** coeficiente de correlación de Matthews (CCM). Los 7 parámetros fueron evaluados a tres puntos de corte del FDR: 0,05, 0,01 y 0,001. Para cada parámetro se representa la mediana ± MAD de los cinco escenarios de expresión génica diferencial: LCA-T0 vs. LCB-T0, LCA-T1 vs. LCA-T0, LCA-T2 vs. LCA-T0, LCB-T1 vs. LCB-T0 y LCB-T2 vs. LCB-T0.*

El análisis individual de cada uno de estos parámetros reveló diferencias entre los 17 métodos de expresión diferencial. Así, por lo que respecta a la FVP los métodos basados en *NOISeq* son los que presentaron un mejor comportamiento y, por tanto, una mayor sensibilidad a los tres niveles de FDR, seguidos de *EBSeq* y los métodos basados en *edgeR*. En el extremo opuesto se situó el método *Ballgown*, en el que la detección de verdaderos positivos fue muy baja a los tres niveles de significancia (**Figura 4.14a**).

El resultado obtenido para el VPP (**Figura 4.14c**) fue muy similar al del parámetro FVP, salvo en métodos como *baySeq* o *Cuffdiff*, que en el VPP presentaron un buen desempeño alcanzando para el nivel FDR < 0,05 valores de 0,88 y 0,87, respectivamente. Del mismo modo, también a FDR< 0,05 *Ballgown* alcanzó un VPP = 0,91, lo que indica que en el 91% de las ocasiones, si el método *Ballgown* detecta un gen diferencialmente expresado a FDR < 0,05, este gen estará realmente expresado de forma diferencial. De forma opuesta, el método *Ballgown*, al igual que ocurrió para la FVP, obtuvo un VPP = 0 a FDR < 0,01 y a FDR < 0,001, indicando este hecho el mal funcionamiento en la detección de verdaderos positivos de este método a ambos niveles de significancia.

En cuanto a la FVN (**Figura 4.14b**), los métodos con una mejor detección de verdaderos negativos fueron *SAMseq* a FDR < 0,001 y *Ballgown* a los tres niveles de FDR, en todos los casos consiguiendo un 100% de detección. Sin embargo, este buen

rendimiento fue debido a la ausencia de detección de genes diferencialmente expresados (verdaderos positivos) (**Figura 4.14a**). En este caso, los métodos no paramétricos, representados por *NOISeq* y *SAMseq*, fueron los que presentaron un peor comportamiento, principalmente a FDR < 0,05 y a FDR < 0,01.

Respecto al VPN (**Figura 4.14d**), todos los métodos obtuvieron sus mejores resultados a FDR < 0,001 rondando un VPN de 0,8, lo que indica que en el 80% de las ocasiones, cuando uno de estos métodos detecte que un gen no está diferencialmente expresado, este gen no presentará expresión diferencial entre las condiciones experimentales contrastadas.

En lo que atañe a los parámetros ACC (**Figura 4.14e**) y AUC (**Figura 4.14f**), los valores más elevados de ambos parámetros se alcanzaron también en la mayor parte de los métodos a FDR < 0,001, salvo *NOISeq* FPKM y *SAMseq* en el caso de ACC, y *Cuffdiff* y *SAMseq* en el caso del AUC. Particularmente, los métodos con un mayor valor de ACC, es decir, más exactos, fueron *limma trend*, *limma voom* y *baySeq* a FDR < 0,001 con ACC = 0,78, mientras que el método con un mayor AUC fue *baySeq* a FDR < 0,001 con AUC = 0,81.

Finalmente, en el análisis del CCM (**Figura 4.14g**), que mide la calidad de la clasificación, ninguno de los 17 métodos logró un CCM > 0,7, lo que indica que ningún método obtuvo una relación muy fuerte y positiva entre el resultado real y el resultado obtenido por dichos métodos[423]. Sin embargo, sí se obtuvo una relación fuerte y positiva (CCM > 0,4) con el método *NOISeq* UQ a FDR < 0,001, y una relación moderada y positiva (CCM > 0,3) con métodos como *Cuffdiff*, *baySeq* y algunas variantes de *edgeR* a los FDR < 0,01 y FDR < 0,001. Se consideró como ausencia de relación un CCM < 0,2, y en ningún caso se registraron valores negativos de este parámetro.

La clasificación completa de todos los métodos a los tres puntos de corte del FDR, para los 7 parámetros puede ser consultada en el **Anexo 7**.

A continuación, se llevó a cabo el análisis del rendimiento de cada uno de los 17 métodos de expresión diferencial considerando de manera simultánea los 7 parámetros estudiados. Se consideraron tres supuestos experimentales para no beneficiar o perjudicar a alguno de los métodos estudiados: rendimiento general, rendimiento por escenario de contraste de expresión génica diferencial y rendimiento por significancia estadística.

Como primera aproximación se procedió con el supuesto experimental de determinación del rendimiento de manera general. Para llevar a cabo este análisis fueron considerados de manera conjunta los cinco escenarios de expresión génica diferencial y los tres niveles de significancia propuestos en este trabajo (**Figura 4.15)**.

| Método de análisis de la expresión génica diferencial | LCA-T0 *vs.* LCB-T0 | | | LCA-T1 *vs.* LCA-T0 | | | LCA-T2 *vs.* LCA-T0 | | | LCB-T1 *vs.* LCA-T0 | | | LCB-T2 *vs.* LCB-T0 | | | Mediana global |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 | |
| **Ballgown** | 12 | 10 | 8 | 17 | 7 | 14 | 12 | 16 | 6 | 6 | 5 | 11 | 5 | 14 | 16 | **11** |
| **baySeq** | 12 | 9 | 7 | 14 | 14 | 2 | 12 | 6 | 3 | 3 | 1 | 1 | 7 | 4 | 5 | **6** |
| **Cuffdiff** | 9 | 11 | 9 | 10 | 5 | 1 | 8 | 13 | 5 | 17 | 14 | 5 | 17 | 5 | 13 | **9** |
| **DESeq2** | 1 | 1 | 11 | 3 | 11 | 11 | 17 | 13 | 11 | 13 | 12 | 14 | 9 | 13 | 6 | **11** |
| **EBSeq** | 11 | 6 | 13 | 11 | 6 | 9 | 15 | 15 | 17 | 7 | 2 | 4 | 1 | 1 | 1 | **7** |
| **edgeR exact RLE** | 8 | 1 | 11 | 3 | 11 | 11 | 16 | 12 | 7 | 15 | 9 | 14 | 12 | 12 | 14 | **11** |
| **edgeR exact TMM** | 1 | 14 | 6 | 6 | 13 | 15 | 10 | 8 | 12 | 13 | 9 | 14 | 12 | 10 | 15 | **12** |
| **edgeR exact UQ** | 3 | 13 | 10 | 1 | 8 | 17 | 2 | 11 | 14 | 16 | 11 | 14 | 3 | 10 | 10 | **10** |
| **edgeR GLM RLE** | 15 | 3 | 4 | 6 | 16 | 7 | 11 | 3 | 13 | 9 | 7 | 7 | 16 | 7 | 9 | **7** |
| **edgeR GLM TMM** | 4 | 3 | 5 | 5 | 2 | 6 | 9 | 10 | 9 | 9 | 7 | 12 | 14 | 7 | 11 | **7** |
| **edgeR GLM UQ** | 4 | 5 | 3 | 8 | 17 | 13 | 2 | 16 | 16 | 8 | 6 | 2 | 4 | 9 | 11 | **8** |
| **limma trend** | 7 | 7 | 1 | 11 | 9 | 4 | 6 | 7 | 15 | 11 | 3 | 3 | 14 | 3 | 6 | **7** |
| **limma voom** | 4 | 12 | 2 | 11 | 9 | 10 | 6 | 8 | 7 | 12 | 3 | 5 | 11 | 2 | 6 | **7** |
| **NOISeq FPKM** | 17 | 8 | 14 | 8 | 1 | 8 | 1 | 1 | 1 | 5 | 13 | 8 | 10 | 6 | 2 | **8** |
| **NOISeq TMM** | 10 | 15 | 15 | 2 | 3 | 5 | 12 | 4 | 9 | 2 | 16 | 10 | 2 | 15 | 4 | **9** |
| **NOISeq UQ** | 14 | 15 | 16 | 16 | 4 | 2 | 4 | 2 | 2 | 4 | 17 | 9 | 6 | 16 | 2 | **6** |
| **SAMseq** | 16 | 17 | 17 | 15 | 14 | 16 | 5 | 5 | 4 | 1 | 15 | 13 | 8 | 17 | 16 | **15** |

1 — Mejor rendimiento

17 — Peor rendimiento

***Figura 4.15.*** *Rendimiento general de los 17 métodos de expresión génica diferencial. Los valores numéricos coloreados según la escala indicada en la leyenda representan el ranking de los 17 métodos en cada uno de los cinco escenarios de expresión génica diferencial (LCA-T0 vs. LCB-T0, LCA-T1 vs. LCA-T0, LCA-T2 vs. LCA-T0, LCB-T1 vs. LCA-T0 y LCB-T2 vs. LCB-T0) para cada nivel de significancia estudiado (FDR < 0,05, FDR < 0,01 y FDR < 0,001). A menor valor del ranking, mayor rendimiento del método de expresión génica diferencial. La columna de mediana global recoge la mediana de todos los valores anteriores, de modo que cuanto menor sea este valor mayor será el rendimiento global del método de expresión diferencial.*

Los métodos que presentaron un mejor rendimiento general fueron *baySeq* y *NOISeq* en su versión UQ, aunque en el escenario con un mayor número de cambios de expresión génica ambos métodos clasificaron por debajo de la novena plaza del ranking en al menos dos niveles de significancia. Sin embargo, su buen desempeño en los escenarios con menor número o con un número intermedio de cambios de expresión génica, respectivamente, consiguió auparlos a las primeras posiciones del ranking. Estos dos métodos estuvieron seguidos muy de cerca por *limma trend*, *limma voom*, *EBseq* y *edgeR* en su versión GLM, aunque en el caso de *EBSeq*, mostró un comportamiento muy irregular, alcanzando las primeras posiciones del ranking en los escenarios con menor número de genes diferencialmente expresados, pero obteniendo rankings muy bajos en los escenarios con cambios de expresión génica intermedios.

A continuación, se llevó a cabo el estudio del segundo supuesto experimental: el análisis del rendimiento por escenario de contraste de expresión génica diferencial. En este caso se observó una gran variabilidad en cuanto al rendimiento de los 17 métodos, de modo que ninguno de los métodos logró repetir la primera posición en más de uno de los de los cinco escenarios estudiados. De esta manera se advirtieron casos extremos como el del método *EBSeq*, que obtuvo un tercer y un primer puesto en los dos escenarios con un menor número de genes diferencialmente expresados, y, sin embargo, se clasificó decimotercero, octavo y decimosexto en los escenarios con un número mayor de diferencias; el caso de *DESeq2*, primero en el contraste con mayores diferencias, pero con rankings por debajo del noveno en el resto de escenarios; o el caso de *baySeq*, que obtuvo sus mejores resultados en los escenarios con un menor número de cambios de expresión génica, pero su rendimiento no fue adecuado en el escenario opuesto en número de genes diferencialmente expresados (**Figura 4.16**).

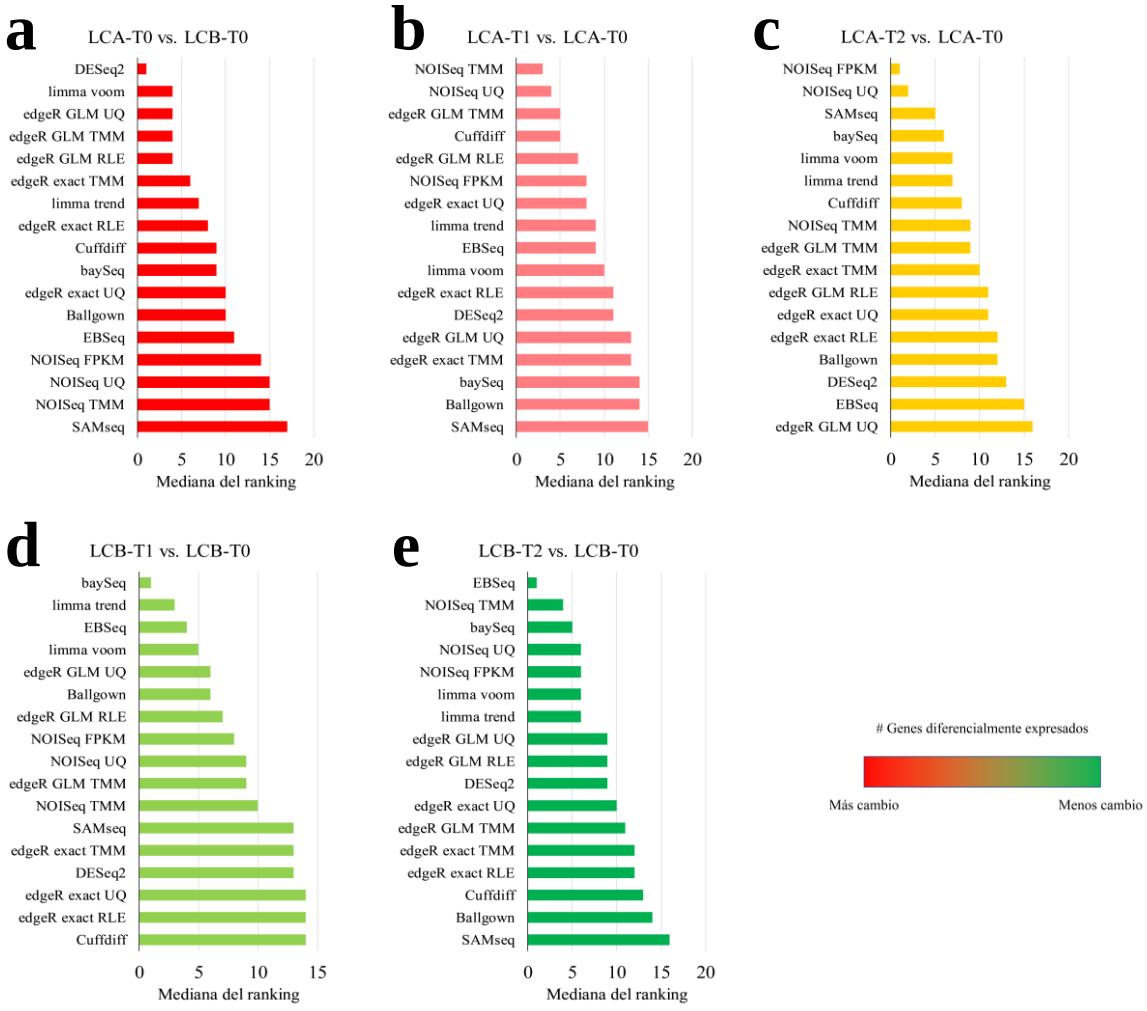



***Figura 4.16.*** *Rendimiento de los 17 métodos de expresión génica diferencial en función de los cinco escenarios de expresión génica diferencial propuestos en este trabajo. Estos cinco escenarios se definen en función del número de cambios en la expresión génica entre las dos condiciones contrastadas, de modo que ordenados de mayor a menor número de cambios son los siguientes:* ***a)*** *LCA-T0 vs. LCB-T0,* ***b)*** *LCA-T1 vs. LCA-T0,* ***c)*** *LCA-T1 vs. LCA-T0,* ***d)*** *LCB-T1 vs. LCB-T0 y* ***e)*** *LCB-T2 vs. LCB-T0.*

En lo que respecta al último de los tres supuestos experimentales, es decir, al análisis en función del nivel de significancia estadística, también se observó variabilidad en las posiciones ocupadas por los 17 métodos, como ocurrió en el caso de *edgeR exact* UQ, que se clasificó en primer lugar a FDR < 0,05 y sin embargo solo pudo alcanzar el decimocuarto y el decimoquinto puesto a FDR < 0,01 y FDR < 0,001, respectivamente. Sin embargo, también hubo métodos que presentaron una buena estabilidad como *NOISeq* FPKM, *limma trend*, *limma voom*, *baySeq* o *edgeR GLM* TMM, siempre clasificados entre los 8 mejores métodos. En el lado opuesto, métodos como *edgeR exact* TMM, *Cuffdiff* o *SAMseq* no lograron mejorar la novena posición. (**Figura 4.17**)

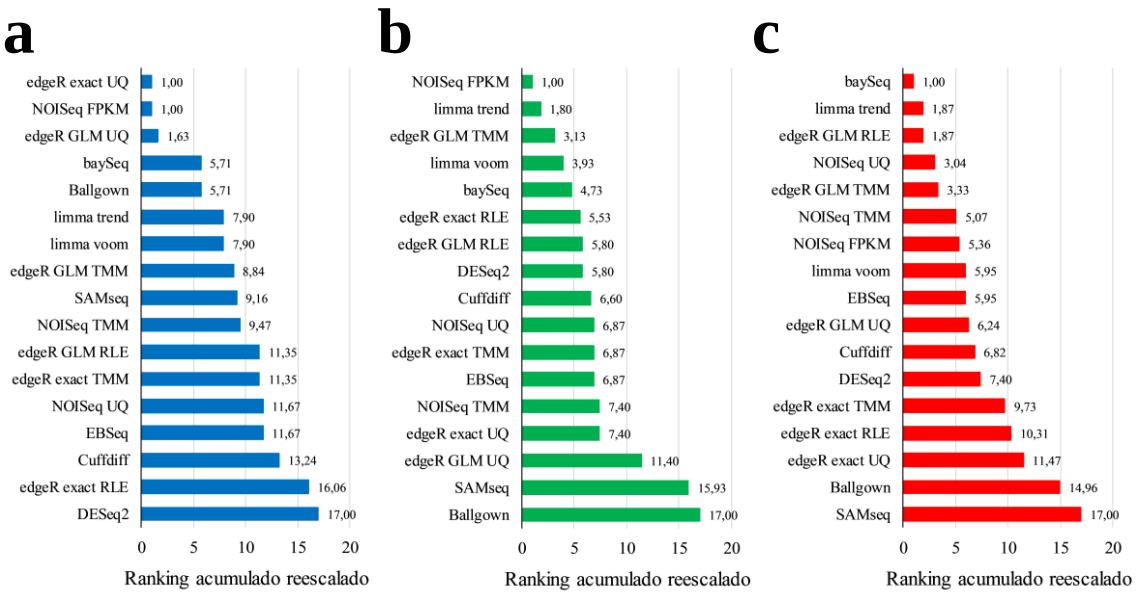



***Figura 4.17.*** *Rendimiento de los 17 métodos de expresión génica diferencial en función de los tres niveles de significancia propuestos en este trabajo:* ***a)*** *FDR < 0,05,* ***b)*** *FDR < 0,01 y* ***c)*** *FDR < 0,001.*

Como resumen final de los tres supuestos experimentales para el análisis del rendimiento, el método que obtuvo un mayor equilibrio (mejores posiciones) considerando tanto el análisis general como los cinco escenarios de forma individual y los tres niveles de FDR, fue *limma trend*, de modo que en todos los casos siempre se clasificó entre los 8 mejores métodos. Otros métodos como *baySeq*, *NOISeq* FPKM, *limma voom* o algunas variantes de *edgeR GLM* también obtuvieron unos buenos resultados, sin embargo, todos estos métodos lograron un ranking superior al noveno puesto en al menos uno de los supuestos estudiados (**Figura 4.18**).

| Método | Rendimiento general | Rendimiento por escenario | | | | | Rendimiento por significancia | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LCA-T0 *vs.* LCB-T0 | LCA-T1 *vs.* LCA-T0 | LCA-T2 *vs.* LCS-T0 | LCB-T1 *vs.* LCB-T0 | LCB-T2 *vs.* LCB-T0 | FDR < 0,05 | FDR < 0,01 | FDR < 0,001 |
| **Ballgown** | 13 | 11 | 15 | 13 | 5 | 16 | 4 | 17 | 16 |
| **baySeq** | **1** | 9 | 15 | 4 | **1** | 3 | 4 | 5 | **1** |
| **Cuffdiff** | 10 | 9 | 3 | 7 | 15 | 15 | 15 | 9 | 11 |
| **DESeq2** | 13 | **1** | 11 | 15 | 12 | 8 | 17 | 7 | 12 |
| **EBSeq** | 3 | 13 | 8 | 16 | 3 | **1** | 13 | 10 | 8 |
| **edgeR exact RLE** | 13 | 8 | 11 | 13 | 15 | 13 | 16 | 6 | 14 |
| **edgeR exact TMM** | 16 | 6 | 13 | 10 | 12 | 13 | 11 | 10 | 13 |
| **edgeR exact UQ** | 12 | 11 | 6 | 11 | 15 | 11 | **1** | 13 | 15 |
| **edgeR GLM RLE** | 3 | 2 | 5 | 11 | 7 | 8 | 11 | 7 | 2 |
| **edgeR GLM TMM** | 3 | 2 | 3 | 8 | 9 | 12 | 8 | 3 | 5 |
| **edgeR GLM UQ** | 8 | 2 | 13 | 17 | 5 | 8 | 3 | 15 | 10 |
| **limma trend** | 3 | 7 | 8 | 5 | 2 | 4 | 6 | 2 | 2 |
| **limma voom** | 3 | 2 | 10 | 5 | 4 | 4 | 6 | 4 | 8 |
| **NOISeq FPKM** | 8 | 14 | 6 | **1** | 8 | 4 | **1** | **1** | 7 |
| **NOISeq TMM** | 10 | 15 | **1** | 8 | 11 | 2 | 10 | 13 | 6 |
| **NOISeq UQ** | **1** | 15 | 2 | 2 | 9 | 4 | 13 | 10 | 4 |
| **SAMseq** | 17 | 17 | 17 | 3 | 12 | 17 | 9 | 16 | 17 |

1 — Mejor rendimiento

17 — Peor rendimiento

***Figura 4.18.*** *Resumen del rendimiento de los 17 métodos de expresión génica diferencial para cada uno de los tres supuestos experimentales estudiados: rendimiento general, rendimiento por escenario de contraste de expresión génica diferencial y rendimiento por significancia estadística.*

Un dato interesante en este punto fue que estos métodos que presentan los mejores rendimientos no asumen la misma distribución *a priori* de los datos de entrada, ya que los métodos *limma* asumen una distribución log-normal, *baySeq* y *edgeR* asumen una distribución binomial negativa, y *NOISeq* no asume ningún tipo de distribución al ser un método no paramétrico. Por tanto, no parece que la distribución asumida por los métodos de análisis tenga una influencia decisiva sobre la bondad de la detección, tal como previamente habían reportado Rapaport y colaboradores[249], aunque sobre métodos diferentes a los estudiados en este trabajo. No obstante, para elucidar este asunto se procedió a la comparación de los rankings obtenidos por los 17 métodos en función de la distribución asumida por los mismos en cada uno de los supuestos de análisis considerados en este trabajo (**Figura 4.19**).

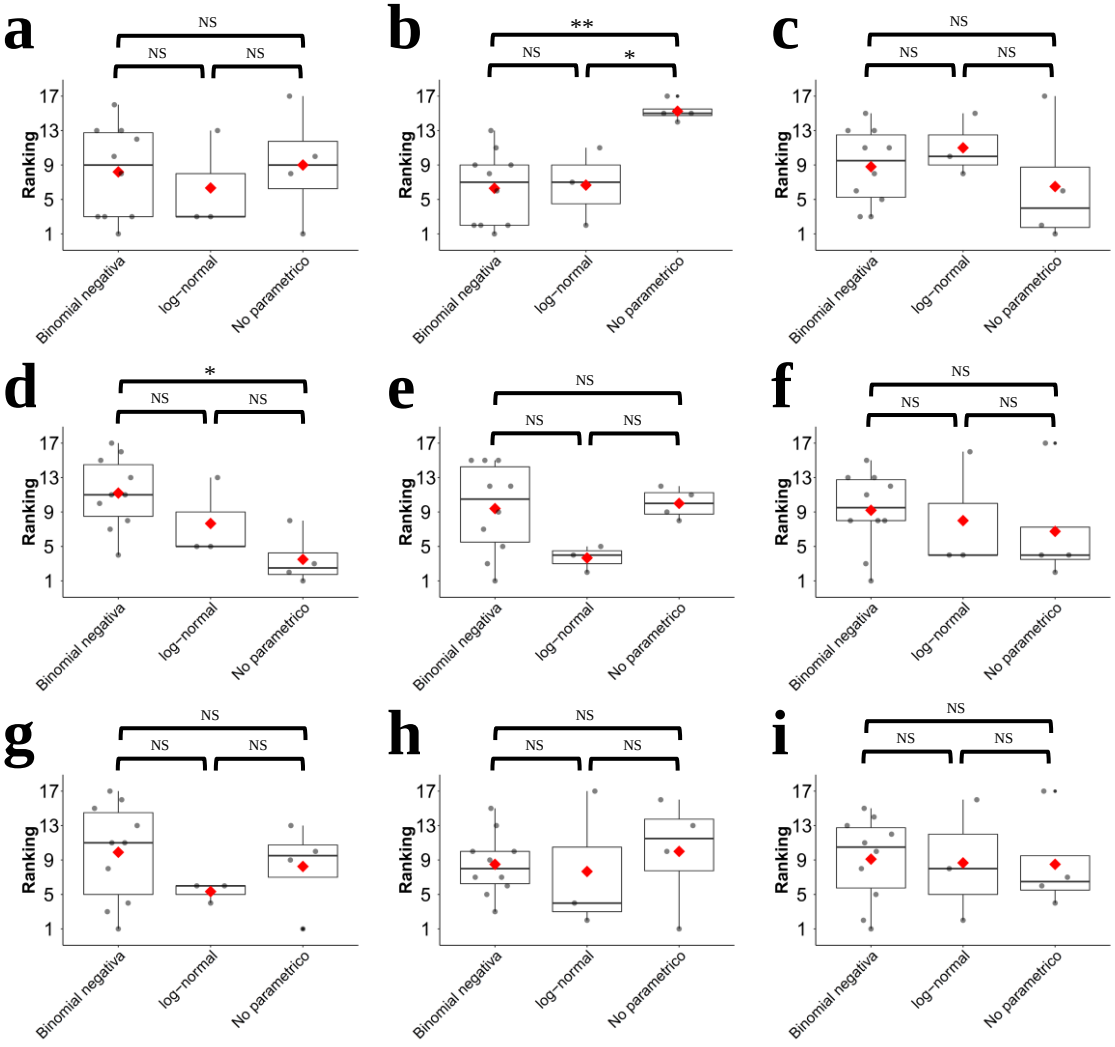



***Figura 4.19.*** *Comparación de los métodos de expresión génica diferencial agrupados en función de la distribución de los datos asumida a priori. Este análisis se llevó a cabo sobre los 9 supuestos experimentales estudiados en este trabajo: a) estudio del rendimiento general; b) a f) cinco escenarios determinados por el número de cambios en la expresión génica de modo que b) corresponde a la comparación LCA-T0 vs. LCB-T0, c) LCA-T1 vs. LCA-T0, d) LCA-T2 vs. LCA-T0, e) LCB-T1 vs. LCB-T0 y f) LCB-T2 vs. LCB-T0; y g) a i) los tres escenarios determinados por el nivel de significancia, de manera que g) corresponde a FDR < 0,05, h) FDR < 0,01 y g) FDR < 0,001. NS = comparación estadísticamente no significativa en el test de Dunn, * = comparación estadísticamente significativa a FDR < 0,05 y ** = comparación estadísticamente significativa a FDR < 0,01. El rombo rojo corresponde al valor medio del ranking de los métodos recogidos en cada caja.*

Este análisis reveló que en la mayoría de los supuestos experimentales y escenarios de análisis no hubo diferencias en cuanto a los rankings ocupados por los métodos de expresión diferencial en función de la distribución asumida. Solamente se detectaron diferencias estadísticamente significativas con los métodos no paramétricos y únicamente en dos escenarios de expresión génica diferencial: el escenario con más cambios de expresión génica (LCA-T0 *vs.* LCB-T0) en el que los métodos no paramétricos tuvieron un rendimiento inferior al resto de métodos, y el escenario LCA-T2 *vs.* LCA-T0 de

cambio intermedio, en el que los métodos no paramétricos funcionaron mejor que los métodos que asumen una distribución binomial negativa.

# 4.2. Capítulo 2.

## Análisis de datos de microarray y comparación con RNA-seq

Se llevó a cabo el análisis comparativo a nivel de expresión génica cruda de los datos procedentes del microarray de transcriptoma HTA2.0 de Affymetrix con los 10 *pipelines* de análisis de la RNA-seq que en nuestro estudio obtuvieron mejores valores de precisión y exactitud determinados en el **Capítulo 1**. De la misma manera, se realizó la comparación a nivel de expresión génica diferencial de los métodos de análisis empleados en el estudio del microarray con el que obtuvo un mejor ranking global en los estudios de RNA-seq.

### 4.2.1. Cuantificación de la expresión génica cruda de los microarrays

La cuantificación de la expresión génica cruda del microarray se efectuó mediante dos técnicas de preprocesamiento diferentes en cuanto a la procedencia del archivo de referencia utilizado. Este archivo indica el lugar de hibridación de cada una de las sondas en el genoma y, por tanto, define qué sondas deben utilizarse para estimar la expresión de un determinado gen. Este análisis, al igual que el que se efectuó en el **Apartado 4.1.1** del **Capítulo 1**, se realizó exclusivamente sobre las tres muestras control de las dos líneas celulares de MM estudiadas.

La primera técnica de preprocesamiento utilizó como archivo de referencia el provisto por BrainArray (BA)[392] específico para el microarray HTA2.0 de Affymetrix a nivel génico bajo la nomenclatura de Ensembl. La ejecución de esta técnica, desde la lectura de los valores de intensidad de fluorescencia en bruto a la normalización de los datos, se realizó en el paquete estadístico R, obteniendo 35.345 "grupos de sondas" o *probesets*, asociados a un idéntico número de identificadores génicos de Ensembl. No hubo identificadores génicos duplicados, ya que el propio diseño del archivo de referencia evita este problema al asignar las sondas a un único identificador.

La segunda técnica de preprocesamiento utilizó el archivo de referencia provisto por el fabricante del microarray (AEC) como referencia para la normalización. El preprocesamiento y normalización se llevó a cabo en la consola de análisis de expresión génica facilitada por el fabricante. Como resultado se obtuvieron 70.523 *probesets*, de los que 52.712 se asociaron de manera no ambigua a algún identificador génico de la base de datos Ensembl, correspondiendo estos 52.712 *probesets* a 37.810 identificadores génicos únicos.

La comparación de ambas técnicas se realizó con los 32.251 genes que fueron comunes a BA y AEC y se muestra en la **Figura 4.20,** donde aparece la correlación de Pearson de la expresión génica entre BA y AEC en cada una de las seis muestras estudiadas. En el caso de AEC, debido a la presencia de genes duplicados, se seleccionó un único *probeset* para representar a cada gen. Esta selección fue realizada de manera aleatoria. Los resultados muestran una buena correlación en todas las muestras entre ambos procedimientos ($r > 0{,}90$, $p < 0{,}0001$), lo que indica que los resultados a nivel de

detección génica son altamente comparables. Sin embargo, en los siguientes apartados trataremos de determinar las diferencias a nivel de detección, así como los pros y los contras de ambas técnicas.

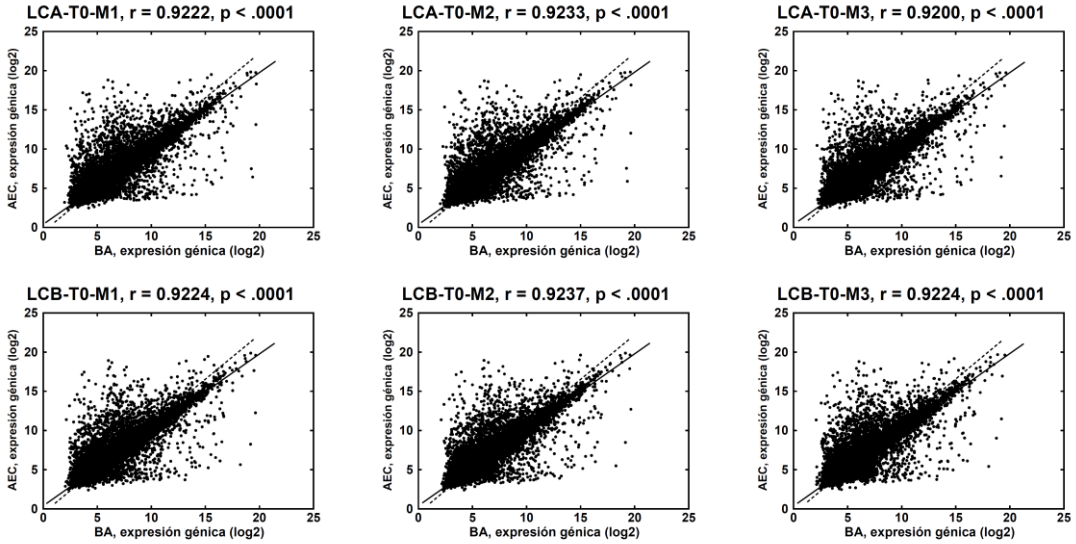



***Figura 4.20.*** *Gráfico de correlación sobre las seis muestras control para el estudio mediante microarrays. En el eje de abscisas (X) se muestran los valores de expresión génica en $log_2$ de los genes utilizando como referencia BA. En el eje de ordenadas (Y) se muestran los valores de expresión génica en $log_2$ de los genes utilizando la referencia AEC.*

### 4.2.2. Cuantificación de la expresión génica diferencial de los microarrays

El análisis de expresión génica diferencial se llevó a cabo utilizando las mismas 18 muestras en las que se realizó el análisis de la RNA-seq en el **Capítulo 1**, mediante la aplicación de dos métodos estadísticos paramétricos: *limma* y *SAMr*. Este análisis se realizó por separado sobre los datos normalizados con las referencias BA y AEC. De este modo, el número de genes analizados fue dependiente del método de preprocesamiento considerado, analizando 35.345 genes en el caso de BA y, en el caso de AEC, considerando dos listas iniciales de genes, una primera de 52.712 genes, que contuvo genes duplicados, y una segunda lista de 37.810 genes, sin elementos repetidos. En ambos casos, de manera similar al estudio de la expresión diferencial realizado en el **Capítulo 1**, se realizó el estudio sobre cinco escenarios de análisis o contrastes de condiciones experimentales, considerando las combinaciones descritas en la **Figura 3.1** y en la **Sección de Material y métodos** entre los cinco grupos de tratamiento utilizados en este trabajo.

Como paso previo al análisis de expresión diferencial se comprobó la similitud entre las muestras de los cinco grupos de tratamiento a través del análisis no supervisado de la expresión génica mediante técnicas de escalado multidimensional (MDS) (**Figura 4.21**).

Como puede observarse, en ambas líneas celulares, KMS12-BM (LCA) y JJN-3 (LCB), e independientemente de la técnica seleccionada para el procesado de los datos, el grupo de tratamiento con el compuesto TG003 (T2) muestra una mayor similitud al grupo control (T0) que el grupo de tratamiento con amilorida (T1). Mediante el análisis de expresión diferencial se tratará de confirmar esta diferencia

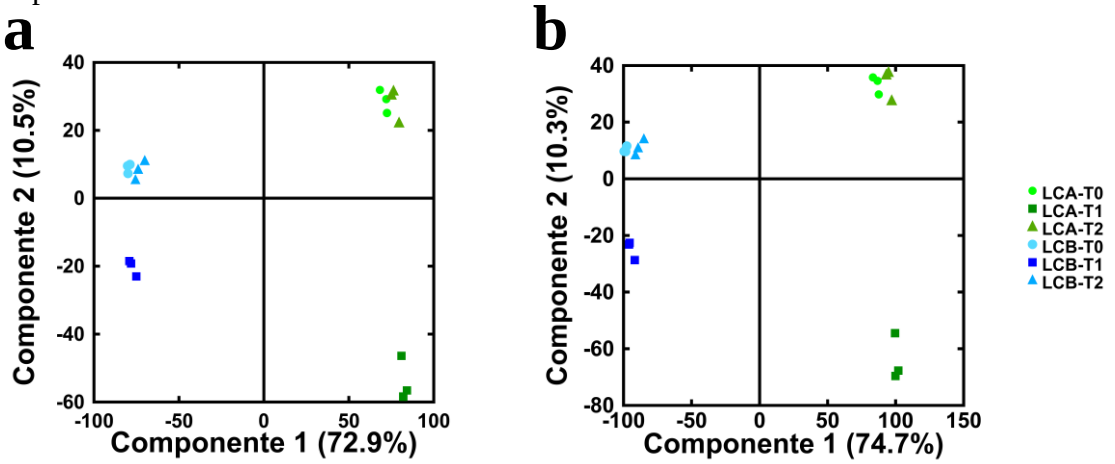



***Figura 4.21.*** *Análisis de escalado multidimensional (MDS) no supervisado de las muestras estudiadas mediante microarray en las líneas celulares de mieloma múltiple KMS12-BM y JJN-3 (LCA y LCB, respectivamente) bajo el tratamiento con amilorida (T1), TG003 (T2) o control (T0).* ***a)*** *MDS utilizando la referencia de normalización BA.* ***b)*** *MDS utilizando la referencia de normalización AEC.*

#### 4.2.2.1. Análisis de expresión génica diferencial con datos preprocesados mediante BA

Se partió de 35.345 genes para el análisis de la expresión génica diferencial. La principal ventaja del uso de la referencia de normalización BA fue la eliminación de redundancia y posibles ambigüedades en la determinación de la expresión de los genes, ya que esta técnica evitó la aparición de identificadores génicos duplicados. Mediante el análisis de expresión génica diferencial, tanto por el método *limma* como por el método *SAMr*, se pudo comprobar que, tal y como había sido previsto por el análisis no supervisado, el fármaco que introdujo un mayor cambio de expresión génica fue la amilorida (T1), tanto en la LCA como en la LCB, si bien los cambios inducidos por ambos compuestos fueron mayores en la LCA. Además, se observó que el número de genes diferencialmente expresados detectado por *SAMr* siempre fue mayor que el detectado por *limma* en todas las comparaciones (**Figura 4.22)**.

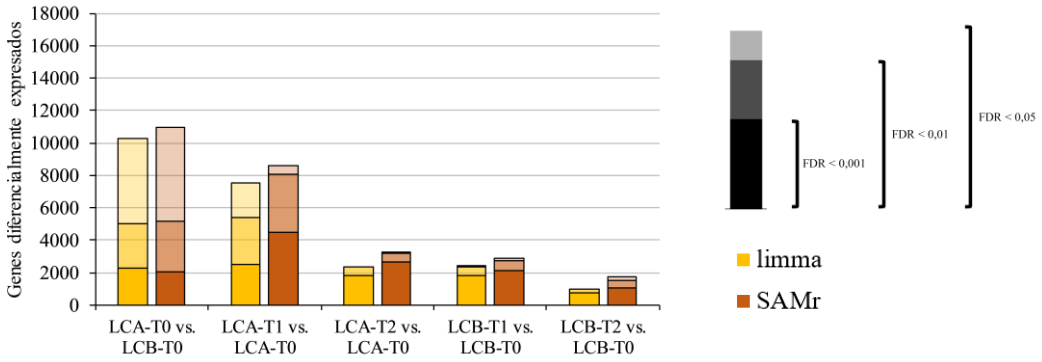



***Figura 4.22.*** *Resultados de los métodos de expresión diferencial para microarrays en las cinco comparaciones resultantes de las combinaciones de los datos procedentes de las líneas celulares CLA y CLB y los grupos de tratamiento T0, T1 y T2, utilizando la técnica de preprocesamiento BA.*

#### 4.2.2.2. Análisis de expresión génica diferencial con datos preprocesados mediante AEC

Hay que destacar que ya *a priori* podemos detectar una desventaja de la utilización de la referencia AEC, ya que la presencia de genes duplicados otorga a este método un factor de ambigüedad en la determinación de la expresión génica que no aparece con la referencia BA. Esto puede generar confusión al investigador debido a la posibilidad de tener dos o más valores de expresión diferentes para el mismo gen. Una segunda complicación aparece a la hora de determinar un gen como verdadero positivo debido a que, al ser estudiadas un mayor número de variables, el ajuste del *p*-valor para múltiples análisis va a penalizar en mayor medida a una lista de genes amplia que a una lista menor. Sin embargo, la presencia de estos duplicados estaría justificada en cuanto a que el fabricante provee un *probeset* que hibrida de manera específica con distintos transcritos de un mismo gen. De este modo, aunque a nivel génico esto pueda generar confusión, a nivel del estudio de isoformas es una herramienta útil. Sin embargo, en este estudio, para evaluar este problema se procedió al análisis con dos listas iniciales de genes, la lista completa con duplicados que constaría de 52.712 genes (por simplicidad hablaremos de genes, pero estrictamente hablando deben ser considerados como *probesets*), y una segunda lista de 37.810 genes donde los genes duplicados serían eliminados de manera aleatoria. En ambos casos se confirmaron las mismas tendencias en cuanto a la expresión génica diferencial que las que se describen para el método de preprocesamiento BA en el **Apartado 4.2.2.1**, de modo que el compuesto que indujo un mayor número de cambios de expresión diferencial fue la amilorida, y la mayor detección de genes diferencialmente expresados se produjo mediante el análisis con *SAMr*. Este análisis de expresión génica diferencial se recoge en la **Figura 4.23**.

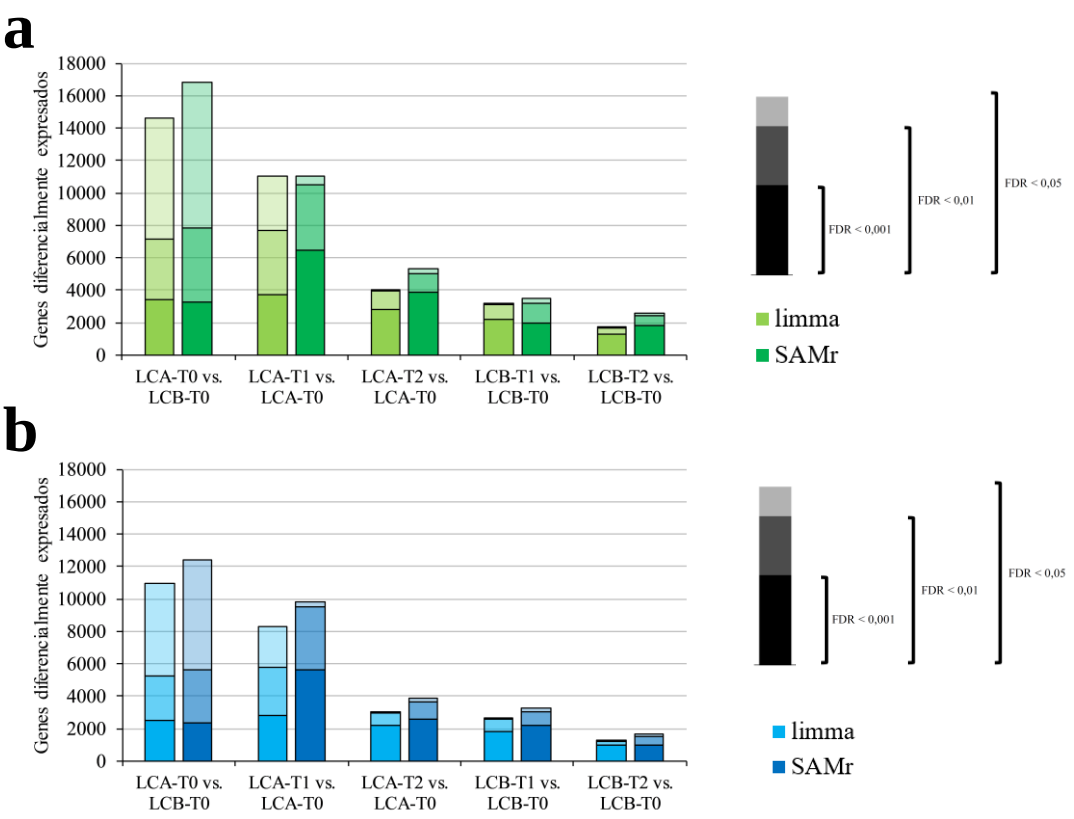



***Figura 4.23.*** *Resultados de los métodos de expresión diferencial para microarrays en las cinco comparaciones resultantes de las combinaciones de los datos procedentes de las líneas celulares LCA y LCB y los grupos de tratamiento T0, T1 y T2, utilizando la referencia AEC.* ***a)*** *Número de genes diferencialmente expresados sin eliminar genes duplicados.* ***b)*** *Número de genes diferencialmente expresados eliminando los genes duplicados.*

#### 4.2.2.3. Comparación de las técnicas de procesamiento de los microarrays.

El estudio comparativo entre las referencias BA y AEC se realizó sobre los resultados de expresión génica diferencial obtenidos con 35.345 genes de BA y con los 37810 genes no duplicados de AEC. Por tanto, en este punto se descartó la comparación utilizando AEC con genes duplicados. En lo que concierne a la detección de diferencias de expresión génica, a pesar de las diferentes técnicas de preprocesamiento y métodos de detección de expresión diferencial empleados, existe un alto grado de solapamiento en los resultados obtenidos. De este modo se detectó un solapamiento entre las técnicas y métodos que osciló desde los 8.626 genes en la comparación LCA-T0 *vs.* LCB-T0 (**Figura 4.24a**), a los 751 genes en la comparación LCB-T2 *vs.* LCB-T0 (**Figura 4.24e**).

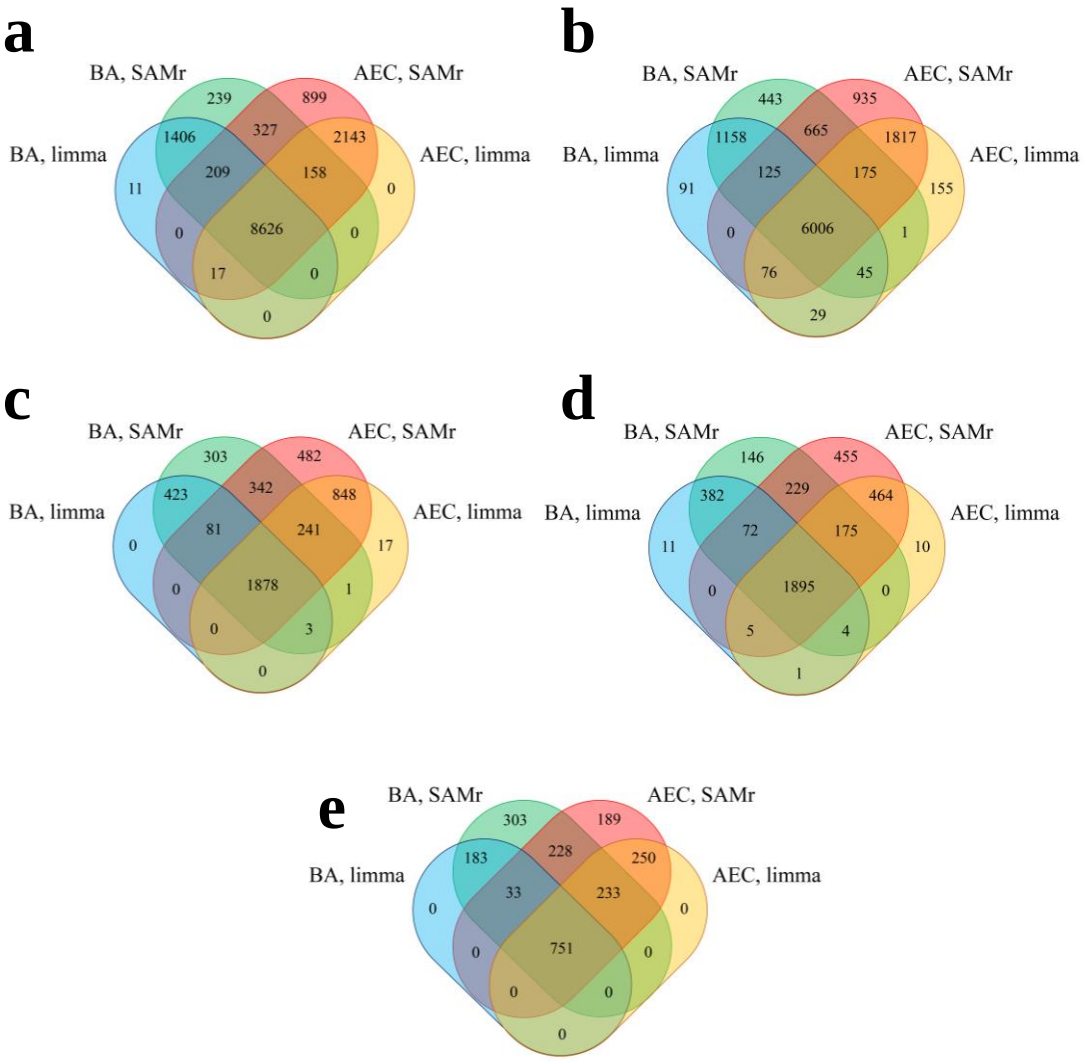



***Figura 4.24.*** *Diagrama de Venn utilizando las listas de genes sin duplicados analizados utilizando las referencias de normalización BA y AEC.* ***a)*** *Comparación LCA-T0 vs. LCB-T0,* ***b)*** *comparación LCA-T1 vs. LCA-T0.* ***c)*** *Comparación LCA-T2 vs. LCA-T0,* ***d)*** *comparación LCB-T1 vs. LCB-T0 y* ***e)*** *comparación LCB-T2 vs. LCB.T0.*

A continuación, se llevó a cabo el análisis de rendimiento de los métodos de análisis de la expresión génica diferencial del microarray a través del estudio de los 7 parámetros para la evaluación del desempeño utilizados en el **Capítulo 1**: VPP, ACC, FVP, FVN, VPN AUC y CCM (**Figura 4.25**).

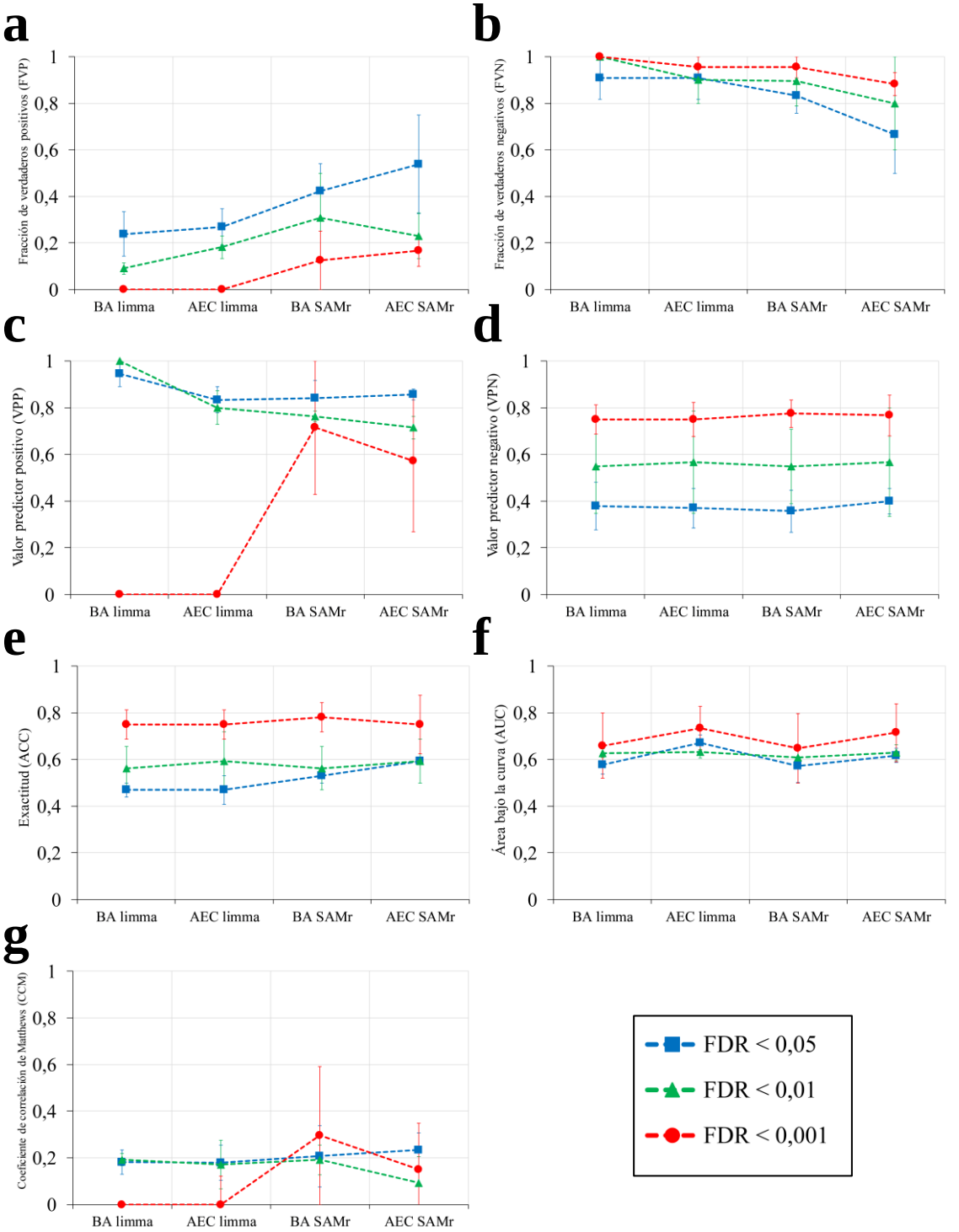



***Figura 4.25.*** *Análisis del desempeño de los métodos de expresión génica diferencial a través de la medición de 7 parámetros.* ***a)*** *Fracción de verdaderos positivos (FVP),* ***b)*** *fracción de verdaderos negativos (FVN).* ***c)*** *Valor predictor positivo (VPP),* ***d)*** *valor predictor negativo (VPN),* ***e)*** *exactitud (ACC),* ***f)*** *área bajo la curva (AUC) y* ***g)*** *coeficiente de correlación de Matthews (CCM). Los 7 parámetros se evaluaron a tres puntos de corte del FDR: 0,05, 0,01 y 0,001. Para cada parámetro se representa la mediana ± MAD de los cinco escenarios de análisis: LCA-T0 vs. LCB-T0, LCA-T1 vs. LCA-T0, LCA-T2 vs. LCA-T0, LCB-T1 vs. LCB-T0 y LCB-T2 vs. LCB-T0.*

En general, el comportamiento de los cuatro métodos de análisis no presentó grandes disparidades para la mayor parte de los 7 parámetros estudiados, excepto cuando se evaluó el nivel de significancia estadística a FDR < 0,001, donde el método *limma* no tuvo buen rendimiento. En lo que respecta a las principales diferencias detectadas a los niveles de significancia FDR < 0,05 y FDR < 0,01, estas se localizaron a nivel de los parámetros FVN y VPP (**Figura 4.25b y 4.25c**), donde el algoritmo *limma* junto con la referencia BA fue el método que presentó un mejor comportamiento, con valores superiores a 0,9 para ambos parámetros. Por su parte, los métodos que utilizaron el algoritmo *SAMr* destacaron en cuanto al valor del FVP obtenido, principalmente a FDR < 0,05, con valores superiores a 0,4 tanto con la referencia BA como con AEC (**Figura 4.25a**), y en los valores de ACC a FDR < 0,05, logrando cerca de un 60% de aciertos totales (ACC = 0,59). En lo que atañe al tipo de referencia utilizada en el preprocesamiento, se detectó una tendencia de los métodos basados en AEC a tener una mayor AUC, tanto a FDR < 0,05 como a FDR < 0,001, que los métodos con referencia BA (**Figura 4.25f**).

A la vista de estos resultados y de la ausencia de discordancias severas, se decidió proceder a la comparación con la RNA-seq considerando los cuatro métodos de análisis del microarray.

## 4.2.3. Comparación de los resultados de microarray y RNA-seq

Para llevar a cabo la comparación entre los resultados obtenidos con el microarray y la RNA-seq se determinó el número de genes comunes entre las dos tecnologías, de manera que para este análisis únicamente se utilizaron 31.471 genes. Con estas dos listas de genes se realizaron los estudios comparativos tanto a nivel de expresión génica cruda como de expresión diferencial.

### 4.2.3.1. Expresión génica cruda

El estudio del grado de asociación entre ambas plataformas se llevó a cabo mediante el coeficiente de correlación de Pearson (*r*), utilizando para esta comparación los 10 *pipelines* que obtuvieron los mejores rankings de precisión y exactitud en el **Capítulo 1**. Los resultados revelaron una correlación fuerte ($r > 0,8$) entre los datos de RNA-seq y microarray para las dos líneas celulares, con las dos referencias de normalización del microarray (BA o AEC), aunque BA mostró en ambos casos una correlación ligeramente superior con la RNA-seq. En la **Figura 4.26** se muestran los resultados de la correlación, así como la explicación del patrón de correlación en una muestra de ejemplo. Este patrón de correlación ya fue observado en estudios previos[228, 294], que comparan los datos de RNA-seq con los datos obtenidos mediante los microarrays de Affymetrix GeneChip HT HG-U133+ PM y HG-U133 Plus 2.0, respectivamente.

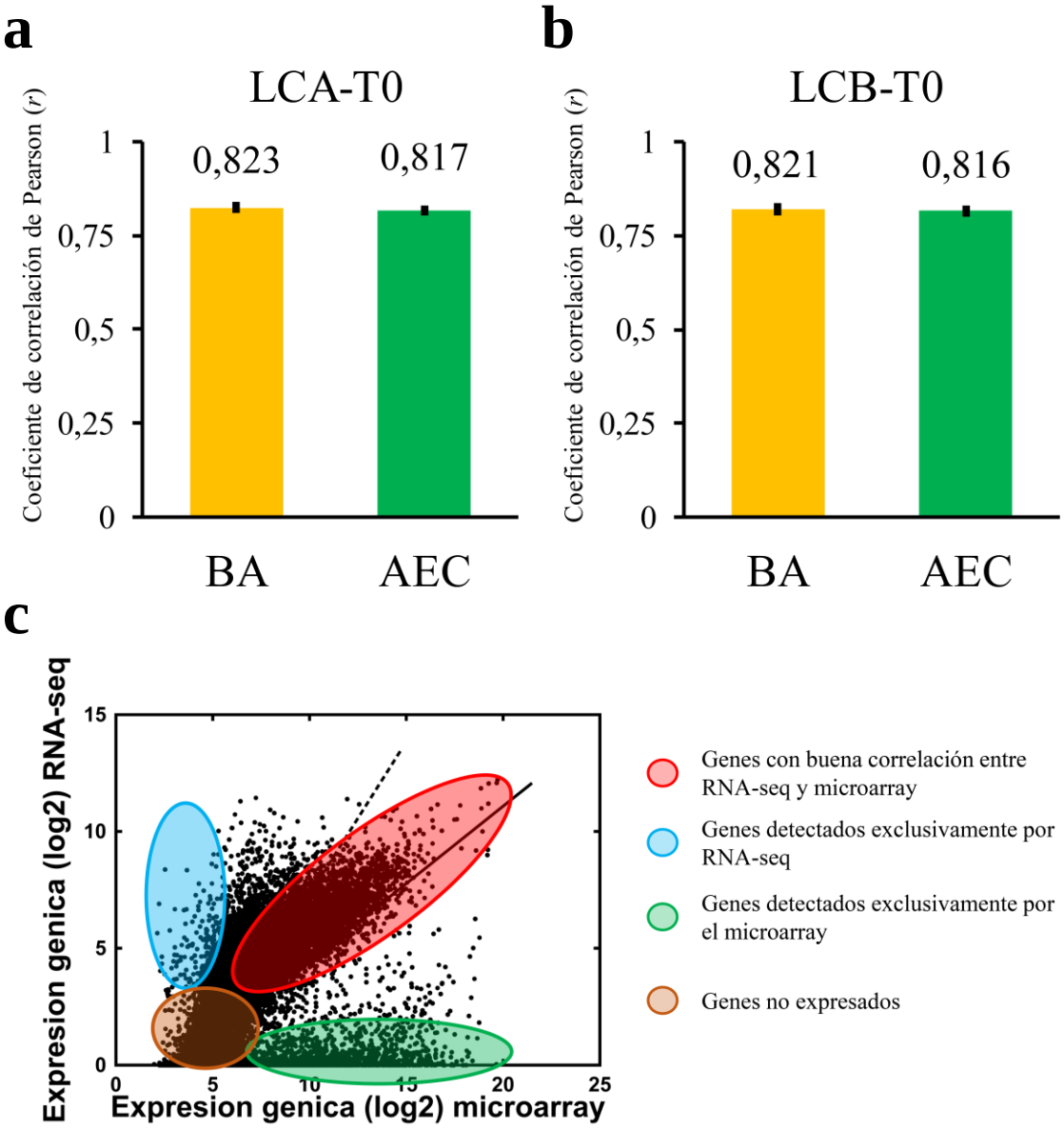



***Figura 4.26.*** *Resultados de la correlación entre los valores de expresión génica obtenidos mediante RNA-seq, y los obtenidos por el microarray.* ***a)*** *y* ***b)*** *Diagramas de barras de los coeficientes de correlación de Pearson (r) entre los datos de RNA-seq y los datos de microarray utilizando las referencias de normalización BA y AEC. Se muestra el valor promedio del r para las seis muestras control de* ***a)*** *LCA-T0 y* ***b)*** *LCB-T0, junto con su desviación estándar.* ***c)*** *Ejemplo de correlación entre la expresión detectada por la RNA-seq y el microarray con la explicación sobreimpresa de cada área.*

A continuación, se realizó el estudio de correlación de ambas plataformas con los datos de qRT-PCR utilizando los 32 genes previamente seleccionados descritos en la **Sección de Material y métodos**. En primer lugar, se comprobó que estos 32 genes cubriesen el rango dinámico tanto de la RNA-seq como del microarray. El resultado se puede comprobar en la **Figura 4.27**, donde se muestran resaltados en rojo los genes

seleccionados sobre la nube global de genes en negro. En general, puede observarse que el rango dinámico de ambas técnicas se cubrió de forma satisfactoria, independientemente del método de preprocesamiento empleado para el análisis del microarray.

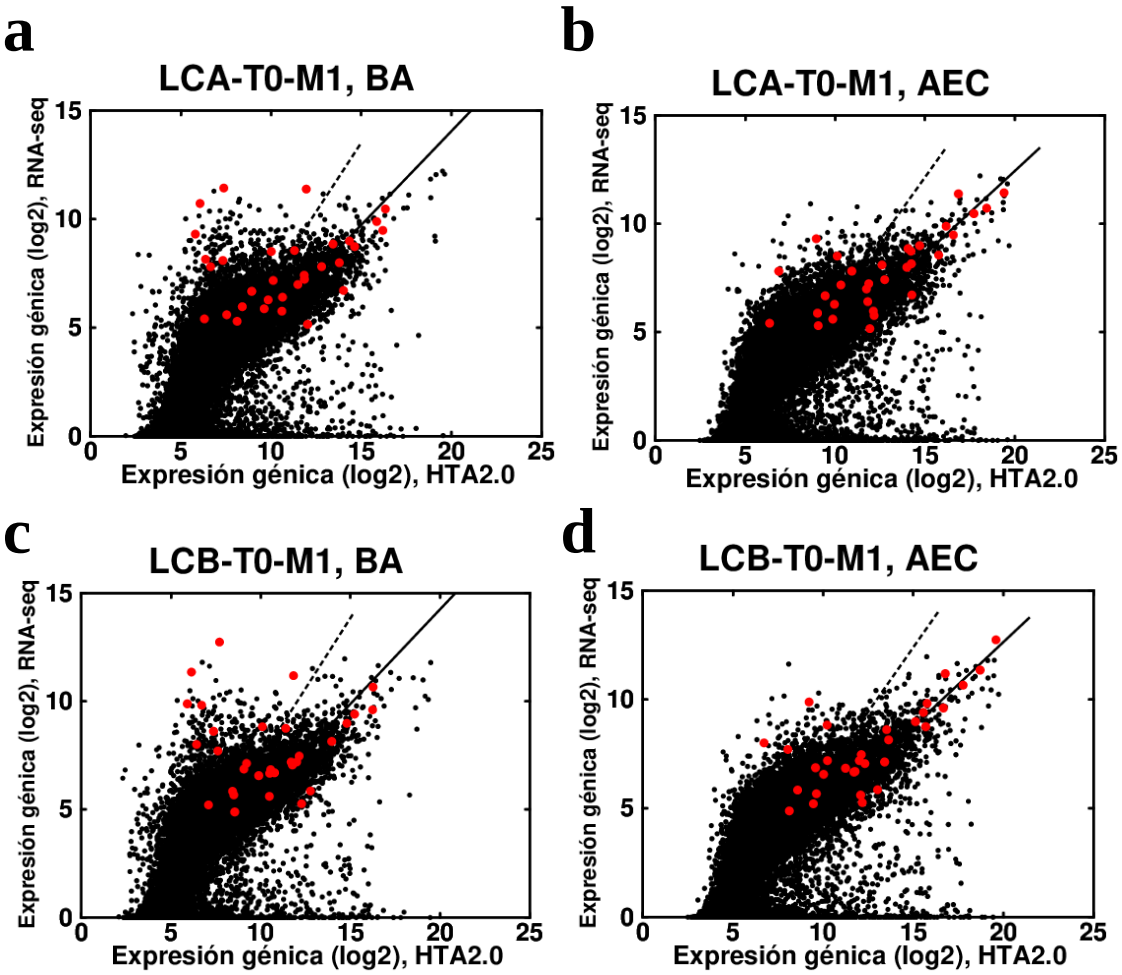



***Figura 4.27**. Posición de los genes seleccionados para validación a través de qRT-PCR (puntos rojos) sobre el gráfico de correlación entre el microarray HTA2.0 y la RNA-seq para las líneas celulares KMS12-BM (LCA) y JJN-3 (LCB). **a)** Correlación en LCA utilizando la referencia de normalización BA para los datos de microarray. **b)** Correlación en LCA utilizando la referencia de normalización AEC para los datos de microarray. **c)** Correlación en LCB utilizando la referencia de normalización BA para los datos de microarray. **d)** Correlación en LCB utilizando la referencia de normalización AEC para los datos de microarray.*

El análisis de correlación se realizó utilizando las seis muestras control presentes en el estudio. Se observó que los datos de RNA-seq se correlacionaron mejor con la expresión génica obtenida por qRT-PCR (**Figura 4.28e**) que por microarrays en las dos HMCLs (**Figura 4.28a y 4.28b**), aunque se detectaron diferencias en función de la referencia de normalización utilizada para este último, obteniendo mejores resultados la referencia AEC que la BA. No obstante, aun considerando ambas referencias, los resultados del microarray siempre estuvieron por debajo de los de RNA-seq, ya que mientras los coeficientes de correlación RNA-seq *vs.* qRT-PCR se situaron por encima

de 0,75, en el caso del microarray *vs.* qRT-PCR en ningún caso se logró un *r* superior a 0,6. Finalmente, se comprobó la correlación directamente entre la RNA-seq y el microarray observándose que el método BA obtenía unos coeficientes de correlación inferiores a los obtenidos por el método AEC (**Figura 4.28c y 4.28d**).

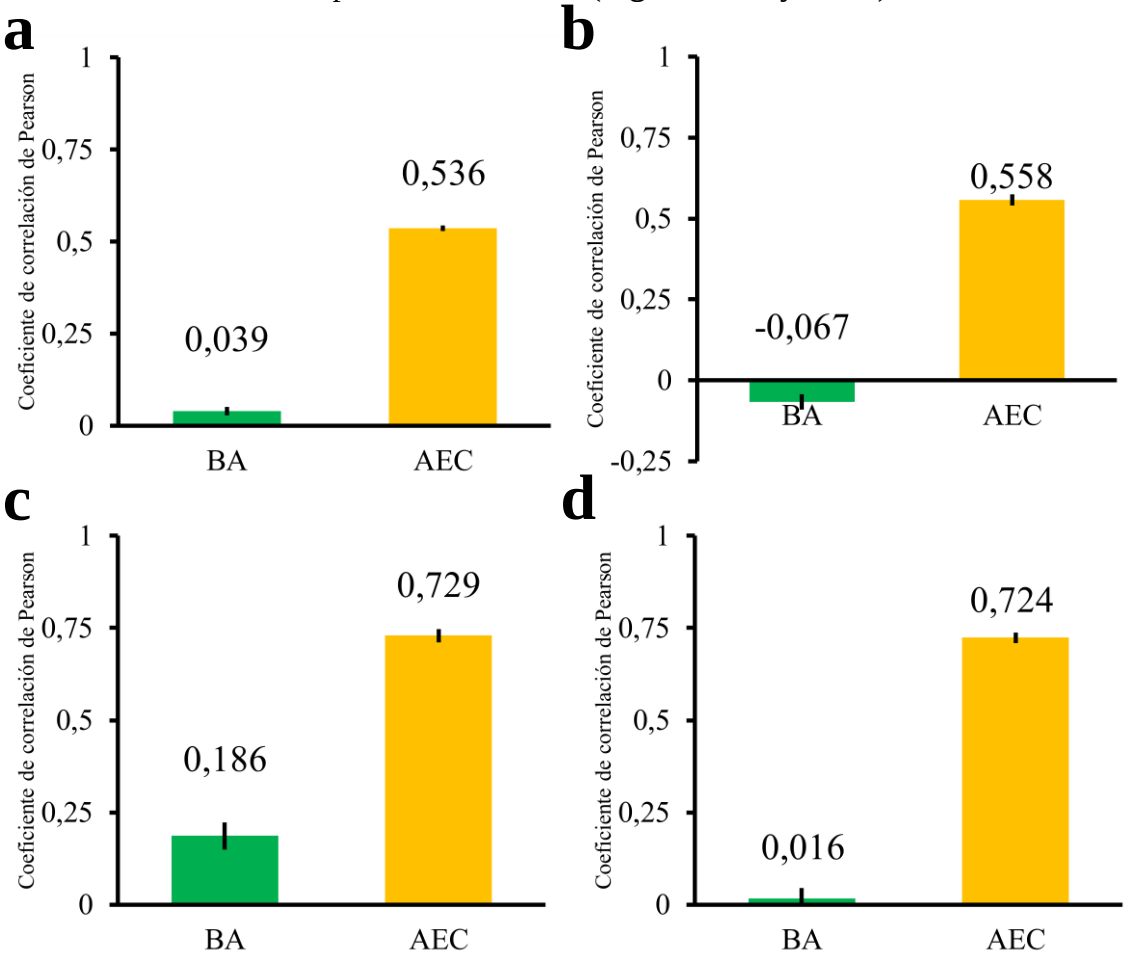



***Figura 4.28.*** *Estudio de correlación entre las tres técnicas de análisis de expresión génica empleadas en este trabajo. Los diagramas de barras muestran los valores promedio junto con la desviación estándar de los coeficientes de correlación de Pearson utilizando los 32 genes seleccionados para el análisis de la exactitud en la determinación del pipeline de RNA-seq para:* ***a)*** *expresión génica en $log_2$ del microarray frente a valores de ΔCt de qRT-PCR en la línea celular KMS12-BM (LCA);* ***b)*** *expresión génica en $log_2$ del microarray frente a valores de ΔCt de qRT-PCR en la línea celular JJN-3 (LCB);* ***c)*** *expresión génica en $log_2$ del microarray HTA2.0 frente expresión génica en $log_2$ de la RNA-seq en LCA;* ***d)*** *expresión génica en $log_2$ del microarray HTA2.0 frente expresión génica en $log_2$ de la RNA-seq en LCB. BA y AEC hacen referencia a la referencia de normalización utilizada por el microarray.*

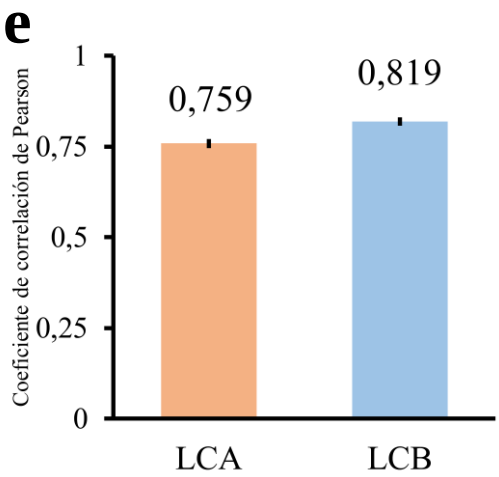



***Figura 4.28 (continuación).*** *Estudio de correlación entre las tres técnicas de análisis de expresión génica empleadas en este trabajo. Los diagramas de barras muestran los valores promedio junto con la desviación estándar de los coeficientes de correlación de Pearson utilizando los 32 genes seleccionados para el análisis de la exactitud en la determinación del pipeline de RNA-seq para:* ***e)*** *expresión génica en* $log_2$ *de RNA-seq frente a valores de ΔCt de qRT-PCR.*

Estos resultados nos llevaron a investigar en profundidad la causa de los bajos índices de correlación del microarray, para ello se procedió a la búsqueda de posibles genes con valores atípicos de expresión (*outliers*). Se consideró a tales efectos todo gen que se situase más allá del percentil 97,5 de acuerdo con la diferencia entre la función empírica de la distribución de la distancia robusta de Mahalanobis y una función teórica simulada. De este modo, identificamos que los genes *GAPDH* y *ACTB*, comúnmente tratados como *HKg* en estudios transcriptómicos y proteómicos, se comportaban como *outliers*, principalmente cuando fue aplicada la normalización BA. Del mismo modo, detectamos que los genes *RPS8*, *SNU13*, *YBX1*, y *RPL13A* también tenían un comportamiento atípico (**Figura 4.29**).

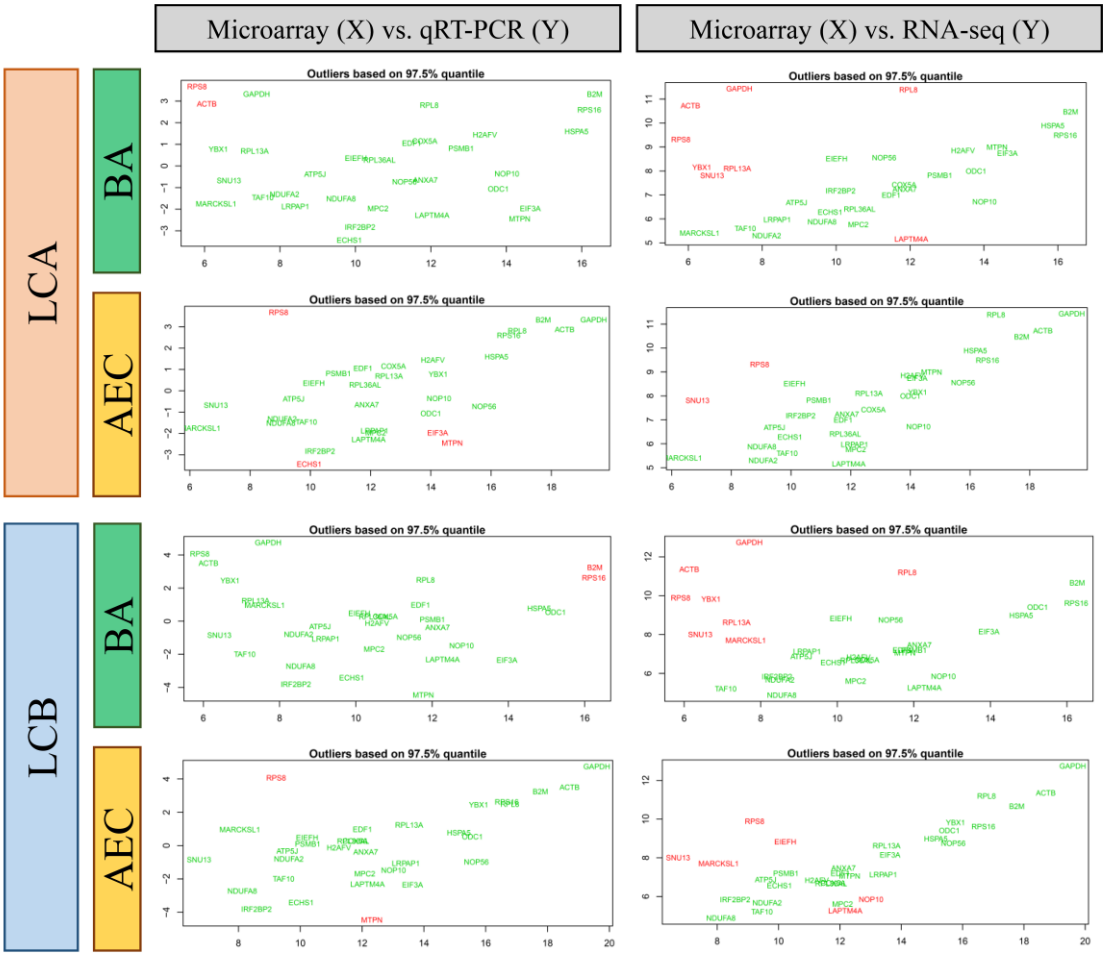



***Figura 4.29.*** *Detección de valores atípicos (outliers, en rojo) para las correlaciones entre microarray y qRT-PCR (izquierda), y microarray y RNA-seq (derecha) sobre las líneas celulares de mieloma múltiple KMS12-BM (LCA) y JJN-3 (LCB). Se muestran los resultados para los dos métodos de preprocesamiento del microarray, BA y AEC, utilizando 32 genes. Todos los ejemplos se han realizado con las muestras LCA-T0-M1 y LCB-T0-M1. En todos los casos el eje de abscisas (X) representa los datos procedentes del microarray, mientras el eje de ordenadas (Y) representa los datos de qRT-PCR (gráficos en la izquierda) o de RNA-seq (gráficos en la derecha).*

Pudimos constatar que este último grupo de genes comparte una función común ya que todos ellos están asociados a la función molecular de unión del ARN con cola de poliadenina y además codifican complejos intracelulares ribonucleoproteicos (Gene Ontology [GO][419]). Este mal comportamiento podría deberse a varias causas, tales como que las sondas correspondientes a estos genes hibriden en múltiples loci del genoma debido a la similitud de la secuencia del gen diana con otros genes del genoma, o que las sondas hibriden con genes que se encuentran incluidos en la secuencia del gen interrogado, como es el caso de *RPS8*, cuya secuencia incluye ARN pequeños nucleolares, como el *SNORD38A*, de manera que en ambos casos se producirían resultados ambiguos a la hora de cuantificar la expresión de estos genes. En lo que respecta al caso particular de los *HKg ACTB* y *GAPDH* sobre la referencia de

normalización BA, la ausencia de precisión a la hora de determinar su expresión apuntaría hacia algún tipo de singularidad en el procesado de ambos genes por parte del paquete *oligo* que conduciría a la infraestimación de la expresión de ambos genes.

Mediante el uso de una técnica robusta de correlación se comprobó la influencia de los genes *outlier* sobre estos análisis de correlación, de modo que los valores de los genes con valores atípicos fueron desestimados. De esta manera, se observó un aumento drástico de los coeficientes de correlación (*r*) en la correlación entre el microarray y la qRT-PCR, con un incremento en el caso de la referencia AEC desde un $r \sim 0{,}5$ en la correlación clásica a un $r \sim 0{,}9$ en la correlación robusta (**Figura 4.30**). En lo que respecta a la referencia BA, la técnica de correlación robusta no resultó efectiva quizá por la presencia de un mayor número de genes *outlier* (**Figura 4.30**). En el caso de la correlación entre el microarray y la RNA-seq también se observó una mejoría de los *r*, consiguiendo *r* superiores a 0,9 con la correlación robusta, tanto con la referencia BA como con la AEC.

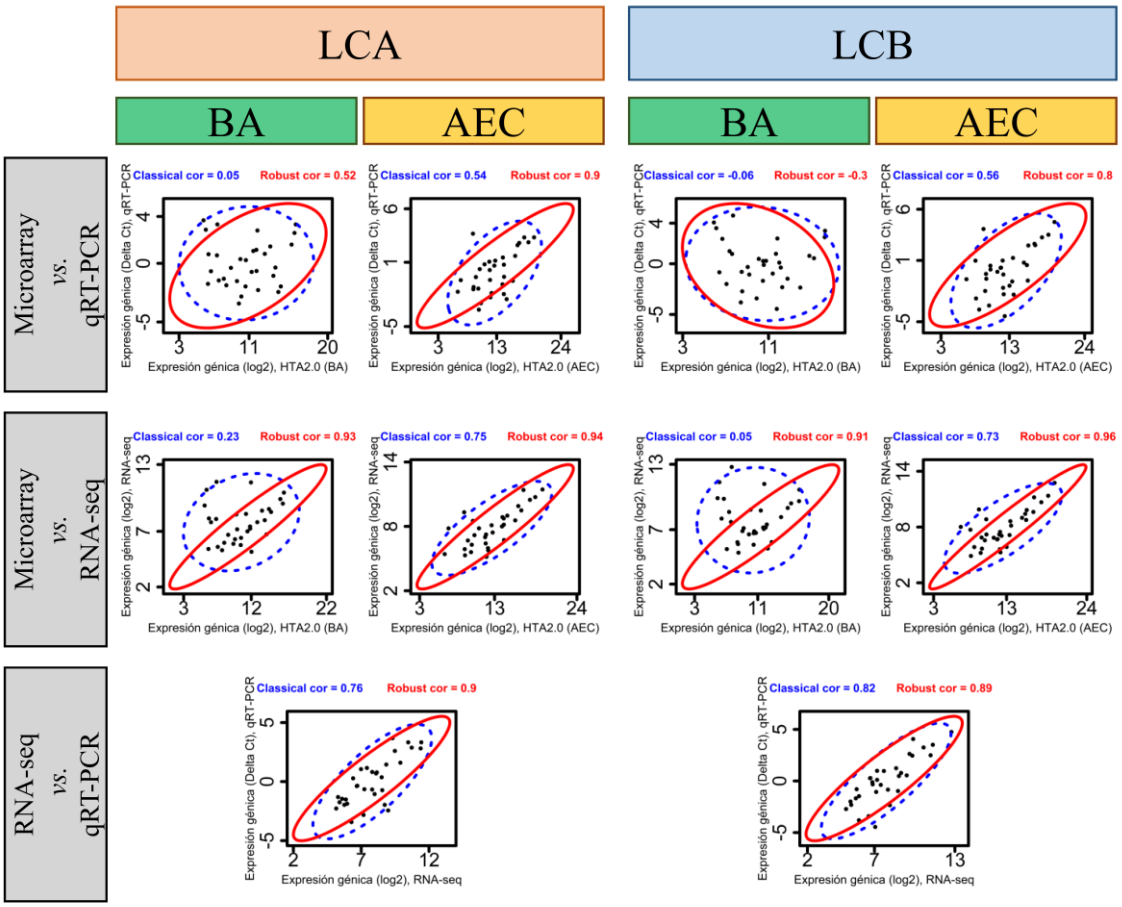



***Figura 4.30.*** *Correlación robusta (en rojo) y comparación con la correlación clásica de Pearson (en azul) para las tres técnicas de análisis de la expresión génica sobre las líneas celulares de mieloma múltiple KMS12-BM (LCA) y JJN-3 (LCB). Se muestran los resultados para los dos métodos de preprocesamiento del microarray, BA y AEC, utilizando 32 genes. Todos los ejemplos se han realizado con las muestras LCA-T0-M1 y LCB-T0-M1.*

En conjunto, los resultados expuestos en este apartado indican la superioridad de la RNA-seq sobre el microarray a la hora de la determinación de la expresión génica cruda, hallazgo que concuerda con lo expuesto en multitud de estudios previos tanto a nivel génico como a nivel transcriptómico[289, 293, 299, 313, 314]. Sin embargo, el microarray HTA2.0 podría funcionar de manera similar a la RNA-seq a nivel de la detección de la expresión génica cruda, siempre que se tengan en consideración posibles genes *outlier*. No obstante, la dificultad de la determinación de estos *outliers* en un análisis *de novo*, hace recomendable la utilización de la RNA-seq en estudios cuyo objetivo sea la determinación de la expresión génica cruda.

#### 4.2.3.2. Expresión génica diferencial

El estudio comparativo de la expresión génica diferencial entre el microarray HTA2.0 y la RNA-seq se llevó a cabo sobre los 31.471 genes comunes a ambas tecnologías, considerando los métodos de normalización BA y AEC del microarray. El análisis de la expresión diferencial del microarray se llevó a cabo utilizando los algoritmos empleados en el **Apartado 4.2.2**, *SAMr* y *limma*. En lo referente a la RNA-seq se procedió al reanálisis de los *pipelines* seleccionados en el estudio de la expresión diferencial (**Apartado 4.1.2**), considerando únicamente los 31.471 genes comunes indicados anteriormente. Este reanálisis de la expresión diferencial de RNA-seq se realizó con el método *limma trend*, al ser el método más equilibrado en todos los escenarios de análisis propuestos en este trabajo (**Apartado 4.1.2.2**).

En un primer paso se procedió al análisis de solapamiento a nivel de detección de genes diferencialmente expresados, utilizando como referencia la detección obtenida por la RNA-seq, a los niveles de significancia FDR < 0,05, FDR < 0,01 y FDR < 0,001, en los escenarios de análisis LCA-T0 *vs.* LCB-T0, LCA-T1 *vs.* LCA-T0, LCA-T2 *vs.* LCS-T0, LCB-T1 *vs.* LCB-T0 y LCB-T2 *vs.* LCB-T0. (**Figura 4.31**). Los porcentajes de solapamiento entre las dos tecnologías observados en los cinco escenarios se situaron siempre por debajo del 80% a todos los niveles de significancia, y fueron especialmente bajos en el escenario con un menor número de cambios de expresión génica (LCB-T2 *vs.* LCB-T0), en el que no llegó a alcanzarse el 40% de solapamiento (**Figura 4.31e**). El bajo grado de solapamiento a FDR < 0,001 en los escenarios correspondientes a las **Figura 4.31c**, **4.31d** y **4.31e** fue debido a la ausencia de resultados a este nivel con ambas tecnologías. Hay que indicar que estos datos deben ser considerados respecto a la referencia de la RNA-seq, y no deben tratarse como indicador de la bondad o el rendimiento de uno u otro método, sino como un indicador descriptivo de la similitud en la detección de la expresión génica diferencial de los métodos de análisis empleados.

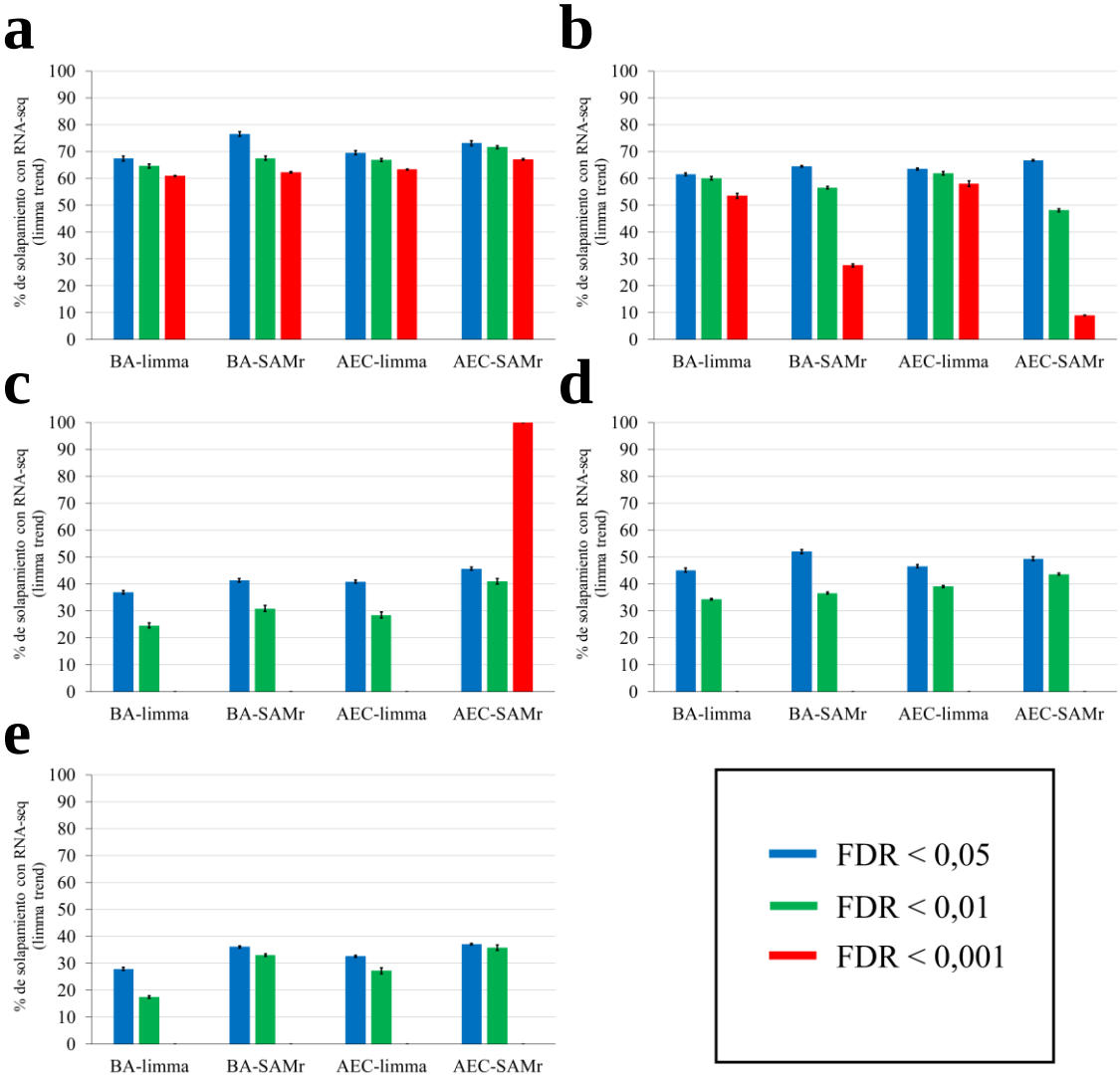



***Figura 4.31.*** *Porcentaje de solapamiento entre los métodos de expresión diferencial empleados en el estudio del microarray frente al método limma trend utilizado en el análisis de RNA-seq. Los datos representan el porcentaje mediano de solapamiento ± MAD considerando los resultados de limma trend en los cinco mejores pipelines como se indica en el* ***Apartado 4.1.2****. Cada panel corresponde a uno de los siguientes escenarios de análisis:* ***a)*** *contraste LCA-T0 vs. LCB-T0,* ***b)*** *contraste LCA-T1 vs. LCA-T0,* ***c)*** *contraste LCA-T2 vs. LCA-T0,* ***d)*** *contraste LCB-T1 vs LCB-T0 y* ***e)*** *contraste LCB-T2 vs. LCB-T0.*

En un segundo paso se procedió a determinar el desempeño de las dos tecnologías. Para ello se llevó a cabo la medición de los parámetros FVP, FVN, VPP, VPN, ACC, AUC y CCM utilizando como referencia el nivel de expresión de 32 genes medidos mediante qRT-PCR. Los resultados considerando de forma conjunta los cinco escenarios de análisis para cada uno de estos parámetros se recogen en la **Figura 4.32**.

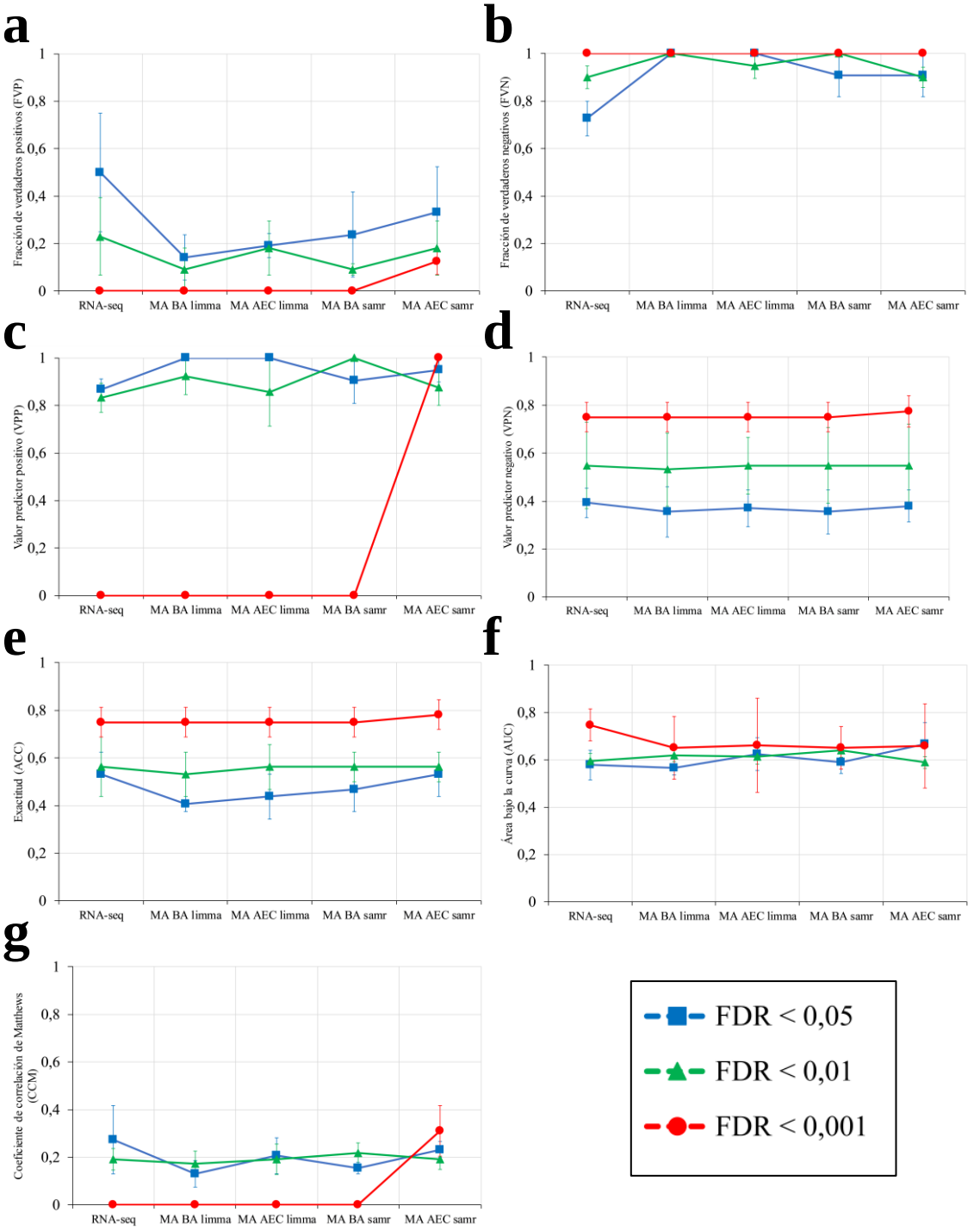



***Figura 4.32.*** *Análisis del desempeño de los métodos de análisis de la expresión génica diferencial mediante microarray (MA) y con el método limma trend utilizado en la RNA-seq a través de la medición de 7 parámetros:* ***a)*** *fracción de verdaderos positivos (FVP),* ***b)*** *fracción de verdaderos negativos (FVN),* ***c)*** *valor predictor positivo (VPP),* ***d)*** *valor predictor negativo (VPN),* ***e)*** *exactitud (ACC),* ***f)*** *área bajo la curva (AUC) y* ***g)*** *coeficiente de correlación de Matthews (CCM). Los 7 parámetros se evaluaron a tres puntos de corte del FDR: 0,05, 0,01 y 0,001. Para cada parámetro se representa la mediana ± MAD de los cinco escenarios de expresión génica diferencial: LCA-T0 vs. LCB-T0, LCA-T1 vs. LCA-T0, LCA-T2 vs. LCA-T0, LCB-T1 vs. LCB-T0 y LCB-T2 vs. LCB-T0.*

En general, en este análisis global no se observa una superioridad clara de ninguna de las tecnologías ni de ningún método de análisis. La RNA-seq se impone al microarray a nivel de los parámetros FVP y MCC, principalmente a FDR < 0,05, con unos valores de 0,5 y 0,27, respectivamente (**Figura 4.32a** y **4.32g** y **Anexo 8**). Por su parte el microarray obtiene mejores resultados que la RNA-seq tanto para el FVN como en el VPP (**Figura 4.32b** y **4.32c** y **Anexo 8**), sin considerar el nivel de significancia FDR < 0,001, ya que como se indicó en el análisis de detección, no funcionó correctamente con ninguna tecnología. En cuanto al resto de parámetros, se obtuvo un empate técnico con valores muy similares entre la RNA-seq y al menos uno de los métodos de análisis del microarray (**Figura 4.32d, 4.32e** y **4.32f** y **Anexo 8**).

En un tercer paso se procedió al estudio del rendimiento de las dos tecnologías en 8 aproximaciones correspondientes a los tres niveles de significancia y a los cinco escenarios de análisis considerados en este trabajo por separado, calculando los rankings de los 7 parámetros (FVP, FVN, VPP, VPN, ACC, AUC y CCM) en cada una de estas aproximaciones (**Figura 4.33**).

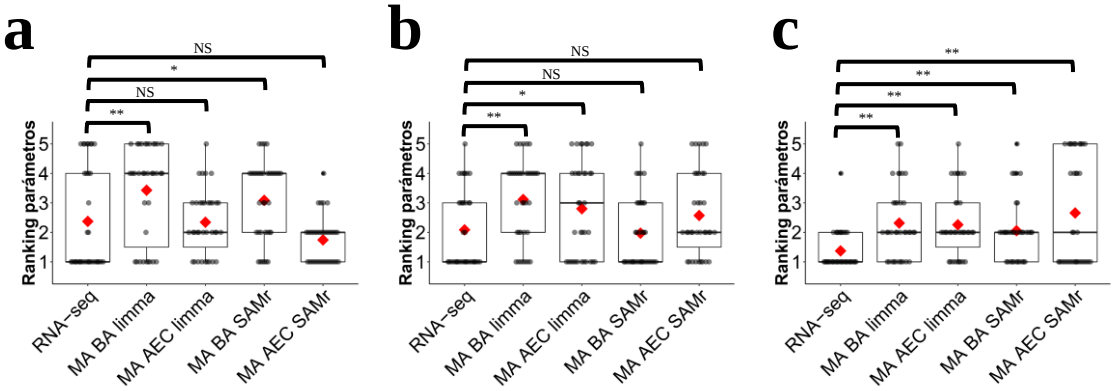



***Figura 4.33****. Rendimiento de las tecnologías de análisis de la expresión génica, RNA-seq y microarray HTA2.0, considerando 8 aproximaciones analíticas correspondientes a tres niveles de significancia,* ***a)*** *FDR < 0,05,* ***b)*** *FDR < 0,01 y* ***c)*** *FDR < 0,001, y a cinco escenarios de análisis,* ***d)*** *LCA-T0 vs. LCB-T0,* ***e)*** *LCA-T1 vs. LCA-T0,* ***f)*** *LCA-T2 vs. LCA-T0,* ***g)*** *LCB-T1 vs. LCB-T0 y* ***h)*** *LCB-T2 vs. LCB-T0. Cada caja contiene los rankings obtenidos por 7 parámetros, FVP, FVN, VPP, VPN, ACC, AUC y CCM, para cada método de análisis y aproximación. El rombo rojo representa la media de los rankings obtenidos por estos 7 parámetors. NS = FDR > 0,05 en el test post-hoc de Dunn (resultados estadísticamente no significativos); * = FDR < 0,05 en el test post-hoc de Dunn; ** = FDR < 0,01 en el test post-hoc de Dunn; *** = FDR < 0,001 en el test post-hoc de Dunn.*

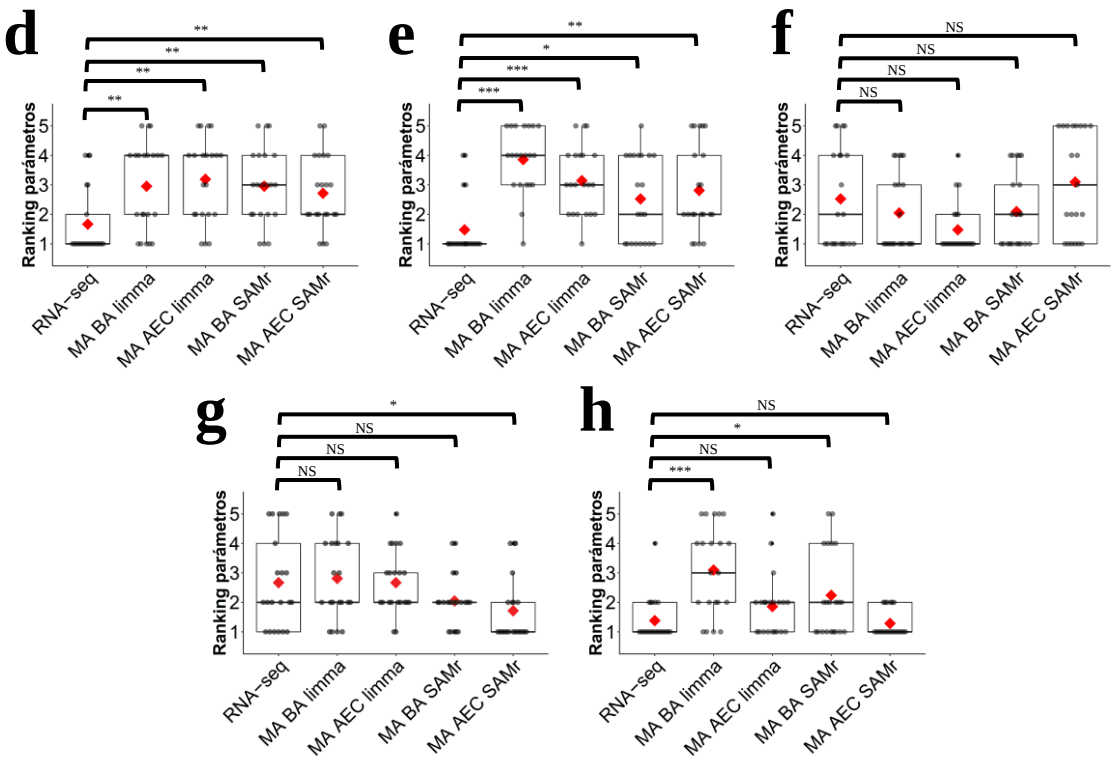



***Figura 4.33 (continuación)***. *Rendimiento de las tecnologías de análisis de la expresión génica, RNA-seq y microarray HTA2.0, considerando 8 aproximaciones analíticas correspondientes a tres niveles de significancia,* ***a)*** *FDR < 0,05,* ***b)*** *FDR < 0,01 y* ***c)*** *FDR < 0,001, y a cinco escenarios de análisis,* ***d)*** *LCA-T0 vs. LCB-T0,* ***e)*** *LCA-T1 vs. LCA-T0,* ***f)*** *LCA-T2 vs. LCA-T0,* ***g)*** *LCB-T1 vs. LCB-T0 y* ***h)*** *LCB-T2 vs. LCB-T0. Cada caja contiene los rankings obtenidos por 7 parámetros, FVP, FVN, VPP, VPN, ACC, AUC y CCM, para cada método de análisis y aproximación. El rombo rojo representa la media de los rankings obtenidos por estos 7 parámetors. NS = FDR > 0,05 en el test post-hoc de Dunn (resultados estadísticamente no significativos); * = FDR < 0,05 en el test post-hoc de Dunn; ** = FDR < 0,01 en el test post-hoc de Dunn; *** = FDR < 0,001 en el test post-hoc de Dunn.*

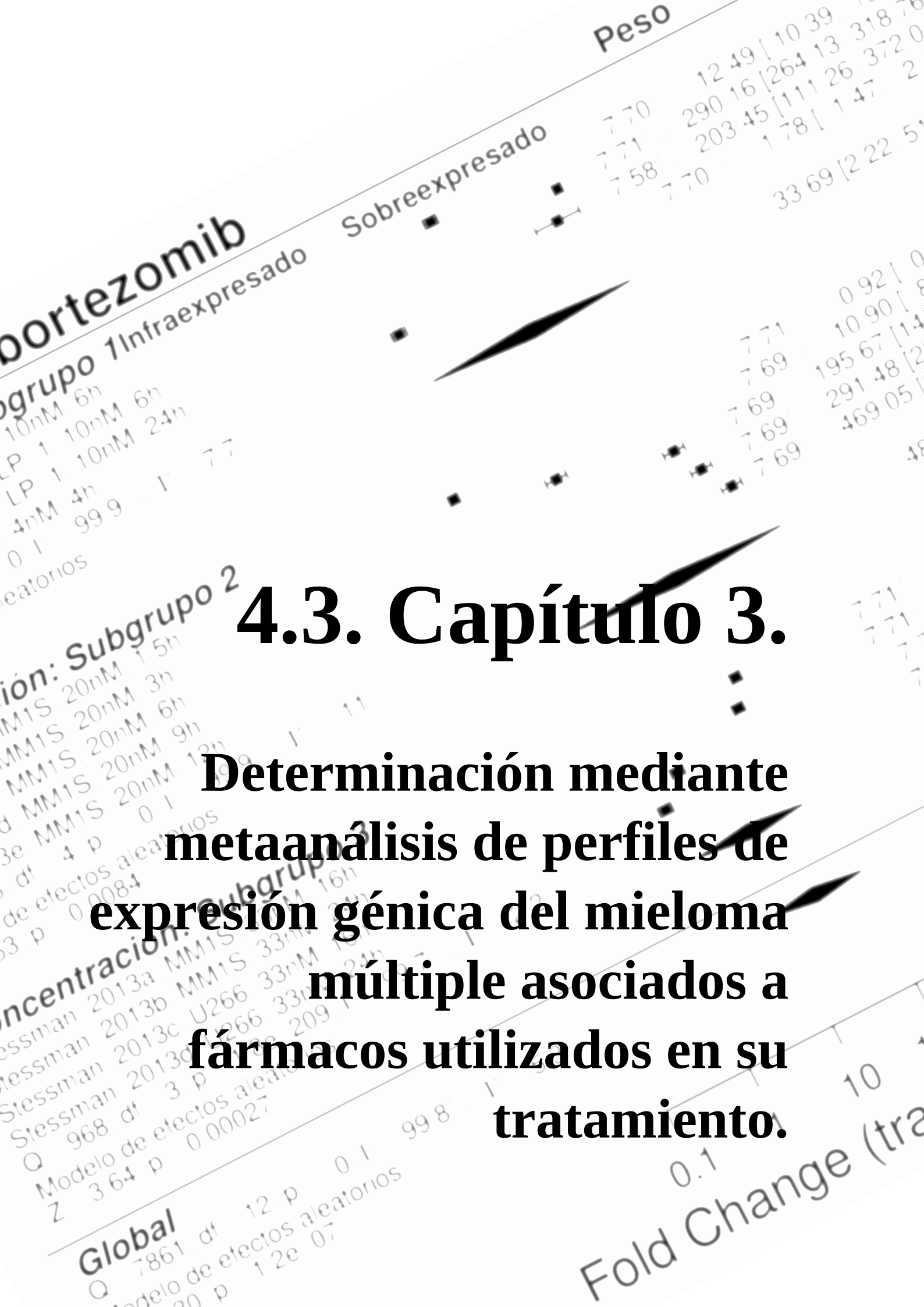
En el caso de las tres aproximaciones relativas a los niveles de significancia, se detectaron diferencias estadísticamente significativas en la aproximación que considera un FDR < 0,001, de manera que la RNA-seq a este nivel superó a todos los métodos de análisis del microarray (**Figura 4.33c**). En lo que respecta a los niveles de significancia FDR < 0,05 y FDR < 0,01, la RNA-seq presentó un comportamiento superior de forma estadísticamente significativa a algunos de los métodos, como a los basados en la referencia BA en el caso de la aproximación a FDR < 0,05, o a los métodos basados en el algoritmo *limma* en la aproximación a FDR < 0,01, pero mantuvo un comportamiento similar al resto de los métodos (**Figura 4.33a y 4.33b**).

En lo concerniente a los cinco escenarios de análisis, la RNA-seq mostró una superioridad estadísticamente significativa frente al microarray en los escenarios que presentaron un mayor número de cambios de expresión génica (LCA-T0 *vs.* LCB-T0

[**Figura 4.33d**] y LCA-T1 *vs.* LCA-T0 [**Figura 4.33e**]), considerando cualquiera de los métodos de análisis del microarray. En los escenarios con un número de cambios de expresión génica intermedio, ambas técnicas tienen un comportamiento similar, llegando a vencer, el método de análisis del microarray basado en la referencia AEC y el algoritmo *SAMr*, a la RNA-seq (FDR = 0,023) en el escenario LCB-T1 *vs.* LCB-T0 (**Figura 4.33g**). Finalmente, en el escenario con un menor número de cambios de expresión génica, LCB-T2 *vs.* LCB-T0, la RNA-seq supera a los dos métodos basados en la referencia BA, pero muestra un comportamiento similar a los métodos basados en AEC (**Figura 4.33h**).

Estos resultados muestran, por tanto, que la RNA-seq fue claramente superior en tres de las 9 aproximaciones estudiadas, mientras que en otras tres fue superior a, al menos, dos de los métodos de análisis del microarray. Solo hubo ligera ventaja del microarray en una de las aproximaciones y total igualdad en las dos restantes. Por tanto, estos análisis otorgan cierta preeminencia a la RNA-seq a la hora de la detección de manera exacta y precisa de la expresión génica diferencial, principalmente si se considera el contraste estadístico de muestras con diferencias en la expresión génica medias o altas. Esta superioridad de la RNA-seq ya ha sido previamente reportada sobre microarrays de Agilent[284], sobre 3′ arrays de Affymetrix[297, 304] y sobre arrays exónicos de Affymetrix[309]. En el caso de los microarrays de transcriptoma como el HTA2.0, estudios recientes como el de Nazarov[282] reportan, en contra de lo expuesto en este trabajo, un alto grado de solapamiento entre las dos tecnologías, tanto a nivel de detección de patrones de expresión génica como a nivel de expresión diferencial de genes codificantes de proteínas. Otros trabajos como el de Yu[286], sin embargo, presentan una buena concordancia con lo expuesto en este Capítulo al detectar una baja similitud de este microarray con la RNA-seq, indicando que esto podría ser debido a la estimación a la baja del FC por el HTA2.0. Sin embargo, a pesar de esta contraposición, la realidad es que ambos trabajos confirman lo expuesto en este Capítulo, ya que la aproximación analítica de Nazarov se situaría entre los escenarios con un número intermedio de cambios de expresión génica de este trabajo, donde las dos plataformas presentan una alta concordancia en sus resultados.

Con todo lo expuesto, se podría concluir que de manera general la RNA-seq tiene un mejor rendimiento que el microarray HTA2.0, tanto a nivel de detección de la expresión génica cruda como de la expresión génica diferencial, lo que unido a la detección de nuevos genes y la posibilidad analizar los eventos de *splicing* alternativo o las fusiones génicas, la convierten en una tecnología óptima para el análisis de expresión génica. No obstante, bajo ciertas circunstancias como las que se exponen en este estudio, y siempre que se quiera trabajar con genes bien establecidos, el microarray puede ser una buena alternativa analítica.



# 4.3. Capítulo 3.

## Determinación mediante metaanálisis de perfiles de expresión génica del mieloma múltiple asociados a fármacos utilizados en su tratamiento.

Se realizó una revisión sistemática con metaanálisis para determinar el tamaño del efecto producido sobre la expresión génica de 20 compuestos farmacológicos utilizados habitualmente en el tratamiento del MM según la Sociedad Americana de Cáncer (https://www.cancer.org/). Estos 20 compuestos se clasifican en las siguientes familias según su efecto farmacológico:

a) Quimioterapia convencional: melfalán, vincristina, ciclofosfamida, etopósido, doxorrubicina y bendamustina.
b) Corticoides: dexametasona y prednisona
c) Agentes inmunomoduladores: talidomida, lenalidomida y pomalidomida
d) Inhibidores del proteasoma: bortezomib, carfilzomib e ixazomib
e) Inhibidores de las deacetilasas de histonas: panobinostat
f) Agentes hipometilantes: azacitidina y decitabina
g) Anticuerpos monoclonales: daratumumab y elotuzumab

Además, se decidió añadir a esta revisión sistemática dos compuestos adicionales, como son el interferón γ, debido a su amplia utilización en el tratamiento del MM durante más de tres décadas, y el compuesto JQ1. Aunque el JQ1 es de uso exclusivamente experimental en MM, se determinó su inclusión debido al elevado número de estudios realizados en la literatura, permitiendo así la incorporación al presente trabajo de la familia de fármacos inhibidores de bromodominio.

Como punto de partida para este análisis, se realizó una búsqueda *online* de series experimentales en las que se hubiese empleado alguno de estos 21 compuestos en monoterapia sobre líneas celulares de MM. Esta búsqueda bibliográfica fue cerrada a fecha 24 de marzo de 2.017, obteniendo como resultado la inclusión de 9 fármacos para el estudio mediante metaanálisis. Los 12 compuestos restantes fueron descartados debido a la ausencia de al menos dos series que cumpliesen los criterios de inclusión y exclusión definidos en la **Sección de Material y métodos**. El resumen de la búsqueda de series experimentales correspondiente a los fármacos excluidos y las causas de su exclusión se recoge en el **Anexo 9**. Por su parte, los 9 fármacos seleccionados para realizar una revisión sistemática con metaanálisis fueron los siguientes:

1) Melfalán
2) Dexametasona
3) Bortezomib
4) Lenalidomida
5) Pomalidomida
6) Panobinostat
7) Azacitidina
8) Decitabina
9) JQ1

Los resultados para cada compuesto serán expuestos en los distintos apartados de este **Capítulo 3**, y de manera adicional se procederá al análisis de los compuestos moduladores del *splicing* alternativo amilorida y TG003, empleados en los **Capítulos 1**

**y 2** de este trabajo, así como del único estudio recogido en las búsquedas bibliográficas para el interferón γ.

### 4.3.1. Melfalán

En este apartado se presentará la revisión sistemática con metaanálisis de las series de expresión génica con muestras de HMCLs tratadas con melfalán en monoterapia. Las búsquedas se llevaron a cabo en los repositorios de datos GEO, ArrayExpress y SRA. Este proceso de búsqueda condujo a la detección 23 series en GEO, 13 series en ArrayExpress y cinco muestras en SRA. Las cinco muestras localizadas en SRA correspondieron a una única serie. El número final de series que fueron seleccionadas para su revisión detallada fue de 24 tras la eliminación de los elementos duplicados en los tres repositorios. Solamente tres de estas 24 series fueron seleccionadas para el posterior metaanálisis tras la evaluación de los criterios de inclusión y exclusión previamente establecidos. En un último paso, se comprobó si las muestras empleadas en cada una de las tres series eran homogéneas en cuanto a las concentraciones de fármaco empleadas, el tiempo de tratamiento o la utilización de distintas líneas celulares, de manera que la presencia de heterogeneidad en alguno de estos tres parámetros permitiría la subdivisión de dicha serie en dos o más estudios. No se halló ninguna subdivisión en las tres series seleccionadas, por tanto, el número de estudios incluidos de manera definitiva en el metaanálisis fue también tres. El diagrama de flujo que se muestra en la **Figura 4.34** detalla el esquema de selección de estudios para el metaanálisis de melfalán en función de los diferentes criterios de inclusión y exclusión.

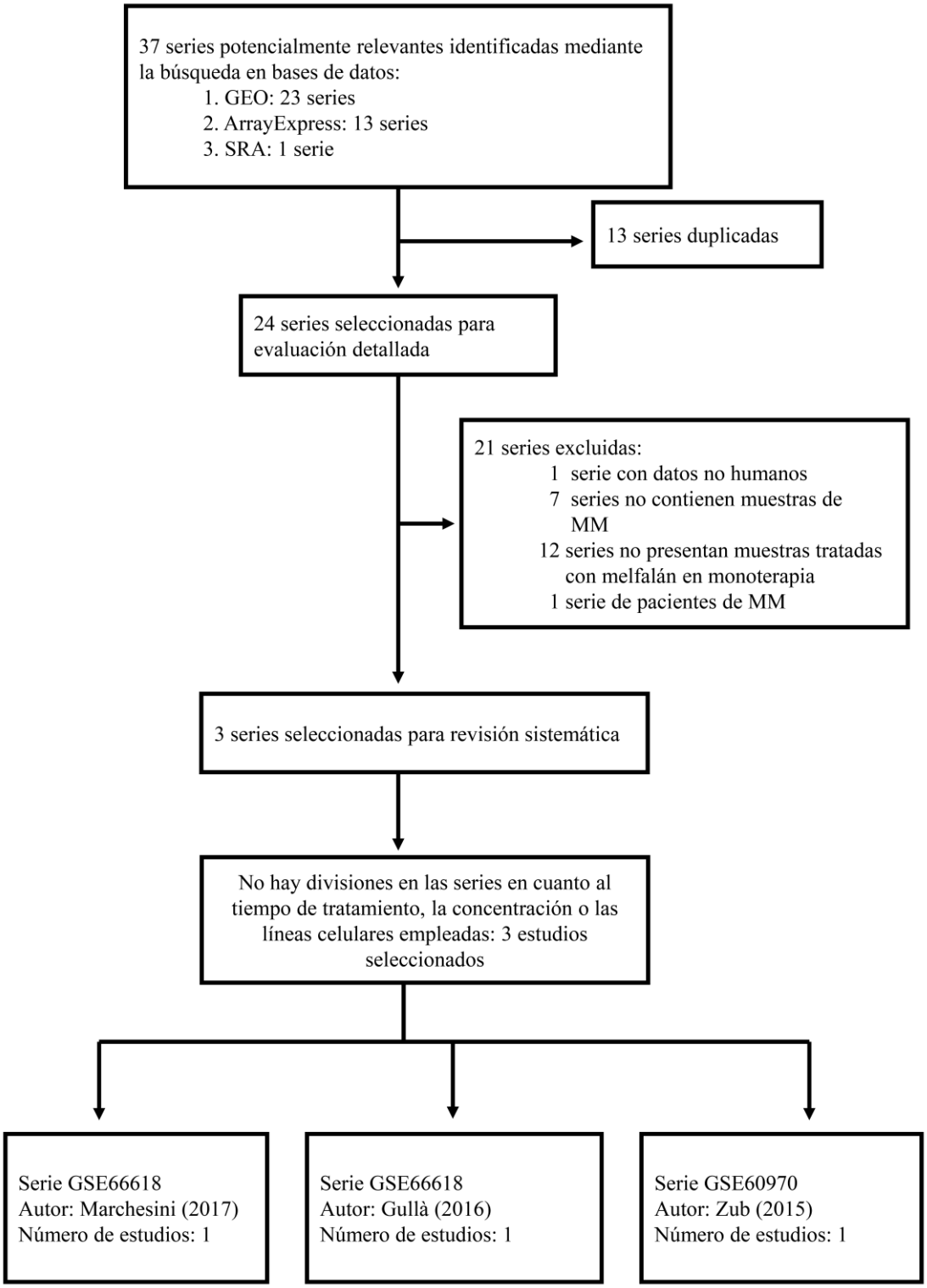



***Figura 4.34**. Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en HMCLs tratadas con melfalán.*

Una vez concluido el proceso de búsqueda sistemática, se procedió a la clasificación de los tres estudios seleccionados en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de melfalán. De este modo, se establecieron como puntos de corte del tiempo de tratamiento las 8 y las 12 horas (mediana de 10 horas), mientras que los puntos de corte para la concentración de melfalán fueron establecidos a 25 y 75 µM (mediana de 50 µM). Los resultados del agrupamiento en subgrupos se recogen en la **Tabla 4.4**.

***Tabla 4.4.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con melfalán.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (µM) |
|---|---|---|---|---|---|---|
| GSE83712 | Marchesini (2017)[424] | JJN-3 | Illumina HiSeq 2000 | 4 | 10 | 25 |
| GSE66618 | Gullà (2016)[190] | U266 | Affymetrix Human Gene 1.0ST | 2 | 12 | 100 |
| GSE60970 | Zub (2015)[191] | RPMI-8226 | Illumina HumanHT-12 V3.0 expression beadchip | 10 | 6 | 50 |

*En verde, estudios seleccionados para el subgrupo de tiempos o concentraciones bajos; en amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedias; en rojo, estudios seleccionados para el subgrupo de tiempos o concentraciones altas.*

A continuación, se llevó a cabo la extracción de los genes candidatos para ser estudiados mediante metaanálisis, de manera que, por un lado, se seleccionaron los genes con un valor absoluto del FC mayor a 1,5 en los tres estudios, y, por otro lado, los genes con un valor absoluto del FC mayor a 1,5 en todos los estudios de, al menos, uno de los subgrupos de tiempo o concentración, excluyéndose los subgrupos que solamente constasen con un estudio- Este procedimiento de selección de genes aparece explicado en detalle en la **Sección de Material y métodos**. Como resultado fueron seleccionados un total de 1.460 genes. La distribución del ln(FC) de estos 1.460 genes en los tres estudios se recoge en la **Figura 4.35**, donde se observa una alta homogeneidad entre los estudios de Marchesini (2017) y Gullà (2016), y una mayor compresión de los ln(FC) en el estudio de Zub (2015), probablemente debida al menor tiempo de tratamiento en relación a los otros dos estudios. Estas posibles diferencias debidas al tiempo de tratamiento se analizarán a continuación mediante el uso de técnicas de metaanálisis por subgrupos.

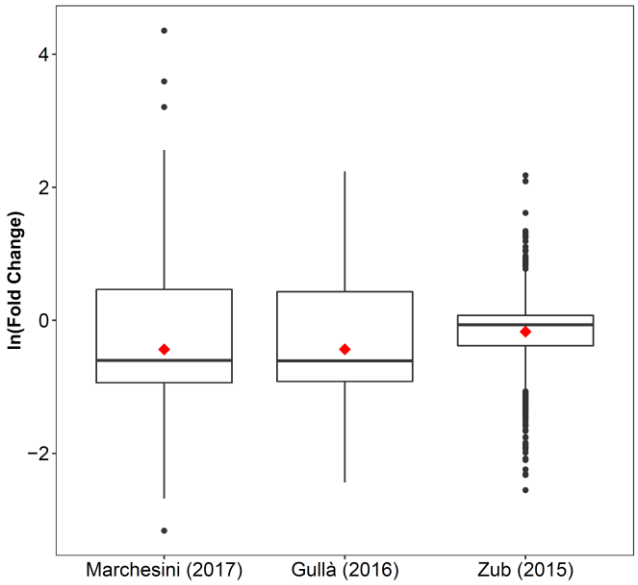


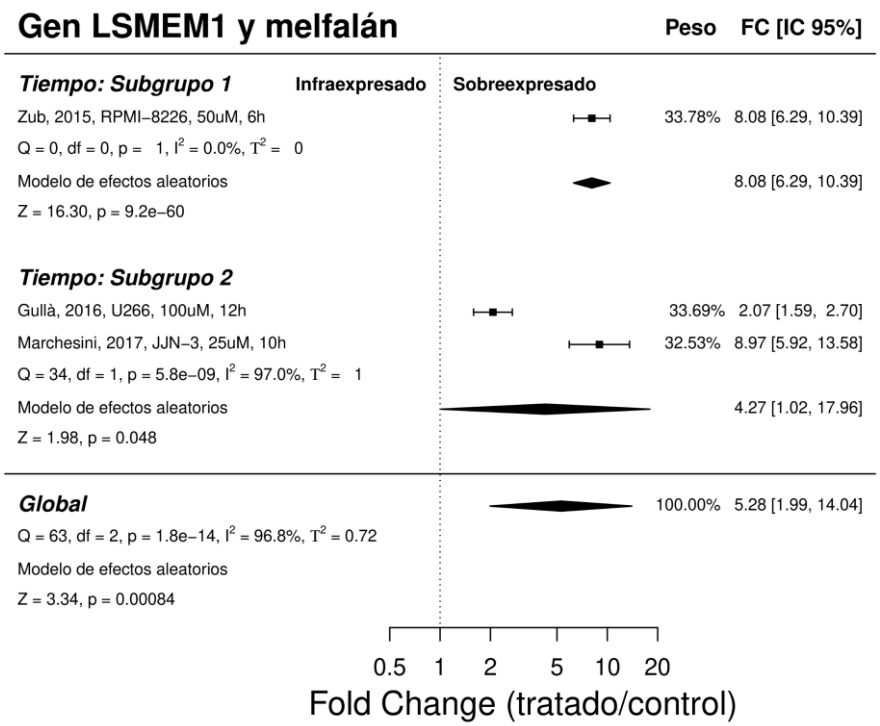

***Figura 4.35.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC]) de los 1.460 genes en los tres estudios seleccionados para el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con melfalán. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.1.1. Metaanálisis por subgrupos: tiempo de tratamiento

Se establecieron dos subgrupos de estudios considerando el tiempo de tratamiento con melfalán. El primer subgrupo (G1) comprendió los estudios con un tiempo de tratamiento menor o igual a las 8 horas, criterio únicamente cumplido por el estudio de Zub (2015). El segundo subgrupo (G2) recogió los estudios cuyo tiempo de tratamiento fue superior a las 8 horas e inferior a las 12 horas, de modo que en este caso fueron dos estudios los que cumplieron los requisitos: Marchesini (2017) y Gullà (2016). Ninguno de los estudios utilizó tiempos de tratamiento superiores a las 12 horas, por lo esté subgrupo no fue considerado.

El metaanálisis por subgrupos reveló diferencias de expresión estadísticamente significativas en 671 genes en G1 y en 954 genes en G2, considerando en ambos subgrupos un *p*-valor < 0,05 (**Anexo 10**). Mediante el cruce de las dos listas de genes se identificaron 521 genes comúnmente desregulados en ambos subgrupos, de los que 65 se encontraban sobreexpresados con el tratamiento, 428 infraexpresados con el tratamiento y 28 con sentido de expresión opuesto en ambos subgrupos. En la **Figura 4.36** se muestran dos ejemplos de diagrama de bosque, considerando los dos subgrupos de tiempo de tratamiento, para los dos genes que presentaron un mayor valor absoluto de la mediana del FC.

**a**




**b**

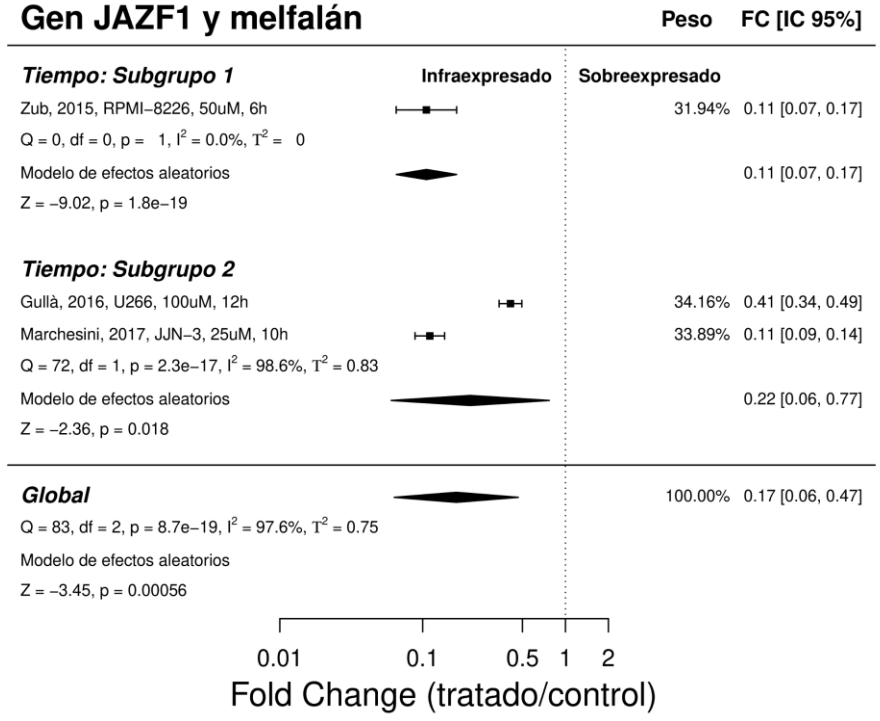



***Figura 4.36.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos para el tiempo de tratamiento con melfalán.* ***a)*** *Diagrama de bosque del gen LSMEM1, que fue el más sobreexpresado considerando la mediana del FC de los tres estudios seleccionados.* ***b)*** *Diagrama de bosque del gen JAZF1, que fue el más infraexpresado considerando la mediana del FC de los tres estudios seleccionados.*

En un siguiente paso se estudiaron las diferencias de expresión génica entre los dos subgrupos de tiempo de tratamiento. Del total de 1.460 genes analizados, 588 presentaron diferencias de FC estadísticamente significativas a *p*-valor < 0,05 entre los dos subgrupos (**Anexo 10**). En la **Figura 4.37** se muestra la comparación de los valores del ln(FC) de los genes estudiados. En rojo se muestran los genes con diferencias estadísticamente significativas entre los dos subgrupos de tiempo estudiados. La mayor parte de los genes que presentan diferencias entre los dos subgrupos muestran, sin embargo, un mismo sentido de la expresión génica, es decir, en ambos subgrupos están sobreexpresados o infraexpresados cuando las células son tratadas con melfalán. No obstante, existe una pequeña proporción de genes con diferente sentido de expresión en función del tiempo de tratamiento, que se recogen en los cuadrantes superior izquierdo e inferior derecho de la **Figura 4.37**.

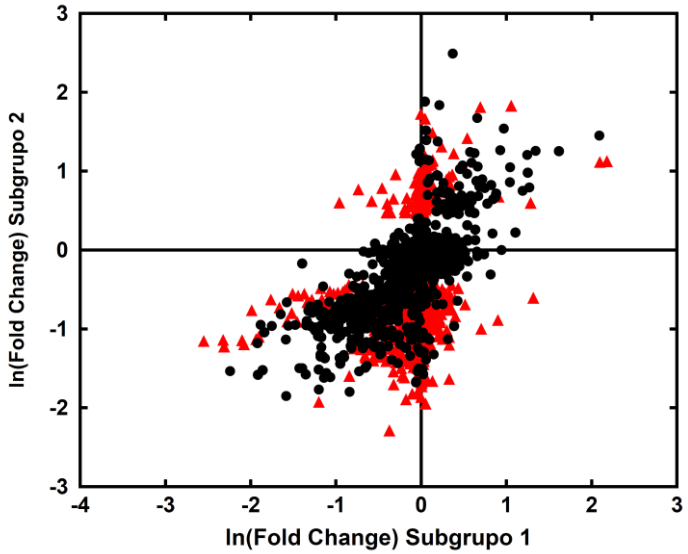



***Figura 4.37.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 1.460 genes estudiados donde se comparan los subgrupos 1 y 2. En rojo se muestran los 588 genes que mostraron diferencias estadísticamente significativas entre ambos subgrupos.*

Para determinar qué vías y funciones biológicas se ven afectadas en función del tiempo de tratamiento se procedió al análisis de sobrerrepresentación (ORA) sobre esta lista de genes diferencialmente expresados. Los resultados de este estudio, considerando como fuente de datos las bases de vías biológicas KEGG y GO, se recogen en la **Figura 4.38**.

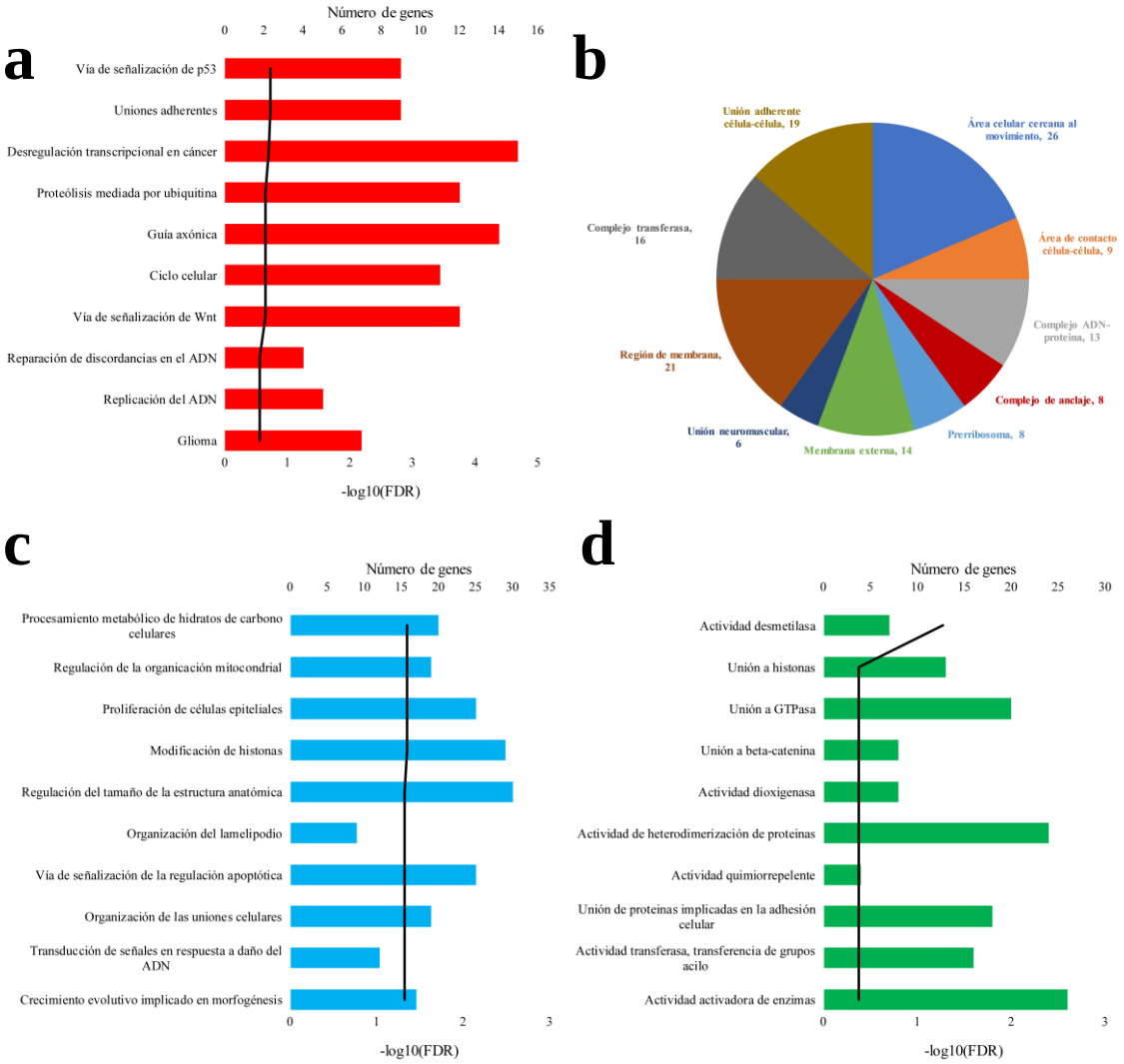



***Figura 4.38.*** *Análisis de sobrerrepresentación de los genes que presentaron diferencias de expresión estadísticamente significativas entre los dos subgrupos de tiempo de tratamiento con melfalán.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

En el análisis ORA de vías biológicas KEGG no se detectó ninguna vía sobrerrepresentada de forma estadísticamente significativa a FDR < 0,05 (**Figura 4.38a**). De la misma manera, ninguno de los términos correspondientes a los componentes celulares (CC) y las funciones moleculares (FM) de la base GO resultó significativamente sobrerrepresentado (**Figura 4.38, paneles b y d**, respectivamente). En cuanto a los procesos biológicos (PB), los 10 procesos que se muestran en la **Figura 4.38c** fueron estadísticamente significativos a FDR < 0,05. Uno de estos procesos es la "vía de señalización de la regulación apoptótica" (FDR = 0.0479). Este hallazgo se sustenta en la apoptosis celular inducida por el melfalán en modelos *in vitro*[425]. El hecho de que este proceso aparezca sobrerrepresentado en este análisis podría ser indicativo de que el tiempo de tratamiento es un factor determinante en la señalización de la apoptosis inducida por el melfalán. Los genes que están sometidos a esta regulación temporal se

recogen en la **Figura 4.39**. En la mayoría de los genes puede comprobarse que, al incrementarse el tiempo de tratamiento, aumenta también la cuantía del cambio de la expresión génica, tanto en sentido de sobreexpresión como de infraexpresión. Esto indicaría que el tiempo al que se ha llevado el estudio del subgrupo G1 ($t \leq 8$ h) podría ser aún muy precoz para producir un cambio en la expresión de estos genes, lo que podría contribuir a que no se observase el efecto deseado del melfalán sobre la apoptosis.

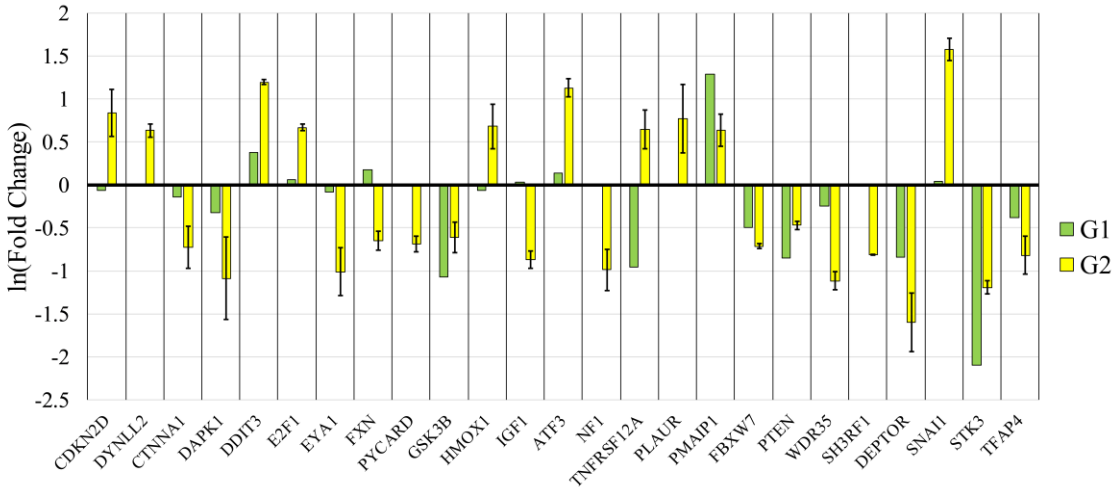



***Figura 4.39.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la función de "señalización de la regulación de la apoptosis" en los dos subgrupos de tiempo de tratamiento con melfalán (G1 y G2). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.1.2. Metaanálisis por subgrupos: concentración

Los puntos de corte para la concentración de melfalán fueron establecidos a 25 y 75 µM en función de la mediana ± MAD de los tres estudios seleccionados. Por esta razón, fue imposible el estudio de metaanálisis por subgrupos de concentración ya que cada uno de los posibles subgrupos constó de un único estudio (**Tabla 4.4)**.

#### 4.3.1.3. Metaanálisis global del melfalán

Una vez determinada la influencia de los subgrupos en la expresión génica se realizó un análisis global considerando todos los estudios seleccionados para el melfalán. Este metaanálisis reveló una diferencia en el tamaño del efecto estadísticamente significativa a *p*-valor < 0,05 para la expresión de 711 genes, de los que 128 presentaron sobreexpresión y 583 infraexpresión en las muestras tratadas con melfalán. El resultado completo de este metaanálisis está recogido en el **Anexo 11**.

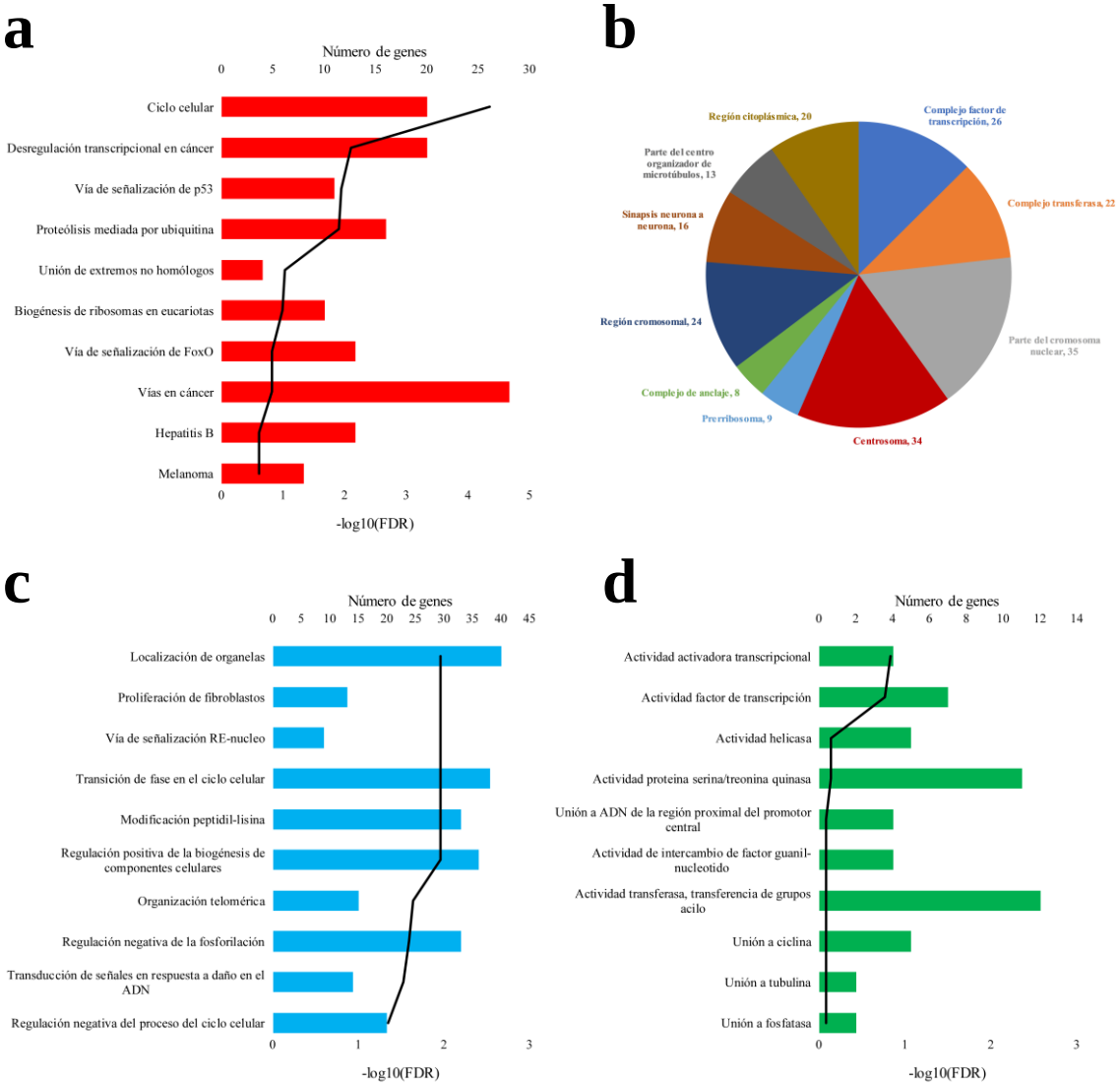



***Figura 4.40.*** *Análisis de sobrerrepresentación sobre rutas KEGG y términos GO considerando los 711 genes con un tamaño del efecto estadísticamente significativo en el metaanálisis de la expresión génica para el tratamiento con melfalán.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

En el análisis de rutas biológicas KEGG (**Figura 4.40a**), la vía más destacada en cuanto a significancia estadística fue el "ciclo celular" (FDR < 0,001), con 20 genes desregulados. Esto concuerda también con el análisis GO donde destaca la presencia de un gran número de procesos que implican al ciclo celular y funciones relacionadas con los procesos de transcripción y replicación.

Uno de los 20 genes desregulados en la vía KEGG del "ciclo celular" fue el oncogén *MYC* (*z*-valor = -6,99, *p*-valor < 0,0001). La infraexpresión de *MYC* producida por el melfalán podría conducir a la reducción de la actividad del complejo Cdk6-Ciclina D[426], lo que produciría la detención del ciclo celular[427]. La supresión de la actividad de este complejo estaría mediada a través de la activación del gen supresor de tumores *CDKN2B*,

sobreexpresado tras el tratamiento con melfalán (*z*-valor = 2,39, *p*-valor = 0,0169), ya que *MYC* ejerce un efecto represor sobre la transcripción de este gen[428].

La reducción de la actividad del complejo Cdk6-Ciclina D también podría ser llevada a cabo por el melfalán a través de la represión directa de la transcripción del gen *CDK6*, ya que en este trabajo se ha observado una marcada infraexpresión del propio gen de esta quinasa (*z*-valor = -2,96, *p*-valor = 0,0031). El resultado de estos mecanismos de regulación de *CDK6* conducirían al control de la progresión del ciclo celular mediante el arresto en la fase G1[429] (**Figura 4.41**).

Otros genes que fueron desregulados en la "vía del ciclo celular" fueron *ATM*, *CHEK2* y *CCNB1*, los tres infraexpresados, o *GADD45B*, sobreexpresado, todos ellos implicados también en la "vía de señalización de p53" (FDR = 0,012) (**Figura 4.42**). La infraexpresión de *ATM* (*z*-valor = -2,25, *p*-valor = 0,0244) y *CHEK2* (*z*-valor = -3,44, *p*-valor = 0,0006) resulta de gran interés, ya que, podría suponer el desarrollo por parte de la célula de un mecanismo antibloqueo del ciclo celular. Por un lado, al estar infraexpresado *ATM*, la proteína codificada por *CHEK2* no podría ser activada por fosforilación en respuesta a roturas de doble cadena[430], y, por otro lado, la propia producción de la proteína codificada por *CHEK2* (Chk2) también podría verse reducida al no estar promovida la transcripción del gen. En ambos casos, este proceso conduciría a la ausencia de estabilización de la proteína p53 con lo que se favorecería la unión con la proteína Mdm2, y por tanto la degradación de p53 por la vía del proteasoma[431]. Sin embargo, esta hipótesis no concuerda con los aumentos de expresión observados en los genes *GADD45B* (*z*-valor = 2,42, *p*-valor = 0,0154) y *CDKN1A*, ya que p53 es un regulador transcripcional positivo de ambos genes[432, 433]. Esto, unido a que previamente se ha reportado que el tratamiento con melfalán produce un gran incremento de la proteína fosforilada Chk2[434] podría sugerir que la regulación negativa tanto de *ATM* como de *CHEK2* se trate de un mecanismo compensatorio al tratamiento con este fármaco, cuyo fin sería estabilizar los niveles de la proteína Chk2 para corregir la parada del ciclo celular.

No obstante, de los genes que aparecen desregulados en la vía anterior puede deducirse una vía alternativa de detención del ciclo celular llevada a cabo a través de la sobreexpresión del gen *GADD45B*, ya que la proteína codificada por este gen inhibe la actividad quinasa del complejo Cdk1-Ciclina Bl[435], responsable del punto de control G2/M[436]. Esto, unido a la infraexpresión de la propia ciclina *CCNB1* (*z*-valor = -5,37, *p*-valor < 0,0001) podría conducir a la célula al arresto en la fase G2/M del ciclo celular.

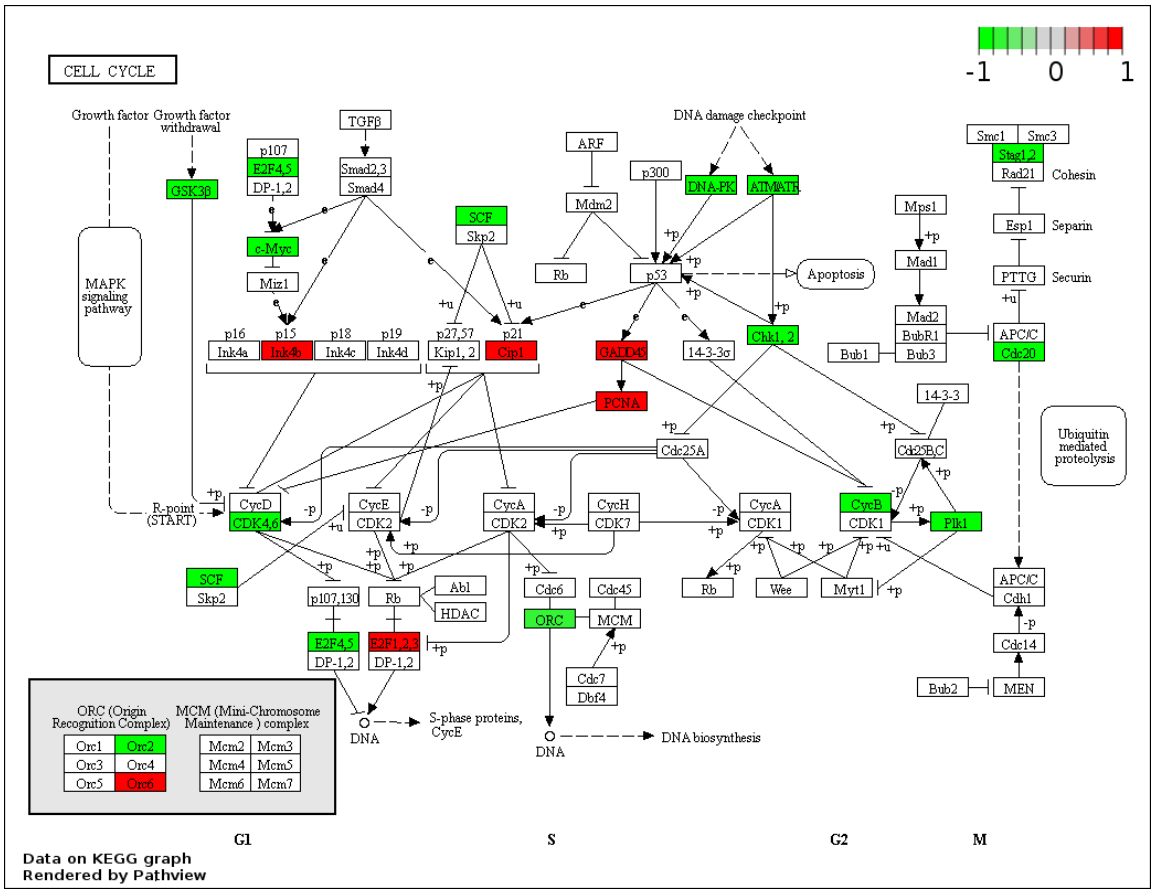



***Figura 4.41.*** *Vía del ciclo celular según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del melfalán.*

Dentro de la "vía de señalización de p53" también se observó la infraexpresión del gen supresor de tumores *PTEN* después de la exposición al melfalán (*z*-valor = -4,12, *p*-valor < 0,0001). La deficiencia de este gen supone *a priori* un fenotipo favorable a la carcinogénesis[437], pero al mismo tiempo el melfalán también provocó una infraexpresión de la subunidad beta de PI3K (*PIK3CB*, *z*-valor = -4,64, *p*-valor < 0,0001), integrada en las vías PI3K-AKT y mTOR. Se ha demostrado que la ausencia simultánea de *PTEN* y *PIK3CB* en cáncer produce una fuerte inhibición del crecimiento celular y de la vía de señalización de PI3K-AKT[438], lo que sugiere que el melfalán podría producir el bloqueo del crecimiento celular a través de la inhibición de estos dos genes.

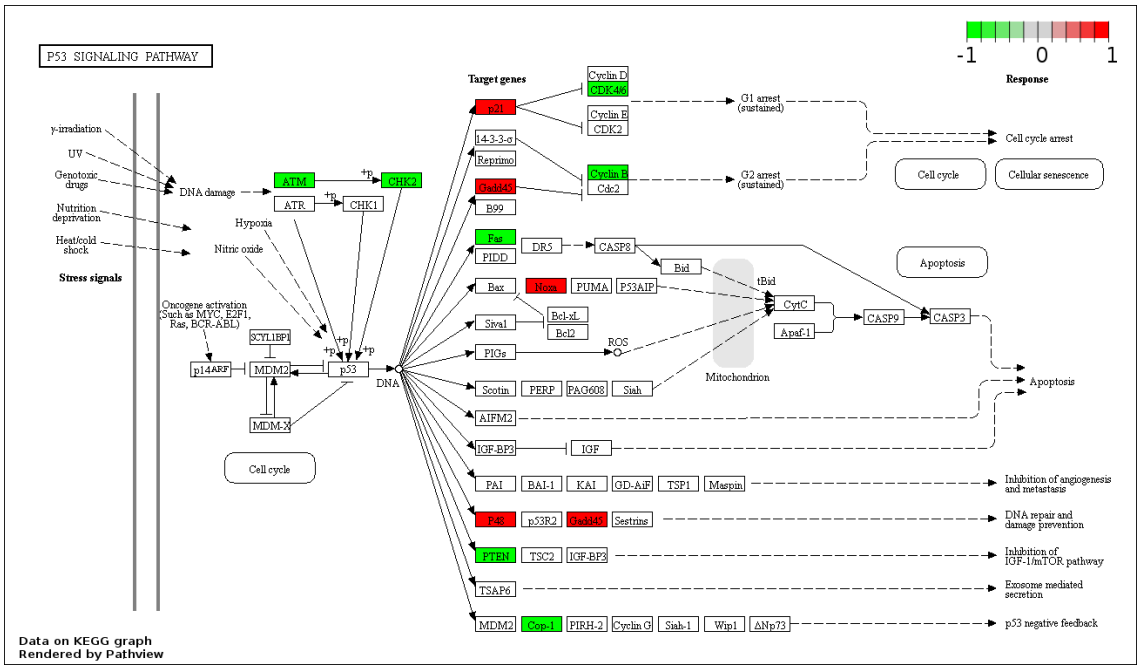



***Figura 4.42.*** *Vía de señalización de p53 según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del melfalán.*

### 4.3.2. Dexametasona

En el proceso de búsqueda sistemática en los repositorios *online* de HMCLs tratadas con dexametasona se detectaron 546 series en GEO, 137 series en ArrayExpress y 38 muestras correspondientes a 10 series en SRA. El total de series localizadas ascendió a 693, y finalmente 585 se eligieron para su evaluación detallada tras la eliminación de los elementos duplicados en los tres repositorios. De las 585 series revisadas, cinco cumplieron los criterios de inclusión y exclusión necesarios para ser incluidas en el metaanálisis. Se añadió además una sexta serie cedida por el Servicio de Hematología de Salamanca. A continuación, se comprobó la posible subdivisión de las seis series en diferentes estudios en función de las concentraciones de fármaco empleadas, el tiempo de tratamiento o la utilización de varias líneas celulares. Solamente la serie GSE30644 tuvo que ser dividida en dos estudios debido a que sobre cuatro de sus muestras se empleó una concentración de 0,025 µM, mientras que en las cuatro muestras restantes, la concentración empleada fue de 2 µM. Por tanto, el número final de estudios considerados para el metaanálisis fue de 7. El diagrama de flujo que se muestra en la **Figura 4.43**, detalla el esquema de selección de estudios para este metaanálisis en función de los diferentes criterios de inclusión y exclusión.

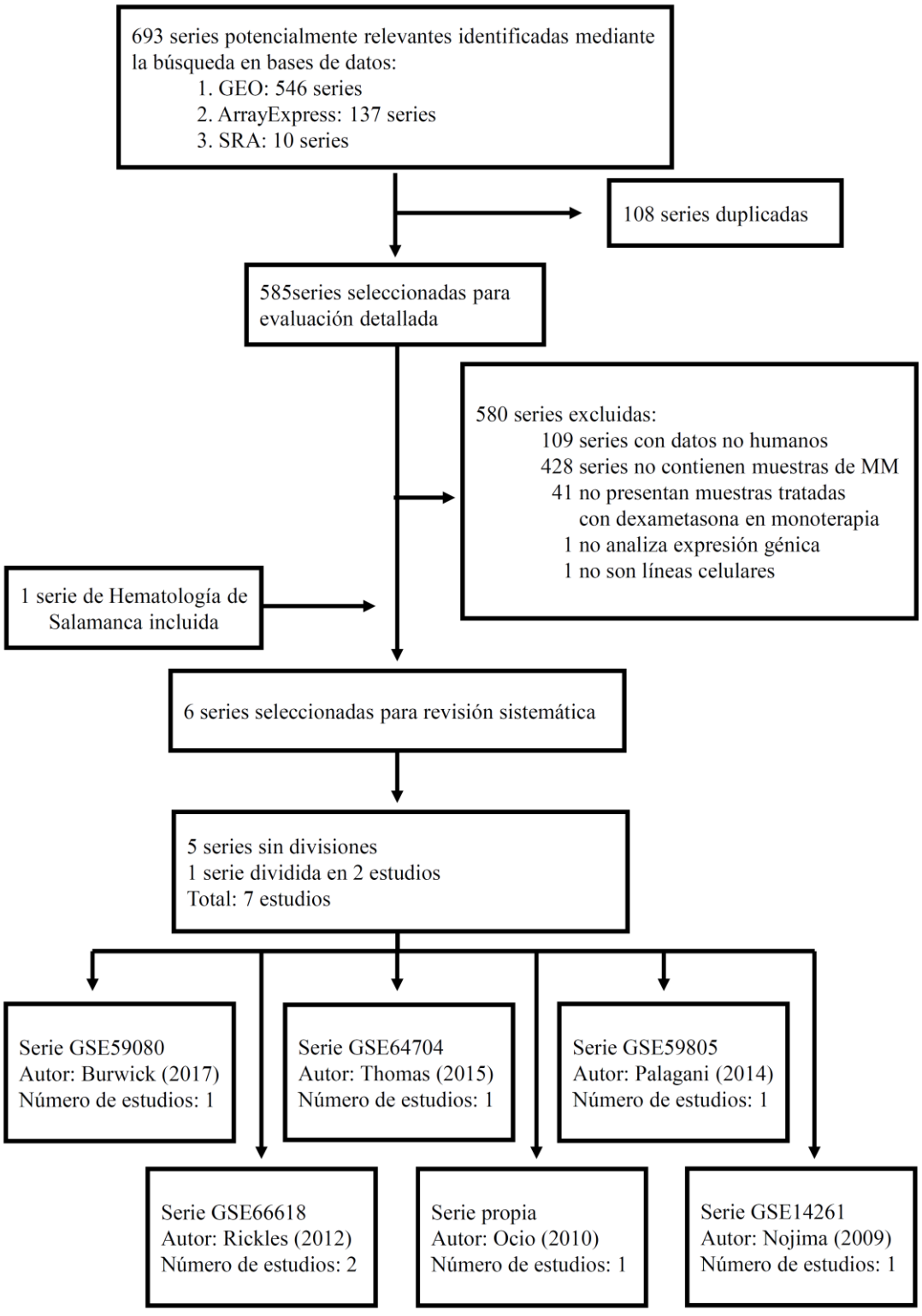



***Figura 4.43.*** *Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares tratadas con dexametasona en monoterapia.*

Los 7 estudios seleccionados se clasificaron en tres subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de dexametasona. Los puntos de corte del tiempo de tratamiento se establecieron a tres y a 9 horas (mediana de seis horas). En el caso de la concentración, debido a que la mayoría de los estudios se realizaron a una concentración de 1 µM, el valor de la mediana fue de 1 µM, con un MAD de 0, con lo que este fue el único punto de corte. Los resultados del agrupamiento en subgrupos están recogidos en la **Tabla 4.5**.

***Tabla 4.5**. Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con dexametasona en monoterapia.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (µM) |
|---|---|---|---|---|---|---|
| GSE59080 | Burwick (2017)[439] | MM1-S | Illumina HumanHT-12 V4.0 expression beadchip | 4 | 4 | 1 |
| GSE64704 | Thomas (2015)[440] | MM1-S | Illumina HumanHT-12 V4.0 expression beadchip | 8 | 3 | 1 |
| GSE59805 | Palagani (2014)[441] | MM1-S | Illumina HumanHT-12 V4.0 expression beadchip | 4 | 72 | 1 |
| GSE30644 | Rickles (2102a)[442] | MM1-S | Affymetrix Human Genome U133 Plus 2.0 | 4 | 6 | 0.025 |
| GSE30644 | Rickles (2012b)[442] | MM1-S | Affymetrix Human Genome U133 Plus 2.0 | 4 | 6 | 2 |
| Salamanca | Ocio (2010) | MM1-S | Affymetrix Human Gene 1.0st Array | 4 | 24 | 0.01 |
| GSE14261 | Nojima (2009)[443] | OPM1 | Agilent Whole Human Genome Microarray 4x44K | 2 | 24 | 1 |

*En verde, estudios seleccionados para el subgrupo de tiempos o concentraciones bajos; en amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedios; en rojo, estudios seleccionados para el subgrupo de tiempos o concentraciones altos.*

Tras la determinación de los subgrupos se procedió a la selección de genes candidatos para el metaanálisis, considerando los genes cuyo valor absoluto del FC fue mayor a 1,5 en los 7 estudios propuestos, o en todos los estudios de, al menos, uno de los subgrupos de tiempo o concentración, excluyendo los subgrupos que solamente constasen de un estudio. Así, se seleccionaron un total de 141 genes para los que la distribución de su FC se muestra en la **Figura 4.44**. Esta distribución se mostró muy similar en los 7 estudios. El estudio de Rickles (2012b) mostró una mayor dispersión de los valores del ln(FC) lo que coincidió con la utilización de concentraciones de dexametasona más elevadas. Estas posibles diferencias en cuanto a la concentración empleada, pero también en cuanto al tiempo de tratamiento, serán evaluadas a continuación mediante metaanálisis por subgrupos.

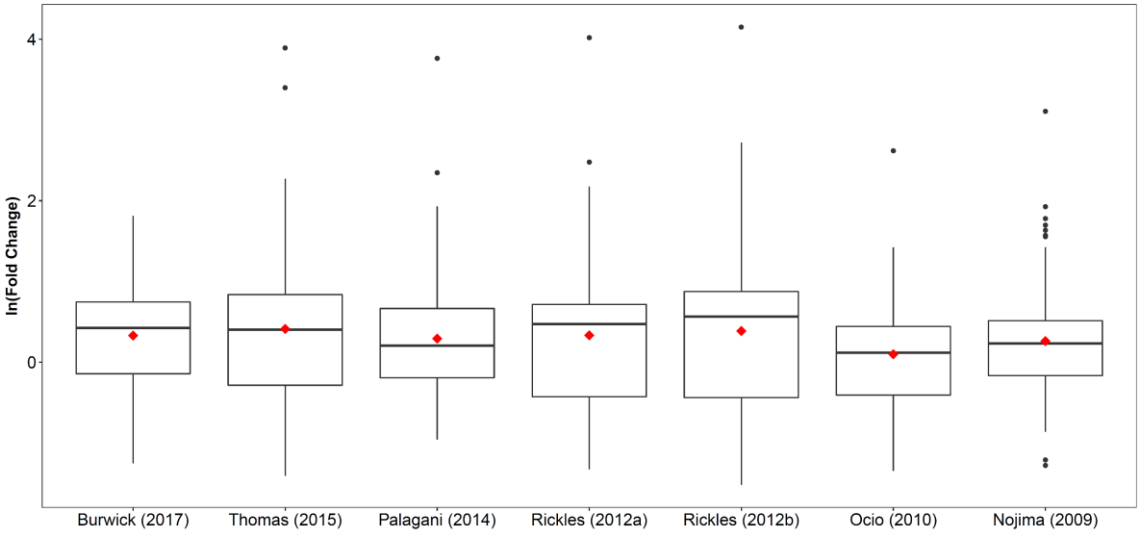
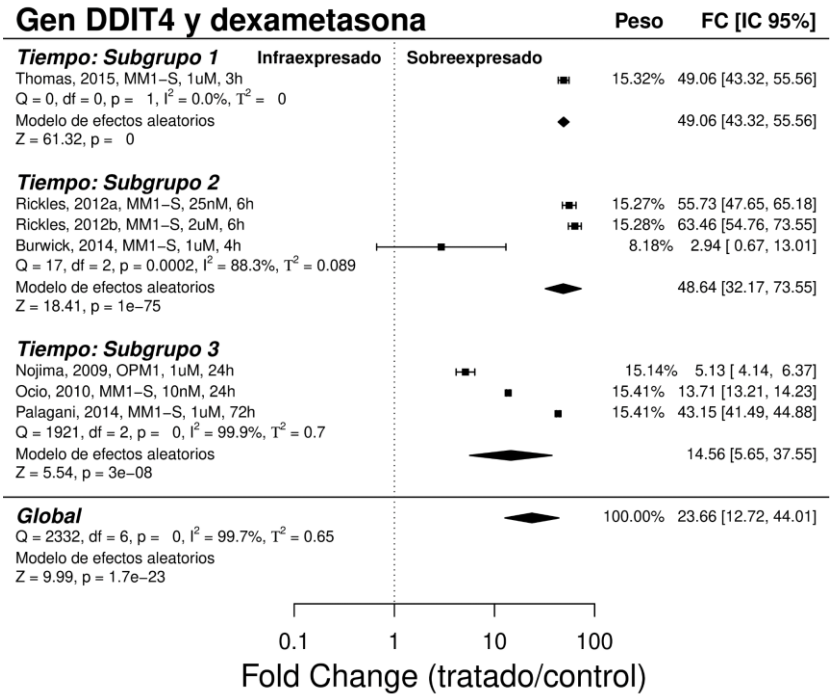



***Figura 4.44.*** *Diagrama de caja (box plot) del ln(Fold Change) de los 141 genes seleccionados para el metaanálisis de dexametasona en monoterapia en líneas celulares de MM. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.2.1. Metaanálisis por subgrupos: tiempo de tratamiento

Para este metaanálisis se determinaron tres subgrupos en función de los puntos de corte establecidos mediante la mediana del tiempo de tratamiento y su MAD a las tres y a las 9 horas. De este modo, el primer subgrupo (G1) solamente recogió el estudio de Thomas (2015) que cumplía el criterio de un tiempo de tratamiento inferior e igual a las tres horas. El segundo subgrupo (G2), que englobó los estudios con un tiempo de tratamiento inferior o igual a 9 horas, pero superior a las tres horas, incluyó tres estudios: Burwick (2017) y los estudios (a) y (b) de Rickles (2012). Por último, el tercer subgrupo (G3) integró los estudios a tiempo superior a 9 horas: Palagani (2014), Ocio (2010) y Nojima (2009). Mediante el metaanálisis se identificaron 115 genes con *p*-valor < 0,05 en el G1, 131 genes en el G2 y 99 genes en el G3 (**Anexo 12**). De estos genes, 76 fueron comunes a los tres subgrupos, de los que 54 presentaron sobreexpresión al tratar con dexametasona, y 22 infraexpresión. En la **Figura 4.45** se muestran dos ejemplos de diagramas de bosque por subgrupos de concentración considerando los dos genes con mayor valor absoluto de la mediana de FC.

**a**




**b**

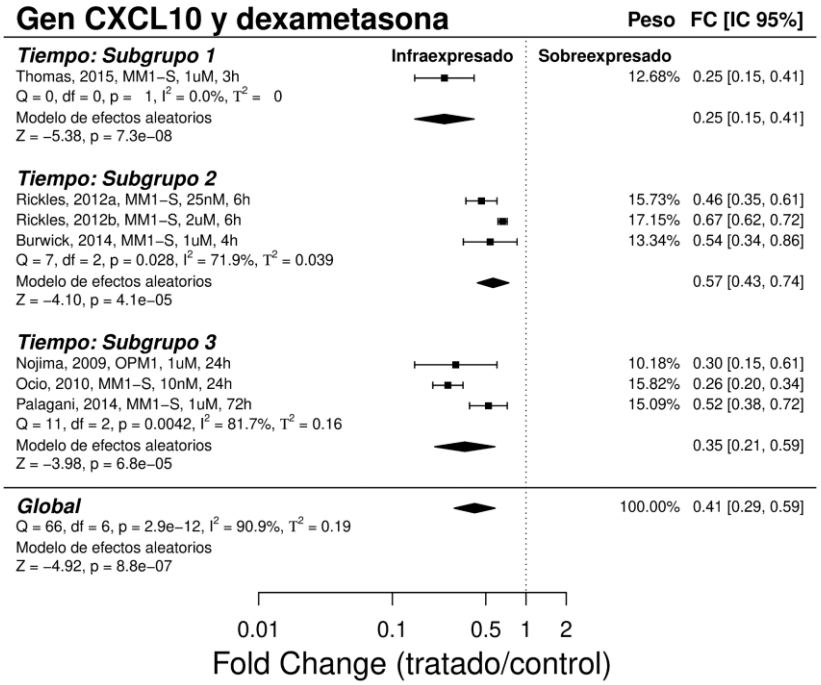



***Figura 4.45.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con dexametasona.* ***a)*** *Diagrama de bosque del gen DDIT4, que fue el más sobreexpresado considerando la mediana del FC de los 7 estudios seleccionados 7 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen CXCL10, que fue el más infraexpresado considerando la mediana del FC de los 7 estudios seleccionados.*

La evaluación de las diferencias entre los tres subgrupos de tiempo de tratamiento mostró que, de los 141 genes analizados, 41 genes presentaron diferencias estadísticamente significativas entre G1 y G2; 67 genes entre G1 y G3; y 69 genes entre los G2 y G3. En la **Figura 4.46** se muestran las comparaciones entre los valores de ln(FC) obtenidos en los tres subgrupos.

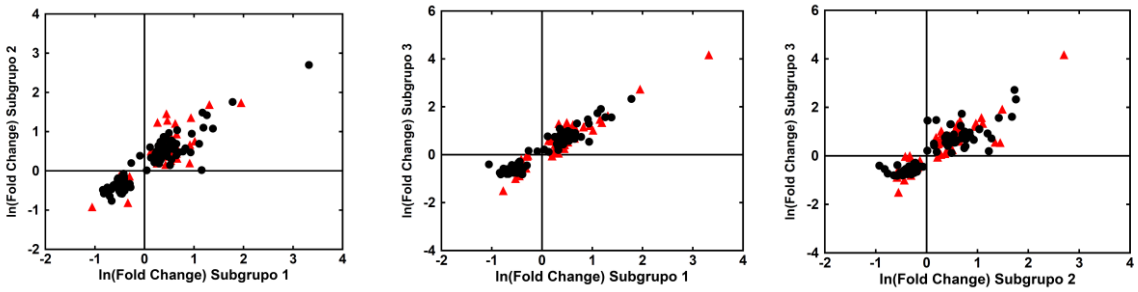



***Figura 4.46.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 141 genes estudiados donde se comparan los subgrupos 1, 2 y 3. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

En las tres comparaciones no se observaron grandes discrepancias en cuanto al sentido de la expresión de estos genes, por lo que se evaluará la naturaleza de las diferencias de expresión génica sobre las rutas KEGG y los procesos, funciones y componentes GO mediante un análisis ORA (**Figura 4.47**).

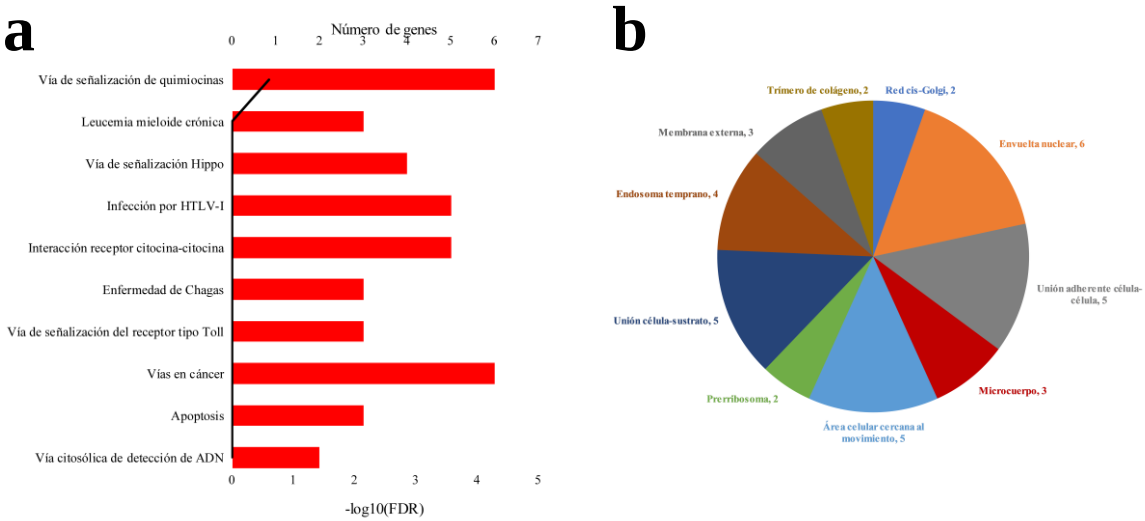



***Figura 4.47.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con dexametasona.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

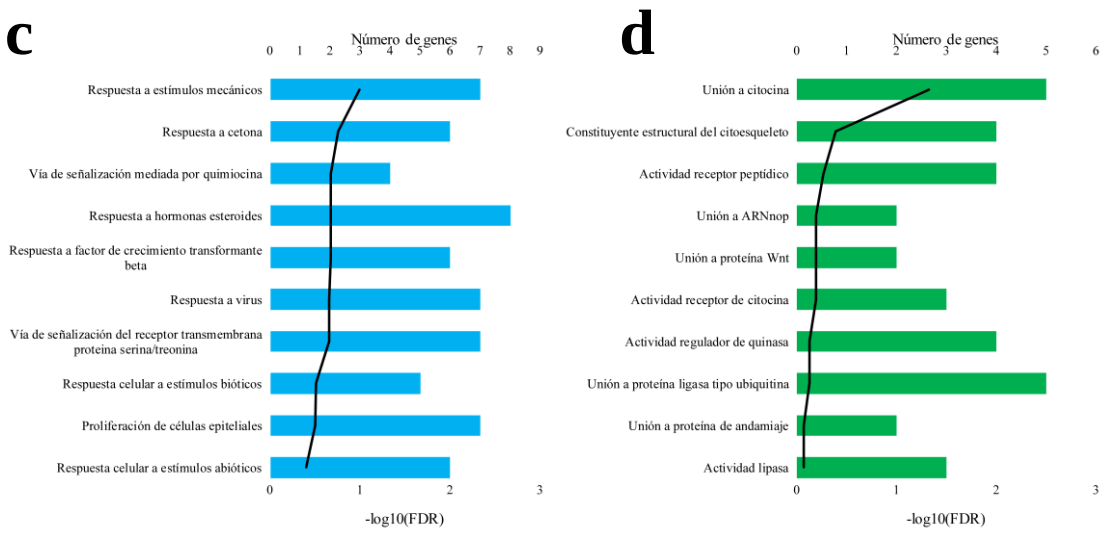



***Figura 4.47 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con dexametasona.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Solamente la FM de "unión a citocinas" logró una sobrerrepresentación estadísticamente significativa (**Figura 4.47d**), de manera que el efecto del tiempo sobre la expresión génica se estudió de forma específica sobre los genes esta función (**Figura 4.48**), observándose que, a medida que se incrementó el tiempo de tratamiento, el efecto de la dexametasona sobre la expresión se atenuaba en la mayor parte de los genes. Este efecto fue acusado de forma particular en genes que codifican receptores de citocinas como *CCR1*, *TGFBR2* y *CXCR4* y en el gen *FZD4* que codifica una proteína transmembrana receptor de proteínas Wnt. Esto sugiere que el efecto máximo de la dexametasona sobre la transcripción de los genes de estos receptores se produciría entre las tres primeras horas de tratamiento del G1 y las 3-9 horas del G2, con una reducción del efecto del fármaco sobre la transcripción a más de 9 horas (G3). Pese a que existen diferencias estadísticamente significativas entre G1 y G2 en todos los genes mencionados, salvo *CXCR4* (**Anexo 12**), no es conveniente atribuir las diferencias entre ambos subgrupos al tiempo de tratamiento ya que G1 solamente consta de un estudio.

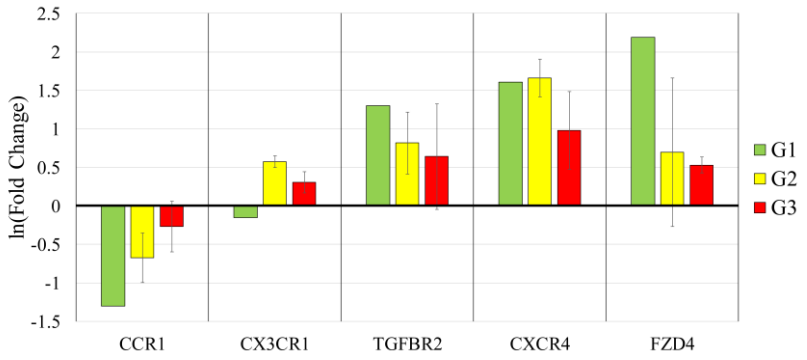



***Figura 4.48.*** *Valores promedio del ln(Fold Change) de los genes desregulados en función molecular "unión a citocina" en los tres subgrupos de tiempo de tratamiento con dexametasona (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.2.2. Metaanálisis por subgrupos: concentración

En cuanto al metaanálisis por subgrupos de concentración se determinaron tres subgrupos considerando como punto de corte la concentración de dexametasona aplicada a 1 µM. El primer subgrupo (G1) comprendió los dos estudios realizados a concentraciones inferiores a 1 µM: Rickles (2012a) y Ocio (2010); en el segundo subgrupo (G2) se agruparon los cuatro estudios a concentraciones iguales a 1 µM: Burwick (2017), Thomas (2015), Palagani (2014) y Nojima (2009); por último, el tercer subgrupo (G3) solamente constó de un estudio en el que la concentración fue superior a 1 µM: Rickles (2012b). Mediante el metaanálisis por subgrupos se identificaron 94 genes con *p*-valor < 0,05 en el subgrupo G1, 111 genes en el caso del subgrupo G2 y 133 genes en el subgrupo G3 (**Anexo 13**). Setenta y ocho de estos genes fueron comunes a los tres subgrupos, de los que 48 presentaron sobreexpresión y 30 infraexpresión al tratamiento con dexametasona. En la **Figura 4.49** se muestra un ejemplo de diagrama de bosque para el metaanálisis por subgrupos de concentración de los dos genes que, siendo significativos en el metaanálisis global, presentaron un mayor valor absoluto de la mediana del FC.

**a**

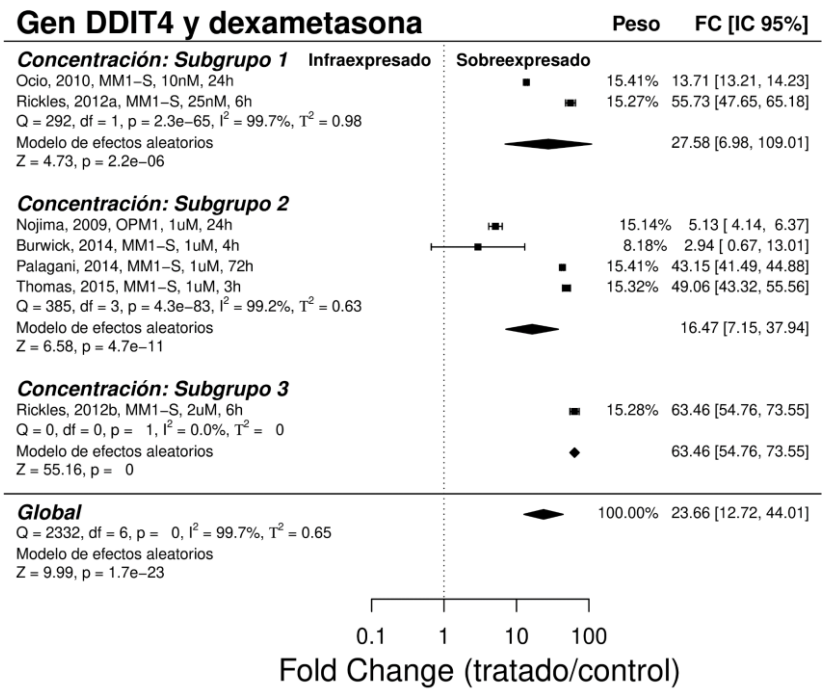



***Figura 4.49.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de dexametasona.* ***a)*** *Diagrama de bosque del gen DDIT4, que fue el más sobreexpresado considerando la mediana del FC de los 7 estudios seleccionados.*

**b**

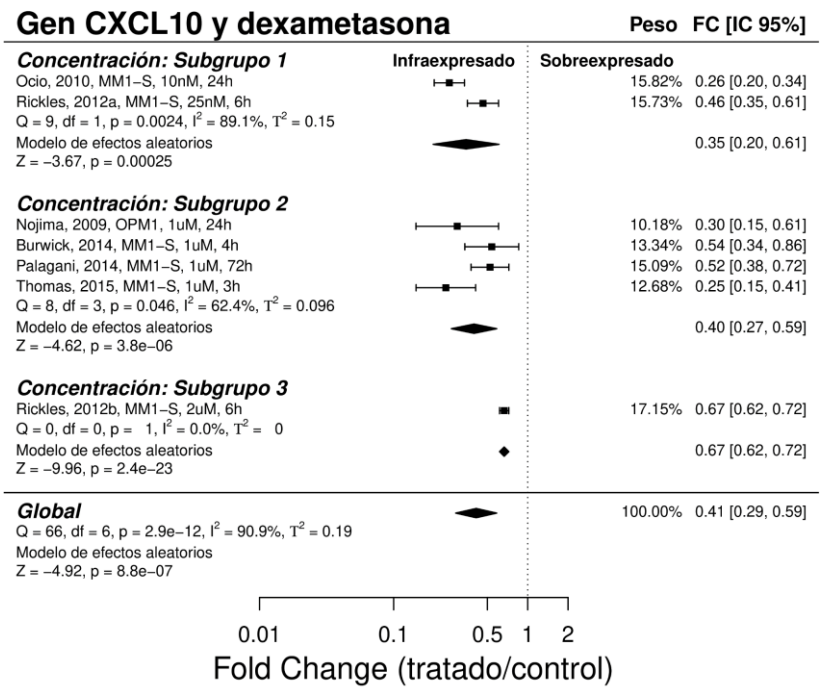



***Figura 4.49 (continuación).*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de dexametasona.* ***b)*** *Diagrama de bosque del gen CXCL10, que fue el más infraexpresado considerando la mediana del FC de los 7 estudios seleccionados.*

A continuación, se compararon los resultados obtenidos en cada uno de los subgrupos para determinar los genes que presentaron diferencias estadísticamente significativas en función de la concentración de fármaco empleada: 19 genes presentaron diferencias estadísticamente significativas entre los G1 y G2, 15 genes entre G1 y G3, y 44 genes entre G2 y G3 (**Anexo 13**). En la **Figura 4.50** se muestra la comparación entre los valores de ln(FC) de los tres subgrupos de concentración de dexametasona. Puede observarse que no aparecen genes estadísticamente significativos entre ninguno de los subgrupos con sentidos opuestos de la expresión génica (cuadrantes superior izquierdo e inferior derecho).

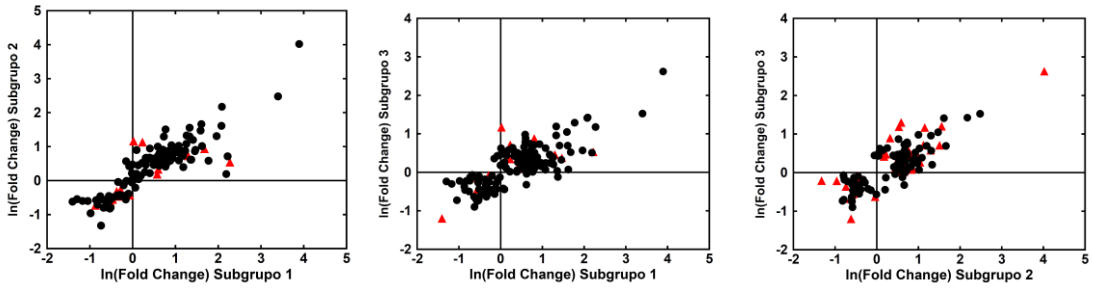



***Figura 4.50.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 141 genes estudiados donde se comparan los subgrupos 1, 2 y 3 de concentración de dexametasona. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

En un último paso se procedió al análisis de vías KEGG y términos GO de los genes con diferencias estadísticamente significativas entre los tres grupos. Este análisis se muestra en la **Figura 4.51**.

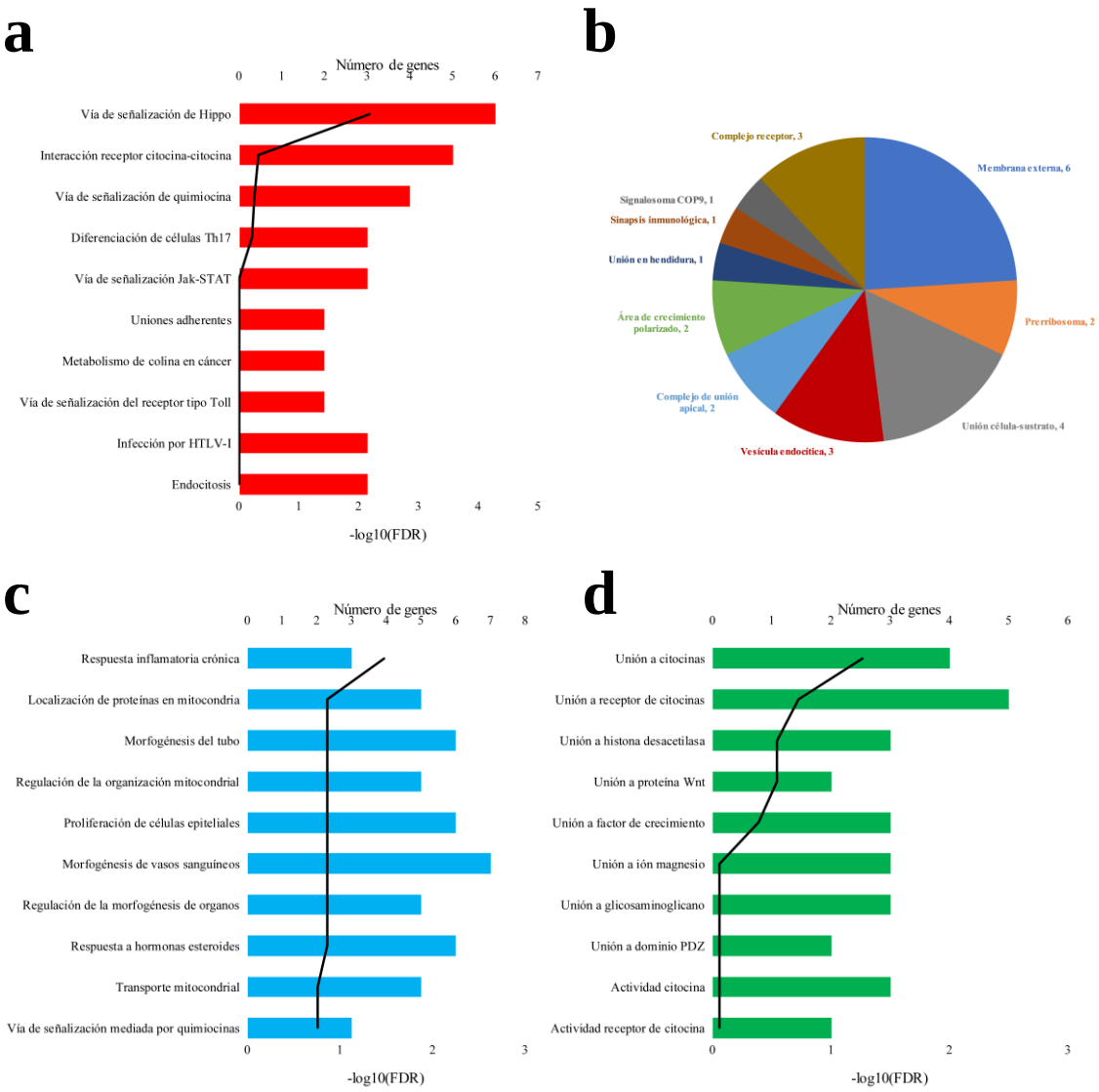



***Figura 4.51.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de dexametasona.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Aunque el número de genes con diferencias estadísticamente significativas entre los subgrupos de concentración fue notablemente menor que en el caso del tiempo de tratamiento, aparece desregulada de nuevo la FM de "unión a citocinas". En este caso se observa una gran variabilidad del efecto de la concentración de fármaco sobre la expresión génica, ya que, mientras para los genes *TGFBR2* y *FZD4* es la concentración aplicada en el G2 la que produce un mayor cambio en la expresión, esto no ocurre con los genes *CX3CR1* y *IL6ST* (**Figura 4.52**), para los que las concentraciones de

dexametasona aplicadas en los subgrupos G1 y G3 son las que inducen unos mayores cambios de expresión.

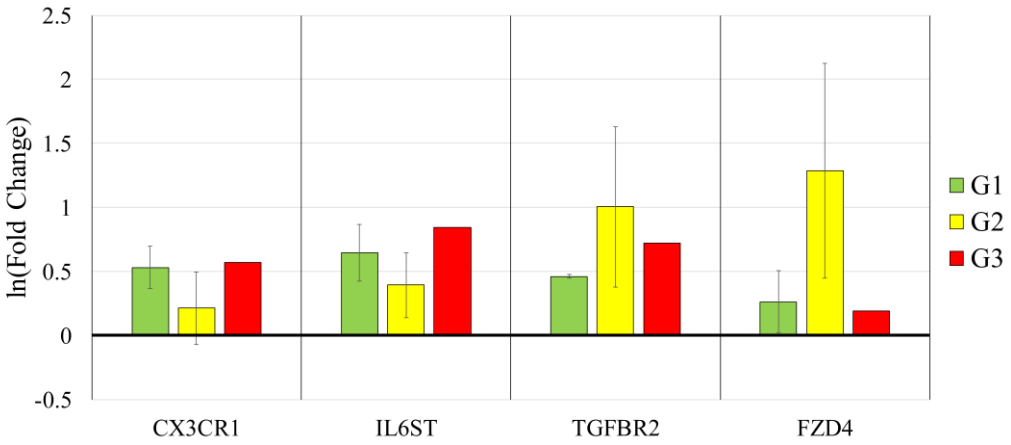



***Figura 4.52.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la función "unión a citocinas" en los tres subgrupos concentración de dexametasona (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

### 4.3.2.3. Metaanálisis global de la dexametasona

El análisis global considerando los 7 estudios reveló una diferencia en el tamaño del efecto estadísticamente significativa a *p*-valor < 0,05 en 132 genes, de los que 91 presentaron sobreexpresión y 41 infraexpresión en las muestras tratadas con dexametasona (**Anexo 14**). En la **Figura 4.53** se recoge el estudio de sobrerrepresentación realizado sobre estos genes en el que se muestran las rutas de señalización KEGG y funciones GO con mayor significancia estadística.

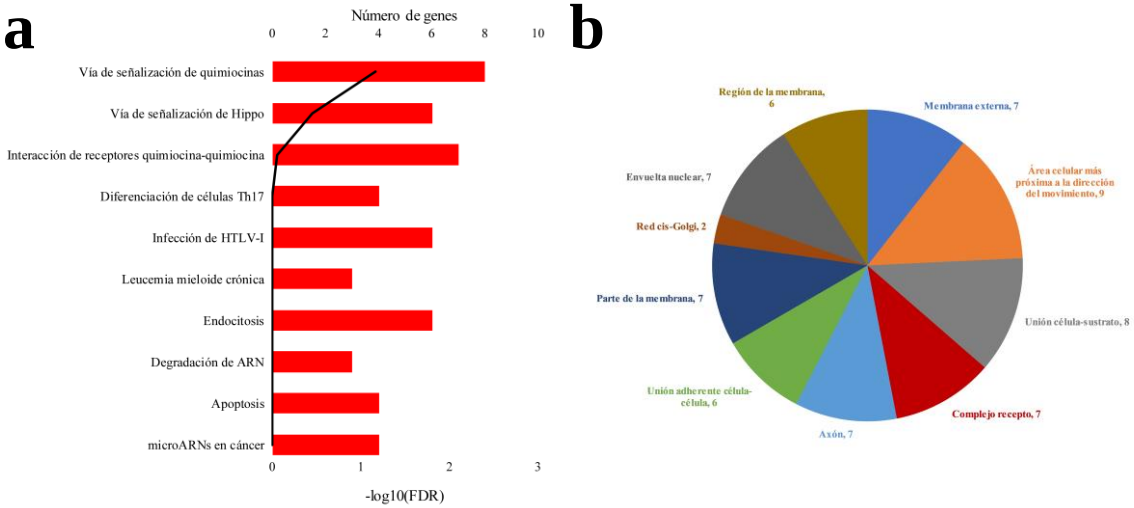



***Figura 4.53.*** *Análisis de sobrerrepresentación de rutas KEGG y funciones GO con los genes con un efecto combinado de la expresión génica estadísticamente significativo en los 7 estudios seleccionados para el metaanálisis de dexametasona.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

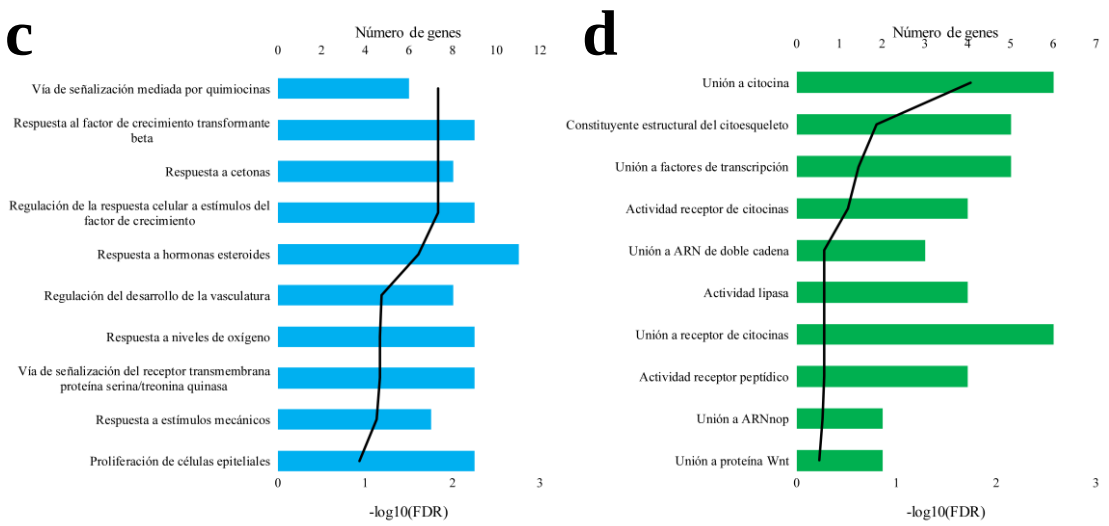



***Figura 4.53 (continuación).*** *Análisis de sobrerrepresentación de rutas KEGG y funciones GO con los genes con un efecto combinado de la expresión génica estadísticamente significativo en los 7 estudios seleccionados para el metaanálisis de dexametasona.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Entre los términos GO estadísticamente significativos del análisis ORA, encontramos la FM "unión a citocina "(FDR = 0,018) y los PB "vía de señalización mediada por quimiocinas" (FDR = 0,0147) y "respuesta al factor de crecimiento transformante beta" (FDR = 0,0147), todos ellos relacionados con procesos asociados a citocinas. Entre los genes desregulados en estas funciones se encuentran varios receptores de citocinas, sobreexpresados con la dexametasona, como *CXCR4*, *CX3CR1*, *TGFR2* y *IL6ST*. La sobreexpresión de algunos de estos receptores como *CXCR4* (*z*-valor = 6,59, *p*-valor < 0,0001)[444] o *CX3CR1* (*z*-valor = 3,80, *p*-valor = 0,0001)[445] ya ha sido reportada tras el tratamiento con este compuesto en células de MM. En el caso de *CX3CR1*, el incremento de su expresión se ha asociado a la activación de mecanismos de resistencia a dexametasona[445]. Además, la bajada de la expresión de la citocina *CXCL10* (*z*-valor = -4,92, *p*-valor < 0,0001), que ha sido descrita como una citocina antimieloma, ya que atenúa el desarrollo tumoral de las células de MM, podría apoyar esta hipótesis de aparición de mecanismos de resistencia tras la aplicación de la dexametasona[446].

Otro gen en el que se produce una disminución de los niveles de ARNm es el receptor de citocinas *CCR1 (z*-valor = -4,57, *p*-valor < 0,0001). *CCR1* tiene un papel relevante en la destrucción ósea y en la progresión del MM[447], por lo que su infraexpresión se considera de potencial interés terapéutico[448].

### 4.3.3. Bortezomib

La revisión sistemática de los estudios de expresión génica en HMCLs tratadas con bortezomib en monoterapia, condujo a la identificación de 75 series en GEO, 52 series en ArrayExpress y 21 muestras correspondientes a tres estudios en SRA. De las 130 series detectadas inicialmente, 82 fueron seleccionadas para su revisión tras ser eliminados los elementos duplicados en los tres repositorios. Solamente cinco de estas series cumplieron los criterios de inclusión y exclusión necesarios para ser incluidas en el metaanálisis. Las series de GSE63520 y GSE48785 fueron divididas en dos y cinco estudios, respectivamente, al presentar muestras con varios tiempos de tratamiento. La serie GSE41929 por su parte fue dividida en cuatro estudios al contener varios tiempos de tratamiento y dos líneas celulares diferentes. El número final de estudios considerados para el metaanálisis fue de 13. El diagrama de flujo que se muestra en la **Figura 4.54**, detalla el esquema de selección de estudios para el metaanálisis del bortezomib en función de los diferentes criterios de inclusión y exclusión.

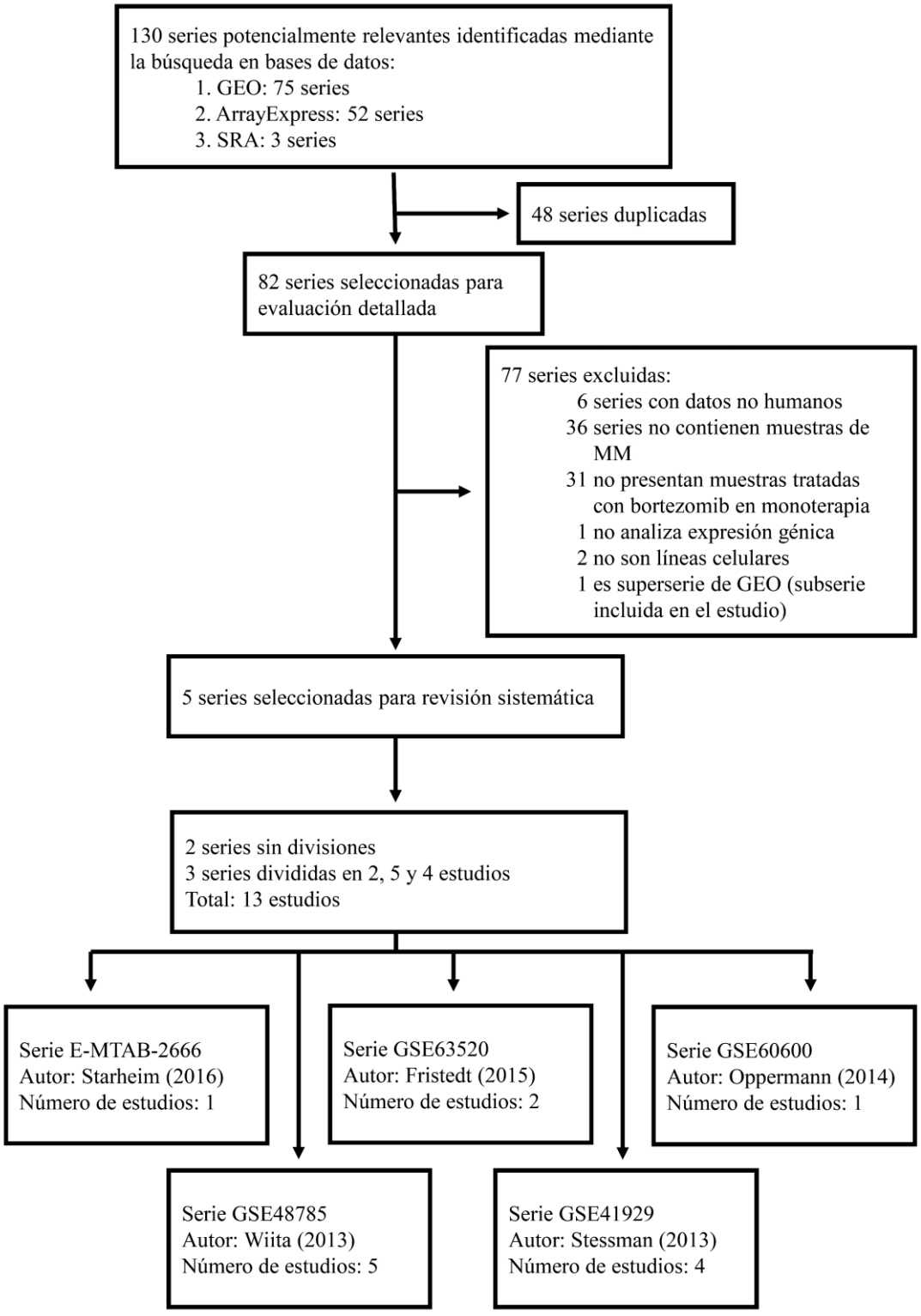



***Figura 4.54.*** *Diagrama de flujo de la selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con bortezomib en monoterapia.*

Los 13 estudios seleccionados fueron a continuación clasificados en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración

aplicada de bortezomib. Los puntos de corte para el tiempo de tratamiento fueron establecidos a las tres y a las15 horas (mediana de 9 horas), mientras que los puntos de corte para la concentración fueron establecidos a 10 y a 30 nM (mediana de 20 nM). Los resultados del agrupamiento en subgrupos pueden observarse en la **Tabla4.6.**

***Tabla 4.6****. Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con bortezomib.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (nM) |
|---|---|---|---|---|---|---|
| E-MTAB-2666 | Starheim (2016)[449] | INA-6 | Illumina HumanHT-12 V4 0_R2_15002873_B | 6 | 4 | 4 |
| GSE63520 | Fristedt (2015a)[450] | LP-1 | Agilent-SurePrint G3 Human GE v2 8x60K | 4 | 6 | 10 |
| GSE63520 | Fristedt (2015b)[450] | LP-1 | Agilent-SurePrint G3 Human GE v2 8x60K | 4 | 24 | 10 |
| GSE60600 | Oppermann (2014)* | JJN-3 | Illumina HumanHT-12 V4.0 expression beadchip | 2 | 6 | 10 |
| GSE48785 | Wiita (2013a)[451] | MM1-S | Illumina HiSeq 2000 | 2 | 1.5 | 20 |
| GSE48785 | Wiita (2013b)[451] | MM1-S | Illumina HiSeq 2000 | 2 | 3 | 20 |
| GSE48785 | Wiita (2013c)[451] | MM1-S | Illumina HiSeq 2000 | 2 | 6 | 20 |
| GSE48785 | Wiita (2013d)[451] | MM1-S | Illumina HiSeq 2000 | 2 | 9 | 20 |
| GSE48785 | Wiita (2013e)[451] | MM1-S | Illumina HiSeq 2000 | 2 | 12 | 20 |
| GSE41929 | Stessman (2103a)[452] | MM1-S | Illumina Human Whole-Genome DASL HT array | 2 | 16 | 33 |
| GSE41929 | Stessman (2013b)[452] | MM1-S | Illumina Human Whole-Genome DASL HT array | 2 | 24 | 33 |
| GSE41929 | Stessman (2013c)[452] | U266 | Illumina Human Whole-Genome DASL HT array | 2 | 16 | 33 |
| GSE41929 | Stessman (2103d)[452] | U266 | Illumina Human Whole-Genome DASL HT array | 2 | 24 | 33 |

*En verde, estudios seleccionados para el subgrupo de tiempos o concentraciones bajos; en amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedios; en rojo, estudios seleccionados para el subgrupo de tiempos o concentraciones altos. * Sin publicación asociada.*

Se consideraron 863 genes candidatos para el metaanálisis seleccionando aquellos genes cuyo valor absoluto del FC fue mayor a 1,5 en los 13 estudios en conjunto o, al menos, en todos los estudios de alguno de los subgrupos de tiempo o concentración, excluyendo los subgrupos que solamente constasen de un estudio. La distribución de los ln(FC) de estos 863 genes en cada uno de los estudios se muestra como diagrama de caja en la **Figura 4.55**. En general se observa que tanto la media como la mediana del ln(FC) de estos genes es muy cercana en todos los estudios, sin embargo, los estudios de Starheim (2016), Fristedt (2015a) y Oppermann (2014) muestran una menor dispersión de los datos. Estos tres estudios, junto con el estudio Fristedt (2015b) forman el subgrupo de menor concentración, aunque, como rasgo diferencial, el tiempo de tratamiento en el estudio de Fristedt (2015b) es muy superior al tiempo de los otros tres estudios. Para tratar

de averiguar la posible influencia tanto del tiempo de tratamiento como de la concentración sobre el efecto del bortezomib, se llevarán a cabo sendos metaanálisis por subgrupos de tiempo y concentración. Finalmente se procederá al metaanálisis global considerando todos los estudios en conjunto.

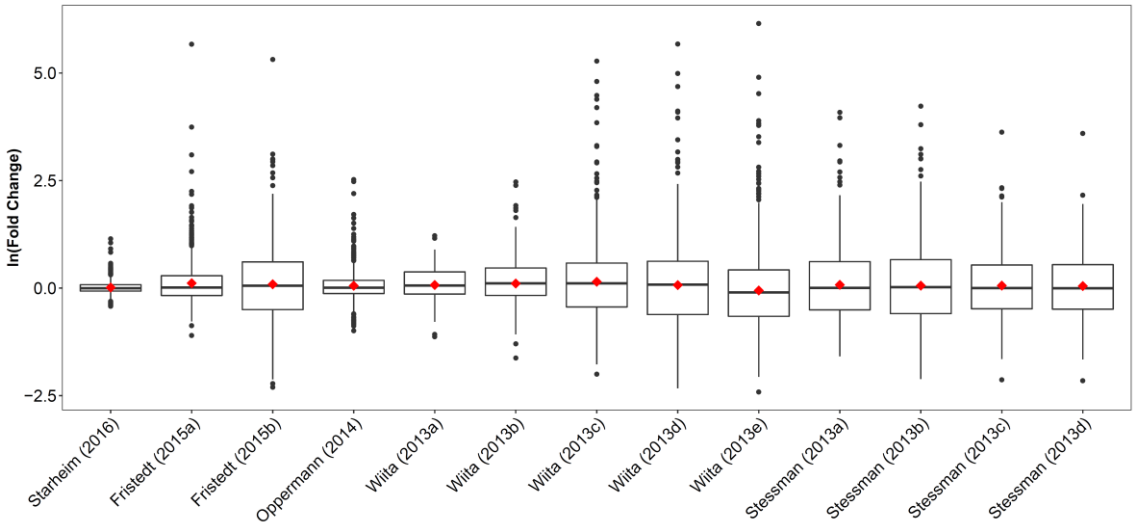
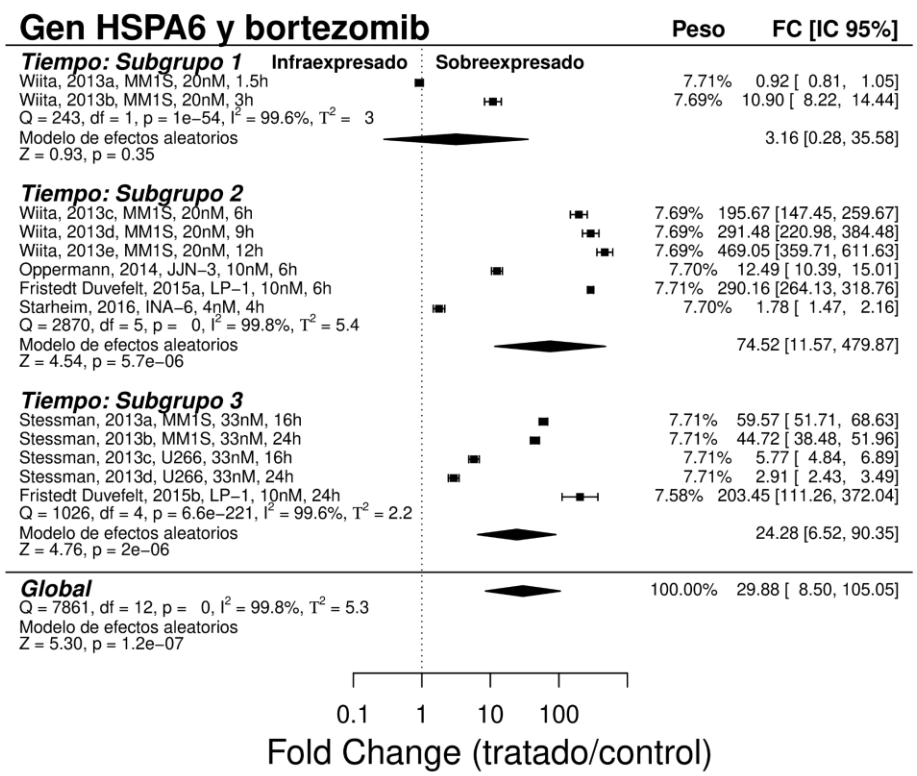



***Figura 4.55.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC]) de los 863 genes seleccionados para el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con bortezomib. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.3.1. Metaanálisis por subgrupos: tiempo de tratamiento

Se establecieron tres subgrupos para este metaanálisis en función del tiempo de tratamiento. El primer subgrupo (G1) comprendió los estudios con un tiempo de tratamiento inferior o igual a las tres horas: estudios de Wiita (2013a y 2013b). En cuanto al segundo subgrupo (G2), lo formaron los seis estudios con tiempos de tratamiento entre las tres y las 15 horas: Starheim (2016), Fristedt (2015a), Oppermann (2014), Wiita (2013c), Wiita (2013d) y Wiita (2013e). El tercer subgrupo (G3) se constituyó con los cinco estudios que tenían un tiempo de tratamiento superior a las 15h: Fristedt (2015b) y los estudios (a), (b), (c) y (d) de Stessmann (2013). Sobre estos subgrupos se realizó el metaanálisis que identificó 513 genes estadísticamente significativos a $p$-valor < 0,05 en G1, 505 genes en el caso de G2 y 416 genes en G3 (**Anexo 15**). Un total de 168 de estos genes fueron comunes a los tres subgrupos, de los que 63 presentaron sobreexpresión y 51 infraexpresión después del tratamiento con bortezomib. Para el resto de los genes, el sentido de la expresión fue opuesto entre al menos dos de los subgrupos. En la **Figura 4.56** se muestran dos ejemplos de diagramas de bosque de los genes con mayor valor absoluto de la mediana de FC considerando los 13 estudios.

**a**




**b**

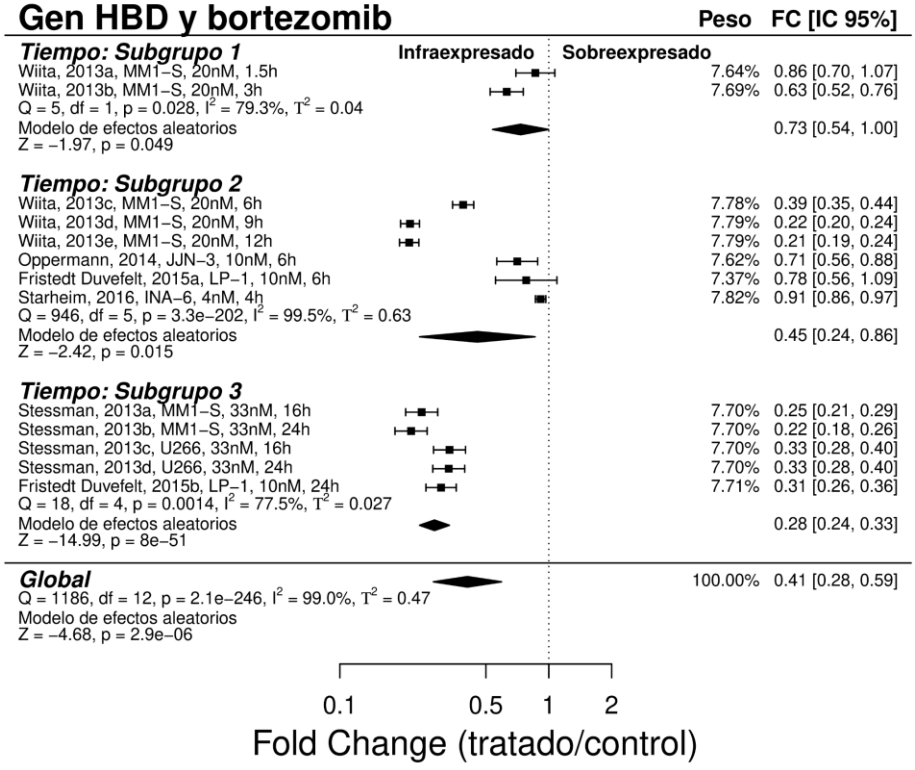



***Figura 4.56.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con bortezomib.* ***a)*** *Diagrama de bosque del gen HSPA6, que fue el más sobreexpresado considerando la mediana del FC de los 13 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen HBD, que fue el más infraexpresado considerando la mediana del FC de los 13 estudios seleccionados.*

La determinación de las diferencias en la expresión génica entre los tres subgrupos arrojó 403 genes diferencialmente expresados entre los subgrupos G1 y G2, 766 genes entre los subgrupos G1 y G3, y 398 genes entre los subgrupos G2 y G3 (**Anexo 15**). En la **Figura 4.57** se muestran, mediante un diagrama de puntos, las diferencias entre los valores de ln(FC) obtenidos en los tres subgrupos.

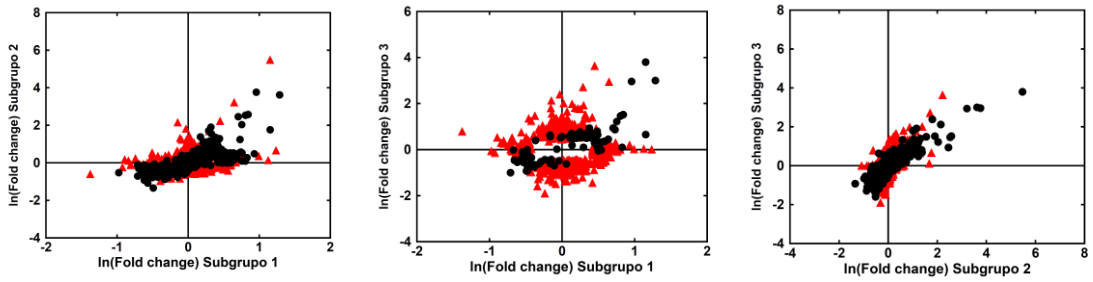



***Figura 4.57.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 863 genes estudiados donde se comparan los subgrupos 1, 2 y 3. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

En un último paso se realizó el análisis ORA utilizando los genes diferencialmente expresados entre los tres subgrupos sobre las vías KEGG y los términos GO. El objetivo de este análisis fue la determinación de las vías o términos más afectados por el tiempo de tratamiento con bortezomib. El resultado de este análisis aparece representado en la **Figura 4.58**.

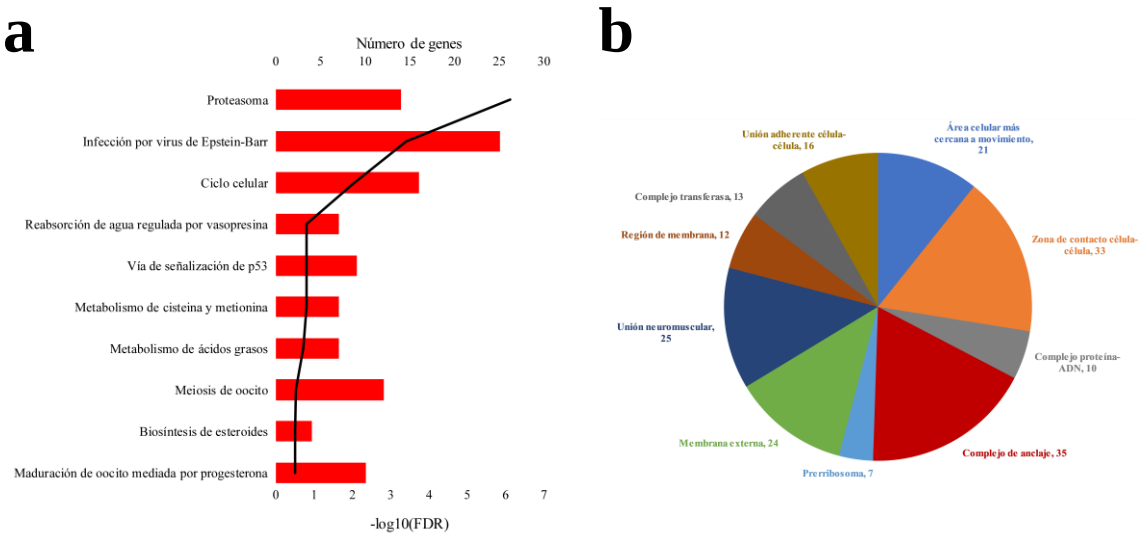



***Figura 4.58.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con bortezomib.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO.*

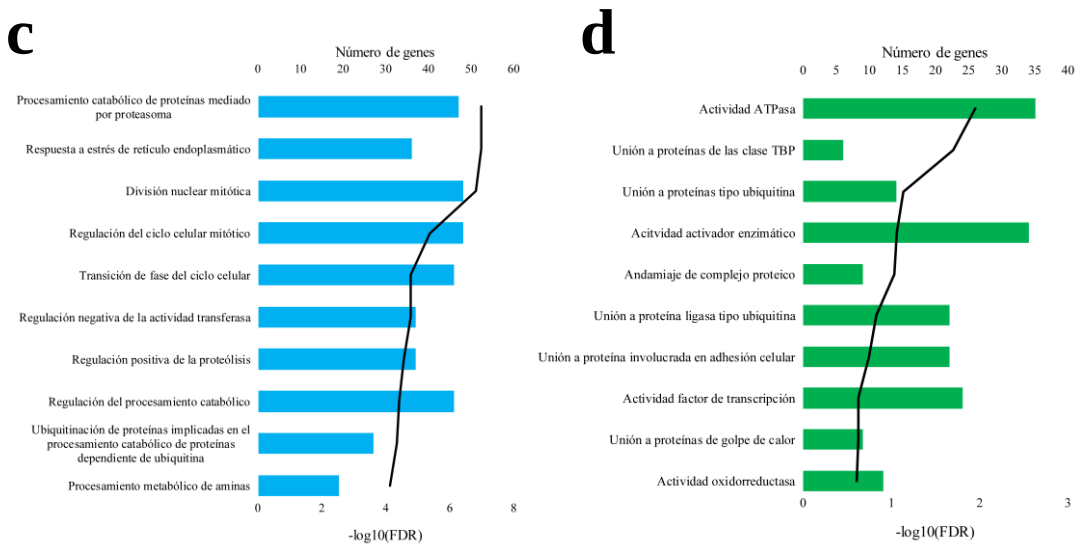



***Figura 4.58 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con bortezomib.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO. Los términos de cada panel están ordenados de mayor a menos significancia estadística.*

El análisis de vías KEGG y funciones GO mostró que las principales vías sobrerrepresentadas corresponden a las relacionadas con el procesamiento proteico. Así encontramos que 14 genes de la "vía del proteasoma" (FDR < 0,0001) presentaron diferencias en al menos dos de los grupos de tiempo. Los valores de ln(FC) de cada uno de estos genes en los tres subgrupos se muestran en la **Figura 4.59a**, donde puede observarse que el cambio de expresión aumenta gradualmente a medida que aumenta el tiempo de tratamiento. Esto supondría que el efecto del bortezomib a tiempos cortos inferiores a las tres horas, como los aplicados en el subgrupo G1, quizá sean insuficientes para que pueda llevar a cabo la sobreexpresión de estos componentes de la "vía del proteasoma". Ampliando estos genes mediante la selección de los 47 genes del PB "procesamiento catabólico de proteína mediado por proteasoma" (FDR < 0,0001) observamos un efecto similar, incluso en los genes que presentaron infraexpresión **Figura 4.59b**.

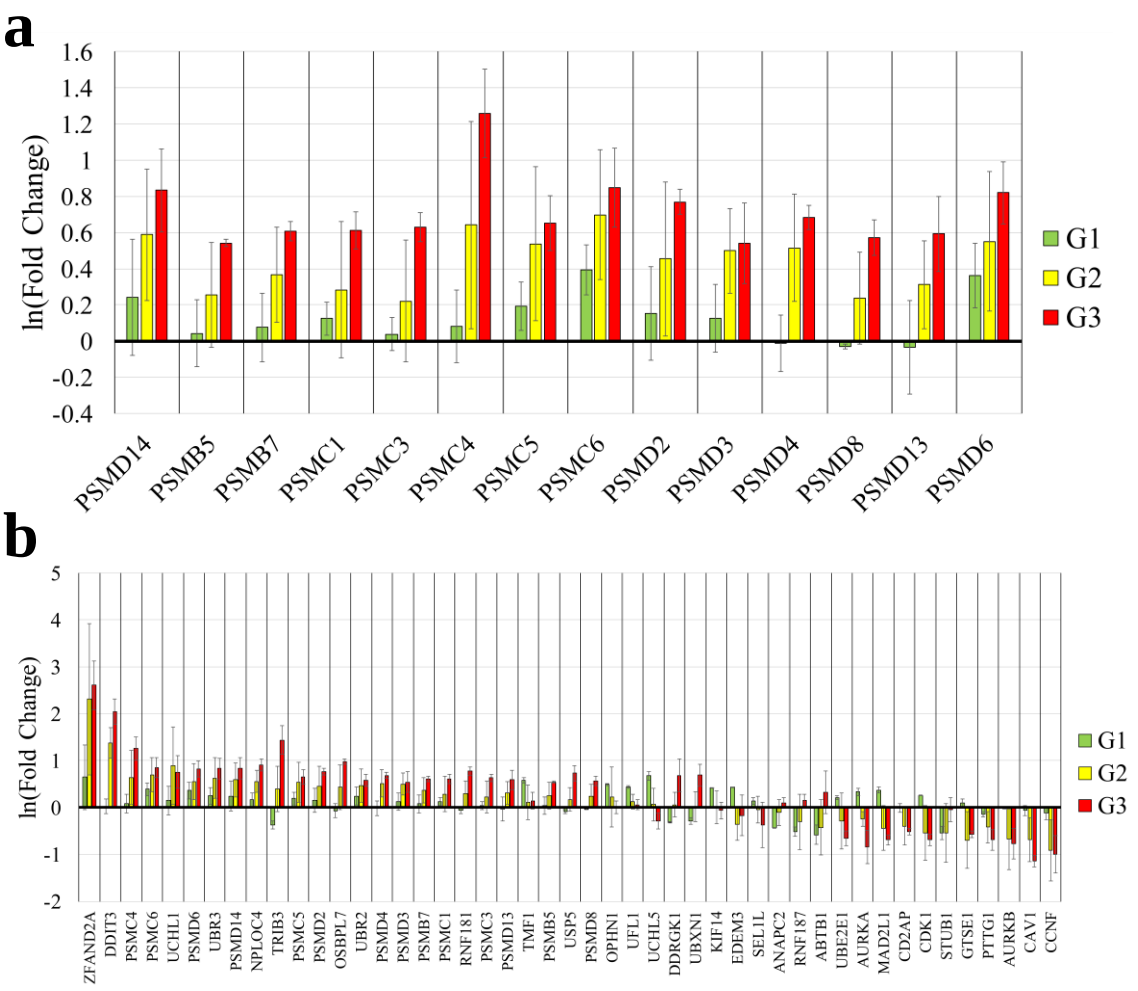


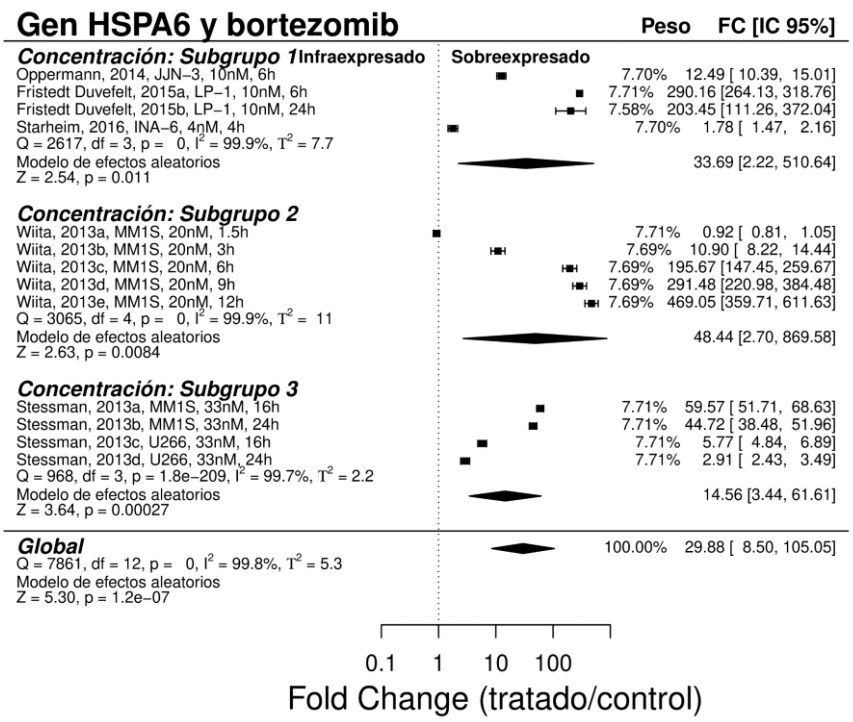

***Figura 4.59.*** *Valores promedio del ln(Fold Change) de los genes desregulados en* ***a)*** *la vía KEGG del "proteasoma" y* ***b)*** *el proceso biológico GO "procesamiento catabólico de proteína mediado por proteasoma", en los tres subgrupos de tiempo de tratamiento con bortezomib (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.3.2. Metaanálisis por subgrupos: concentración

Se establecieron tres subgrupos en función de la concentración aplicada de bortezomib. Estos subgrupos fueron determinados en función del valor de la mediana y la MAD de las concentraciones de los 13 estudios seleccionados. El primer subgrupo (G1) agrupó los estudios realizados con una concentración de bortezomib inferior o igual a 10 nM: estudios de Starheim (2016), Fristedt (2015a), Fristedt (2015b) y Oppermann (2014). El subgrupo 2 (G2) comprendió los cinco estudios en los que la concentración fue superior a 10 nM, pero inferior o igual a 30 nM: estudios [a], [b], [c], [d] y [e] de Wiita [2013]). El tercer subgrupo (G3) recogió los cuatro estudios cuya concentración fue superior a 30 nM: estudios (a), (b), (c) y (d) de Stessman (2013). A través del metaanálisis en cada uno de estos subgrupos se detectaron 277 genes estadísticamente significativos a *p*-valor < 0,05 en el G1, 540 genes en el G2 y 650 genes en el G3 (**Anexo 16**). Finalmente, 156 de estos genes fueron comunes a los tres subgrupos: presentando 87 de ellos estaban sobreexpresados y 55 infraexpresados en los tres subgrupos después de tratar con

bortezomib. En la **Figura 4.60** se muestran dos ejemplos de diagrama de bosque para el metaanálisis por subgrupos de concentración de bortezomib.

**a**




**b**

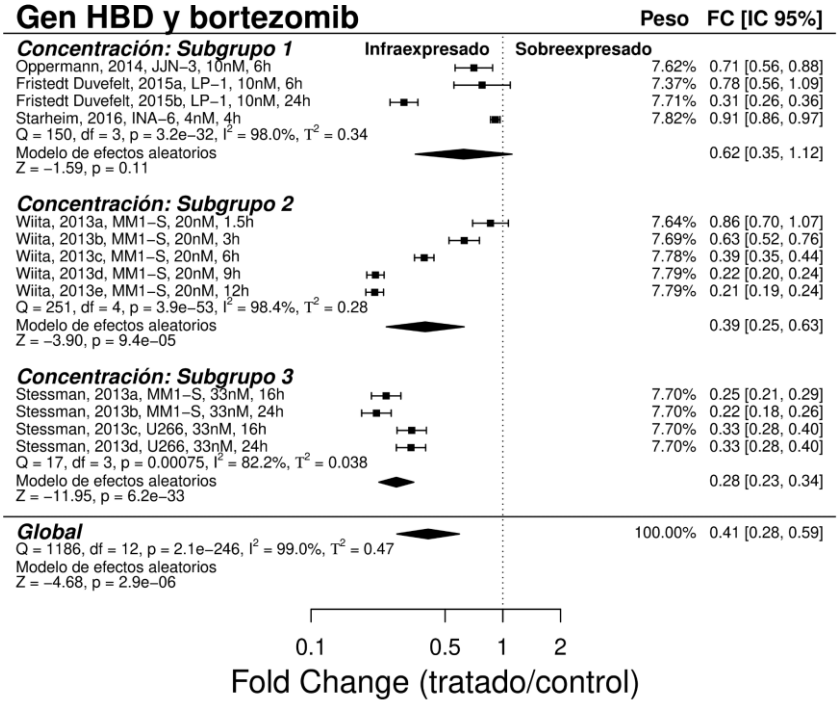



***Figura 4.60.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de bortezomib.* ***a)*** *Diagrama de bosque del gen HSPA6, que fue el más sobreexpresado considerando la mediana del FC de los 13 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen HBD, que fue el más infraexpresado considerando la mediana del FC de los 13 estudios seleccionados.*

Posteriormente, se llevó a cabo la comparación de la expresión génica entre los tres subgrupos. Se detectaron diferencias estadísticamente significativas en 263 genes entre los subgrupos G1 y G2, 429 genes entre los subgrupos G1 y G3 y 522 genes entre los subgrupos G2 y G3 (**Anexo 16**). En la **Figura 4.61** pueden observarse los valores de ln(FC) de los 863 genes seleccionados como candidatos a metaanálisis (**Apartado 4.3.3.1**) en los tres subgrupos de concentración de bortezomib en una representación mediante diagrama de puntos. En la mayoría de los casos el cambio de expresión entre los subgrupos consistió en un cambio de la intensidad de la sobre o la infraexpresión, manteniendo el sentido del cambio inducido por el fármaco, aunque hubo una pequeña proporción de genes, recogidos en los sectores superior izquierdo e inferior derecho, en los que se detectó un cambio de sentido de la expresión.

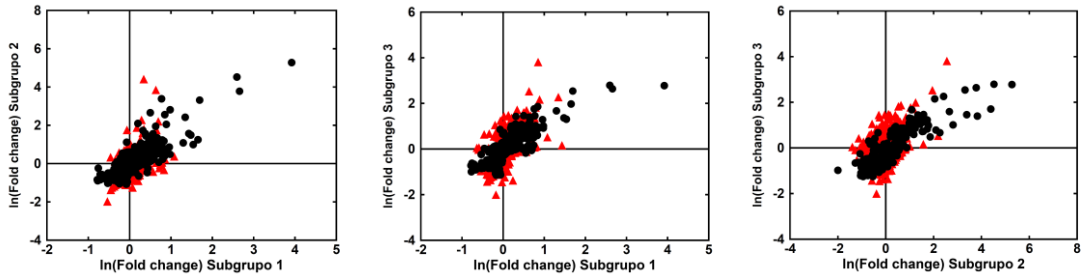



***Figura 4.61.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 863 genes estudiados donde se comparan los subgrupos 1, 2 y 3 de concentración de bortezomib. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

Como último paso de este metaanálisis por subgrupos se llevó a cabo el análisis ORA de rutas biológicas KEGG y funciones GO de los genes que presentaron diferencias estadísticamente significativas entre al menos dos de los grupos de concentración de fármaco, con el fin de determinar qué procesos se ven afectados por la aplicación de diferentes concentraciones de bortezomib. El resultado de este análisis aparece recogido en la **Figura 4.62**.

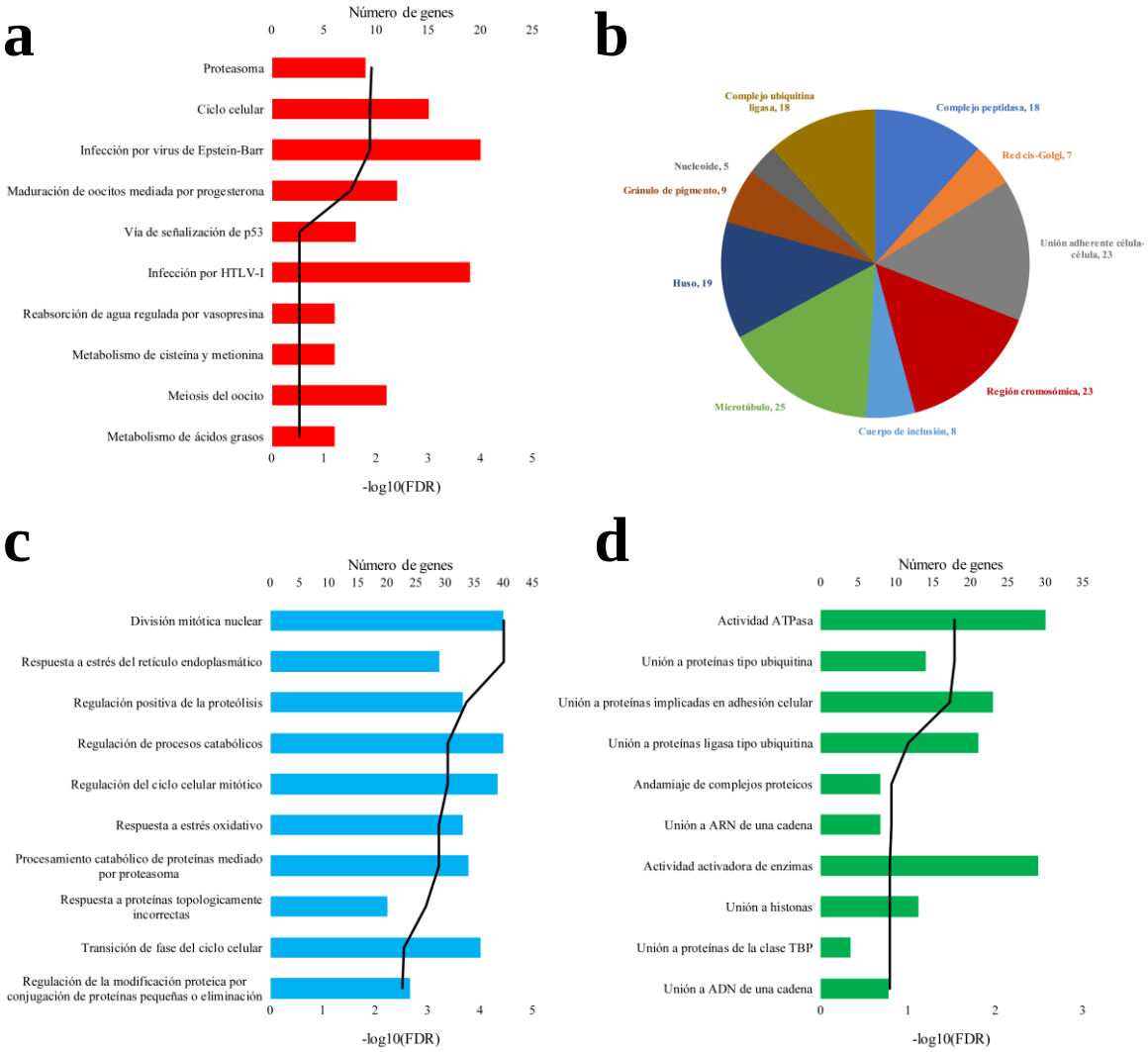



***Figura 4.62.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de bortezomib.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO. Los términos de cada panel están ordenados de mayor a menos significancia estadística.*

De nuevo, la vía KEGG con el valor de FDR más significativo fue la "vía del proteasoma" (FDR = 0,0123) (**Figura 4.62a**). Del mismo modo, entre los PB más significativos volvió a aparecer el "procesamiento catabólico de proteínas mediado por proteasoma" (FDR = 0,0006) (**Figura 4.62c**). Por tanto, ambas vías también se vieron afectadas por la concentración de bortezomib aplicada. Gran parte de los genes que aparecen en este análisis estaban también desregulados en función el tiempo de tratamiento, por lo que su regulación podría depender tanto de la concentración como del tiempo de tratamiento con bortezomib. Sin embargo, aún hay un grupo numeroso de genes desregulados con el tiempo de tratamiento que no se encontraban en este nuevo análisis (**Figura 4.63**), por lo que el grado de expresión de esos genes dependería exclusivamente del tiempo de tratamiento.

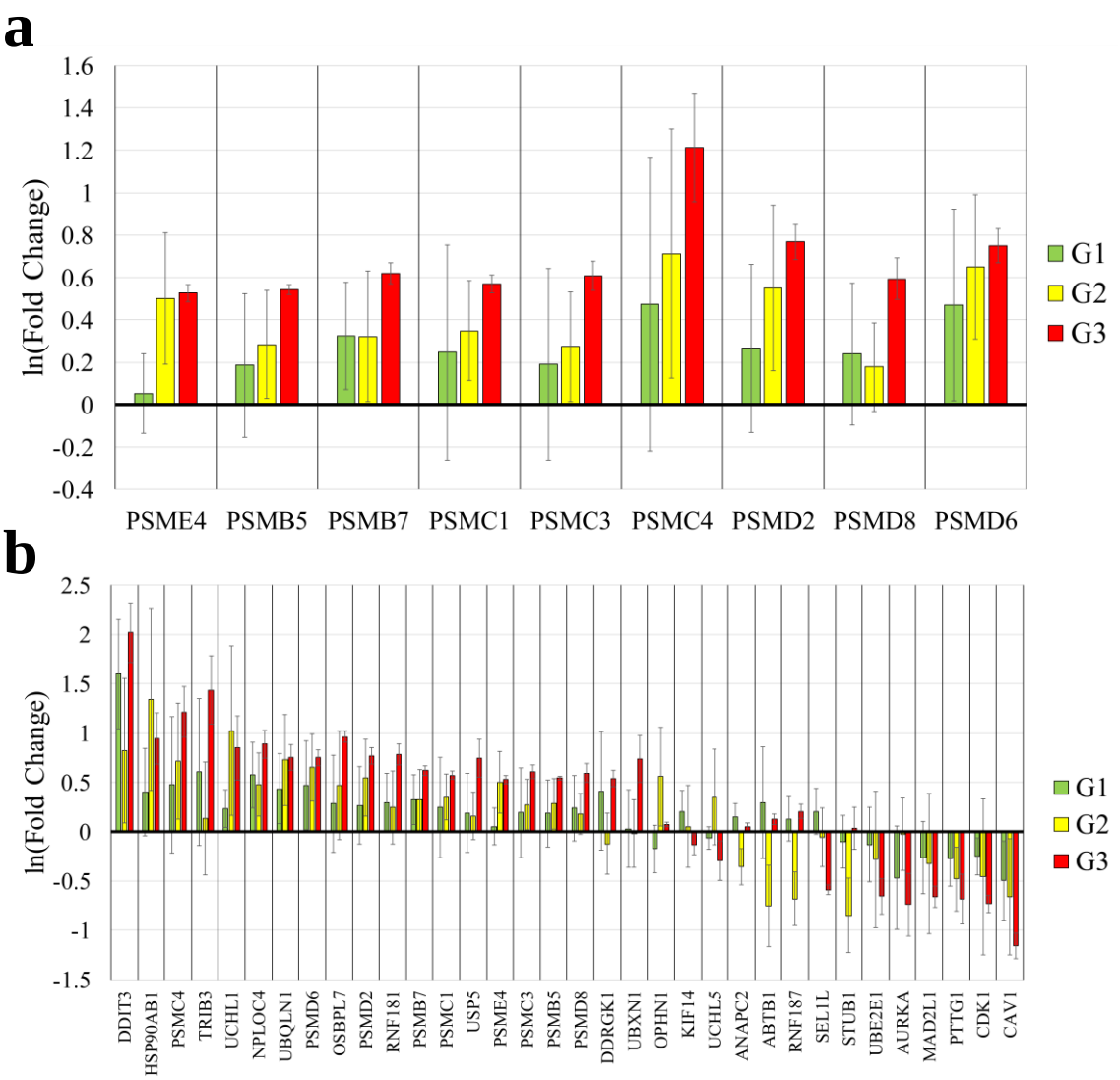



***Figura 4.63.*** *Valores promedio del ln(Fold Change) de los genes desregulados en* ***a)*** *la vía KEGG del "proteasoma" y* ***b)*** *el PB de "procesamiento catabólico de proteínas mediado por proteasoma", en los tres subgrupos de concentración de bortezomib (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.3.3. Metaanálisis global del bortezomib

El metaanálisis global se llevó a cabo considerando los 13 estudios seleccionados como un conjunto para su análisis. De este modo se reveló una diferencia en el tamaño del efecto estadísticamente significativa a *p*-valor < 0,05 en 686 genes, de los que 381 presentaron sobreexpresión y 305 infraexpresión en las muestras tratadas con bortezomib (**Anexo 17**). En la **Figura 4.64** se recogen las 10 rutas biológicas y los 10 términos GO a los niveles de PB, CC y FM, más relevantes relacionadas con estos 686 genes.

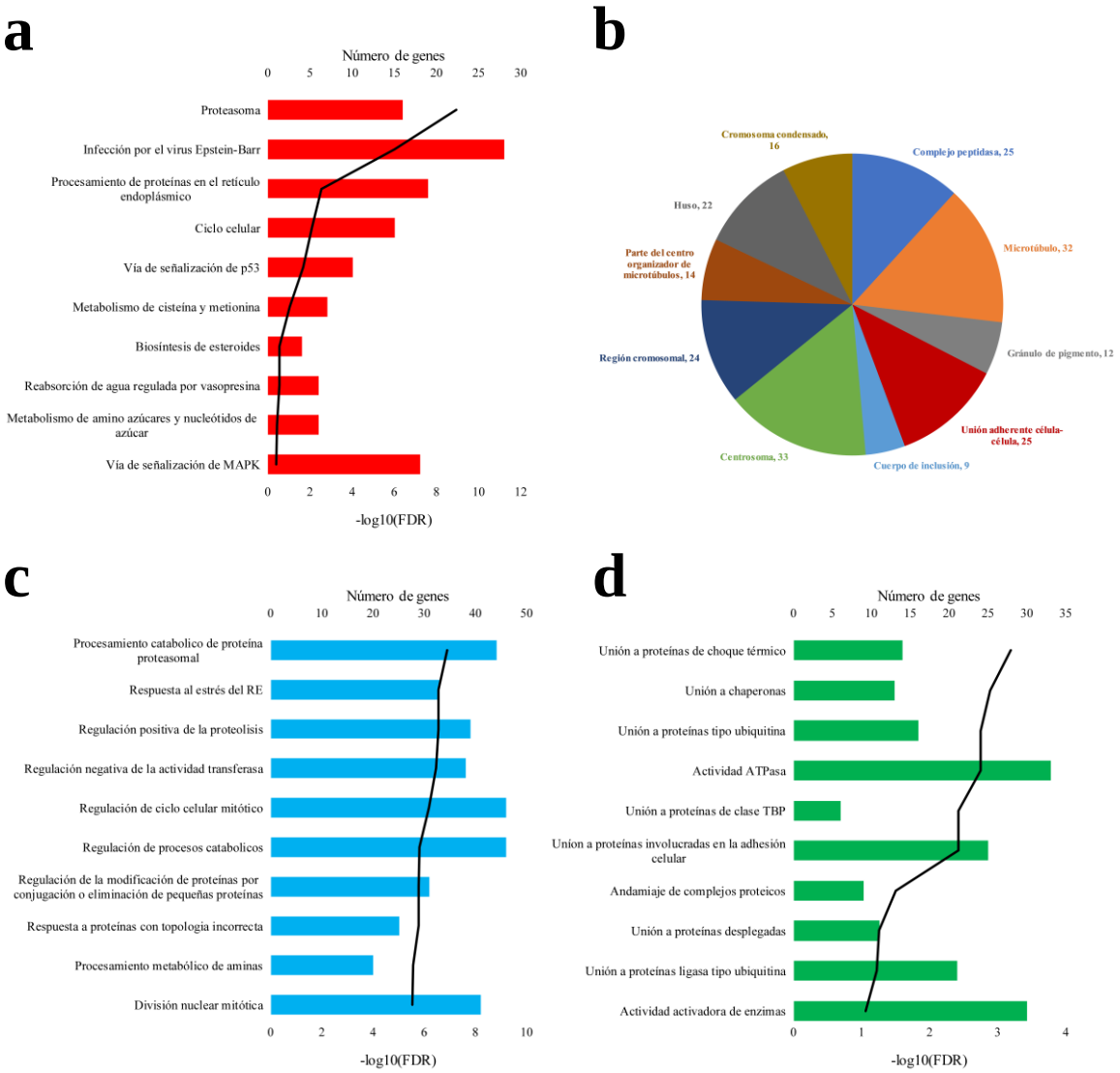



***Figura 4.64.*** *Análisis de sobrerrepresentación en rutas KEGG y términos GO considerando los 686 genes con una diferencia del tamaño del efecto estadísticamente significativa en el metaanálisis de bortezomib.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO. Los términos de cada panel están ordenados de mayor a menos significancia estadística.*

La vía KEGG sobrerrepresentada de manera más significativa fue la "vía del proteasoma" (FDR = < 0,0001) con 16 genes desregulados, todos ellos sobreexpresados tras el tratamiento con bortezomib. Todos estos genes codifican proteínas que actúan como subunidades del complejo del proteasoma (**Figura 4.65**). La dirección del cambio de expresión de estos genes indica que el bortezomib estaría promoviendo su expresión, lo que concuerda con lo publicado en otros trabajos en MM[453] y en leucemia linfocítica crónica[454], pudiendo ser este mecanismo un tipo de respuesta a estrés frente a la exposición de las células a bortezomib[453]. Por otro lado, la sobreexpresión de la proteína codificada por el gen *PSMB5* (*z*-valor = 4,49, *p*-valor < 0,0001) también ha sido asociada con un posible mecanismo de resistencia en estudios que utilizan líneas celulares resistentes a este fármaco[455].

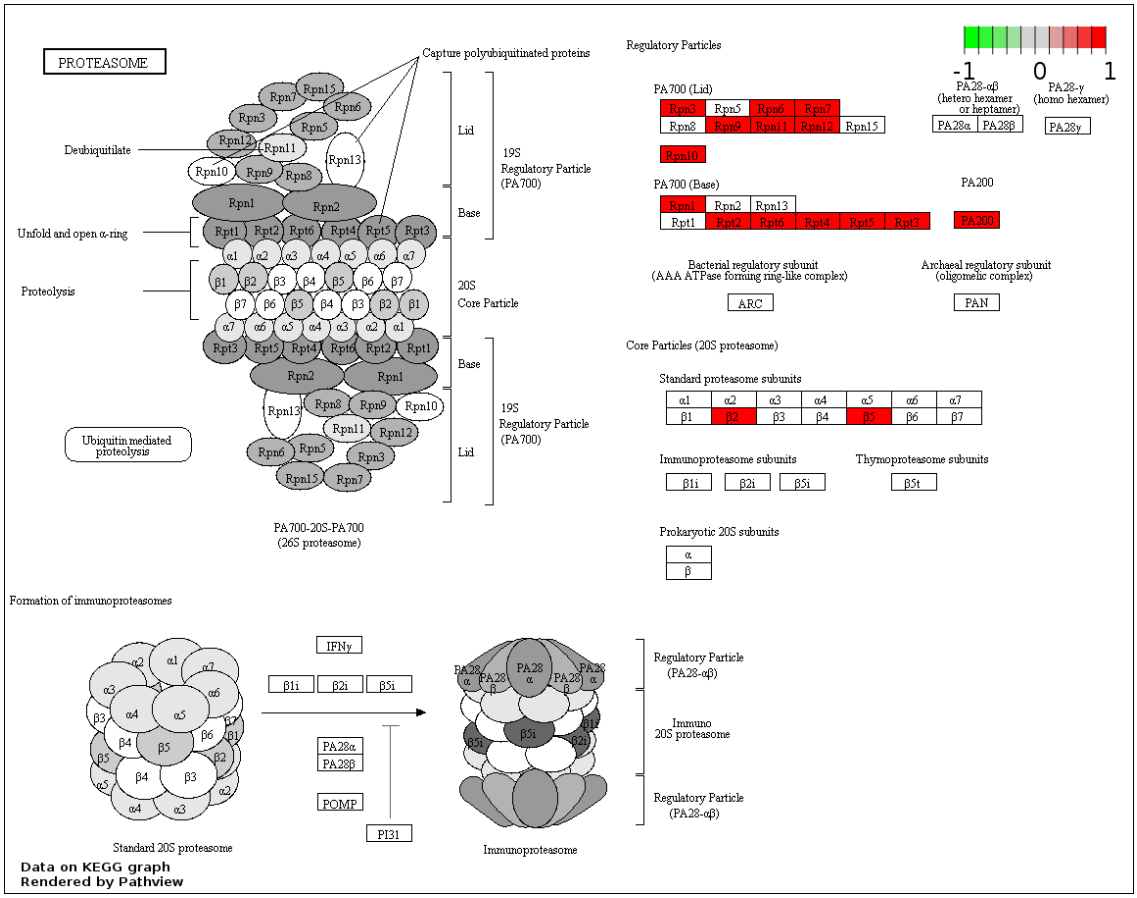



***Figura 4.65.*** *Vía del proteasoma según la base KEGG. En rojo se representan los genes sobreexpresados de manera estadísticamente significativa en el metaanálisis global del bortezomib.*

Otra de las vías KEGG significativamente sobrerrepresentadas en este trabajo fue el "procesamiento de proteínas en el retículo endoplásmico" (FDR = 0,0031) (**Figura 4.66**). Esta vía consta de 19 genes desregulados de los que 7 codifican proteínas de choque térmico (HSP, del inglés *Heat Shock Proteins*), asociadas con la respuesta celular protectora frente a estrés. Además, asociada a las HSP aparece también desregulada de forma significativa (FDR = 0,0007) la FM de "unión a proteínas de choque térmico". La sobreexpresión de estas HSP podría tratarse de un intento de la célula de contrarrestar el efecto apoptótico del bortezomib promoviendo vías favorables a la supervivencia[453]. El mecanismo de apoptosis inducida por bortezomib también podría desencadenarse a partir de los genes desregulados en esta vía. Este sería un mecanismo de respuesta a estrés y estaría mediado por los genes *ATF4* (*z*-valor = 3,69, *p*-valor = 0,0002) y *DDIT3* (*z*-valor = 5,86, *p*-valor < 0,0001). El estrés en el retículo endoplásmico provocaría la sobreexpresión de ambos genes, y esto conduciría a la activación del gen *TRIB3* (*z*-valor = 2,95, *p*-valor = 0,0032), también sobreexpresado, produciendo la apoptosis celular[456].

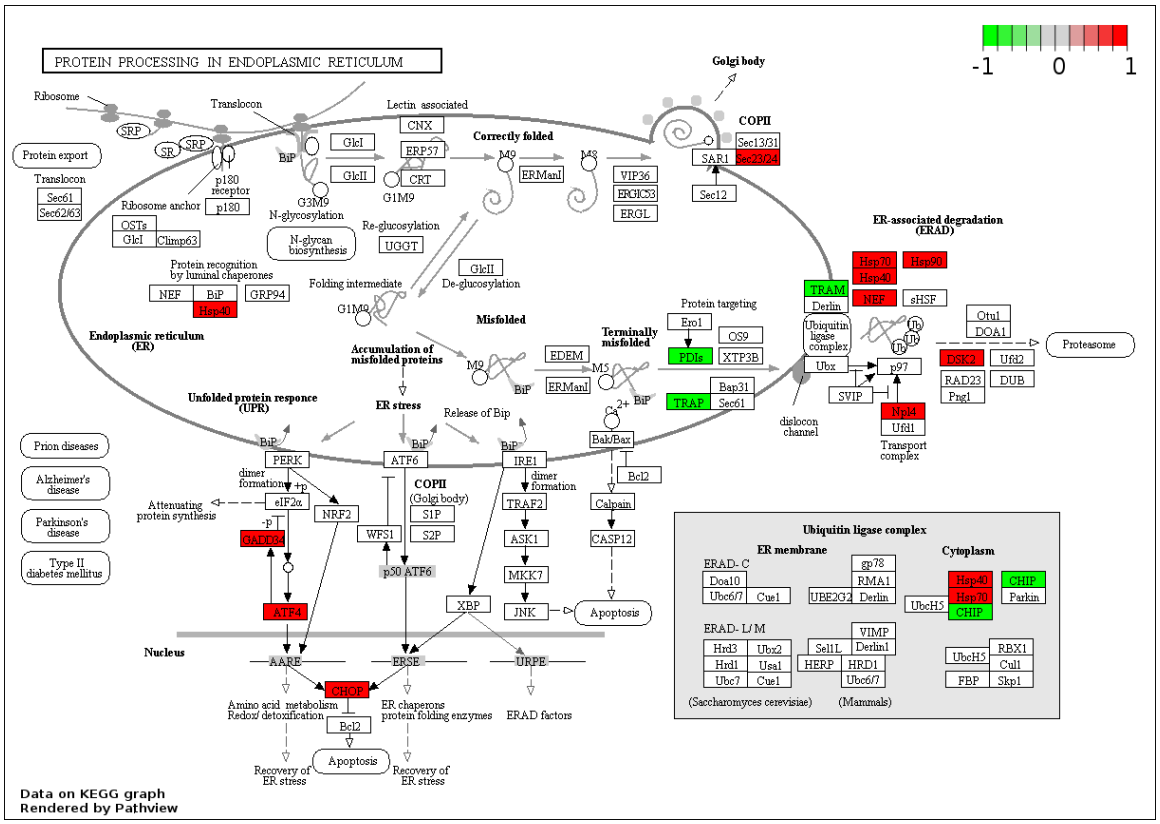



***Figura 4.66.*** *Vía de procesamiento de proteínas en el retículo endoplásmico según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del bortezomib.*

### 4.3.4. Lenalidomida

La acción de la lenalidomida sobre la expresión génica fue evaluada en este trabajo mediante revisión sistemática con metaanálisis. Para ello, se realizó una búsqueda sistemática de estudios en repositorios de datos de expresión génica en la que se hallaron 22 series en GEO, 10 series en ArrayExpress y cuatro muestras correspondientes a una única serie en SRA. Tras la eliminación de los elementos duplicados, 24 series fueron finalmente seleccionadas para su evaluación detallada en función de los criterios de inclusión y exclusión previamente determinados. De las 24 series revisadas, dos cumplieron los criterios de inclusión y exclusión necesarios para ser incluidas en el metaanálisis. Adicionalmente, se añadieron dos series cedidas por el servicio de Hematología de Salamanca. Se comprobó la posible subdivisión de las cuatro series en diferentes estudios en función de las concentraciones de fármaco empleadas, el tiempo de tratamiento o la utilización de varias líneas celulares. Siguiendo estos criterios, la serie GSE31452 fue dividida en dos estudios en función de los tiempos de tratamiento empleados. Por tanto, el número final de estudios considerados para el metaanálisis fue de cinco. El diagrama de flujo que se muestra en la **Figura 4.67**, detalla el esquema de selección de estudios para el metaanálisis de lenalidomida en función de los diferentes criterios de inclusión y exclusión.

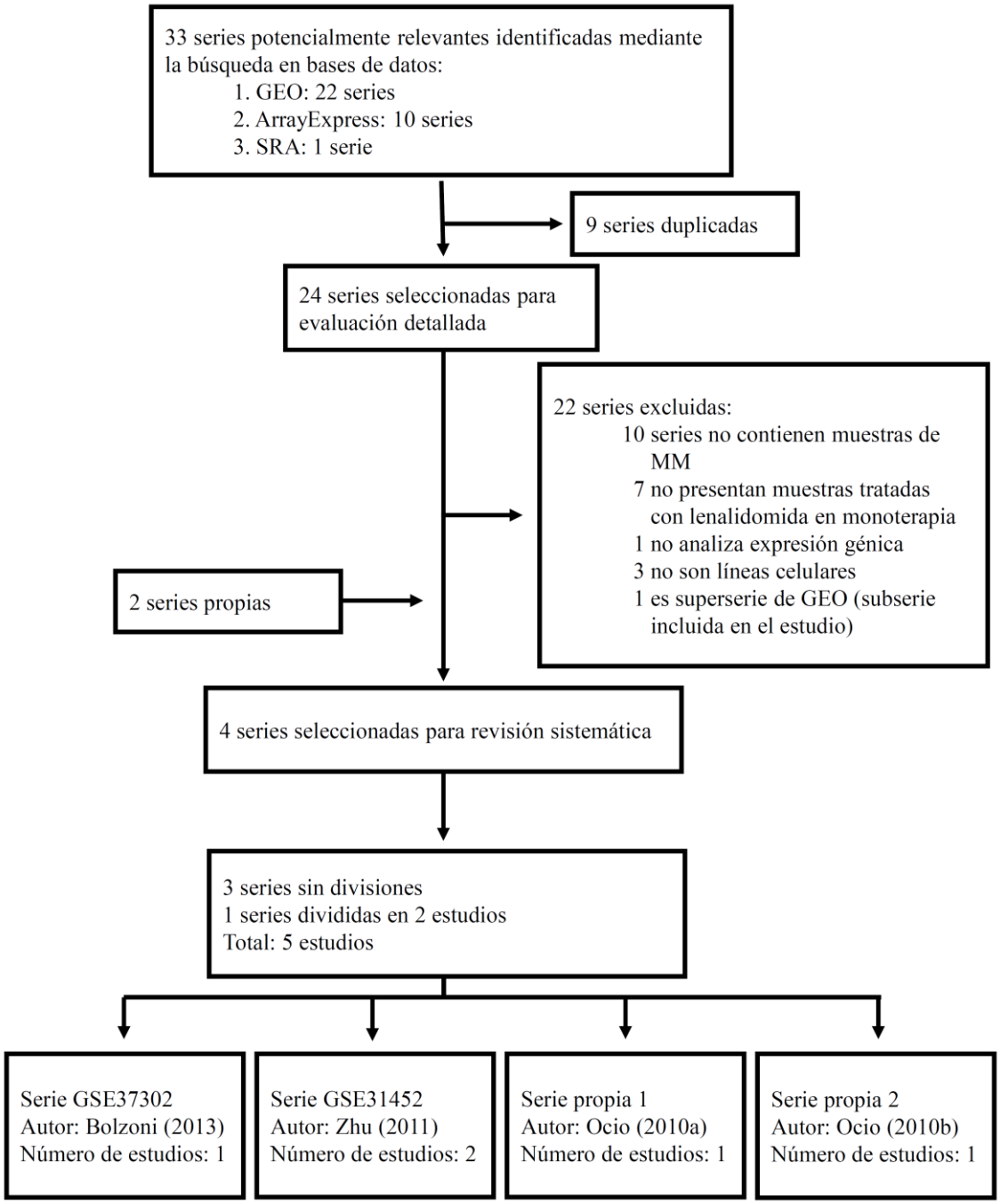



***Figura 4.67.*** *Diagrama de flujo de la selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con lenalidomida.*

Los cinco estudios seleccionados fueron clasificados en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de lenalidomida. Se establecieron los tiempos a 24 y a 72 horas (mediana de 48 horas) como puntos de corte del tiempo de tratamiento, mientras que en el caso de la concentración los puntos de corte se establecieron a 1 µM y a 59 µM (mediana de 30 µM). El resultado del agrupamiento en subgrupos y las características más relevantes de los cinco estudios seleccionados se recogen en la **Tabla 4.7**.

***Tabla 4.7**. Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con lenalidomida.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (µM) |
|---|---|---|---|---|---|---|
| GSE37302 | Bolzoni (2013)[97] | JJN-3 | Affymetrix Human Genome U133 Plus 2.0 | 6 | 24 | 100 |
| GSE31452 | Zhu (2011a)[330] | OPM2 | Affymetrix Human Genome U133 Plus 2.0 | 2 | 48 | 30 |
| GSE31452 | Zhu (2011b)[330] | OPM2 | Affymetrix Human Genome U133 Plus 2.0 | 2 | 72 | 30 |
| Salamanca | Ocio (2010a) | MM1-S | Affymetrix Human Gene 1.0ST | 4 | 24 | 1 |
| Salamanca | Ocio (2010b) | MM1-S | Affymetrix Human Gene 1.0ST | 4 | 120 | 1 |

*En verde, estudios seleccionados para el subgrupo G1, en amarillo, para el subgrupo G2 y en rojo, para el subgrupo G3, de tiempo de tratamiento o concentración de lenalidomida.*

La selección de genes para su estudio mediante metaanálisis se llevó a cabo mediante el criterio del FC explicado en la **Sección de Material y métodos**, de manera que se consideraron 1.164 genes como candidatos. La distribución de los ln(FC) de estos genes aparece reflejada en la **Figura 4.68**, donde se muestra una mayor dispersión de los ln(FC) en los estudios (a) y (b) de Zhu (2011) que en el resto de estudios, ambos pertenecientes a los subgrupos intermedios tanto de tiempo de tratamiento como de concentración. La influencia de estos factores sobre la expresión génica será evaluada a continuación mediante metaanálisis por subgrupos.

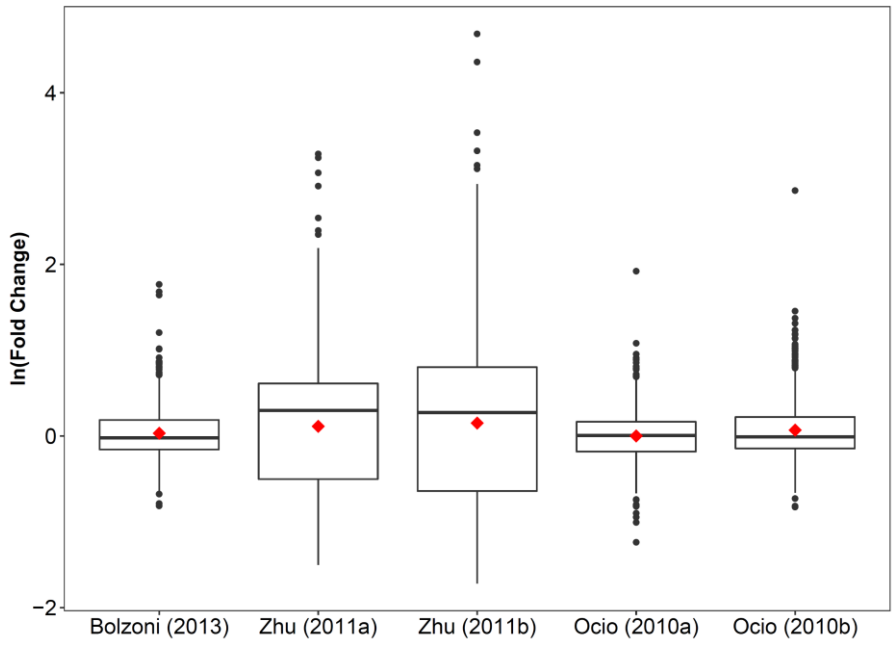



***Figura 4.68.** Diagrama de caja (box plot) del ln(Fold Change) (ln[FC])de los 1.164 genes seleccionados para el metaanálisis de la expresión génica en líneas celulares de MM tratadas con lenalidomida. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.4.1. Metaanálisis por subgrupos: tiempo de tratamiento

Los cinco estudios seleccionados se clasificaron en tres subgrupos en función de los tiempos de tratamiento con lenalidomida. El primer subgrupo (G1) comprendió los estudios con un tiempo menor o igual a las 24 horas: estudios de Bolzoni (2013) y Ocio (2010a). El subgrupo 2 (G2) agrupó los estudios con un tiempo superior a las 24 horas, pero inferior o igual a 72 horas: estudios (a) y (b) de Zhu (2011). Mientras que en el subgrupo 3 (G3) se recogió el único estudio con un tiempo superior a las 72 horas: estudio de Ocio (2010b). El metaanálisis por subgrupos de tiempo detectó 559 genes estadísticamente significativos a *p*-valor < 0,05 en el subgrupo G1, 1.107 genes en el subgrupo G2 y 776 genes en el subgrupo G3 (**Anexo 18**). Trescientos ochenta y cuatro de estos genes presentaron desregulación estadísticamente significativa de forma común a los tres subgrupos, de los que 178 presentaron sobreexpresión y 185 infraexpresión tras el tratamiento con lenalidomida. En la **Figura 4.69** se muestran los dos genes con mayor y menor valor mediano del FC para los cinco estudios de lenalidomida.

**a**

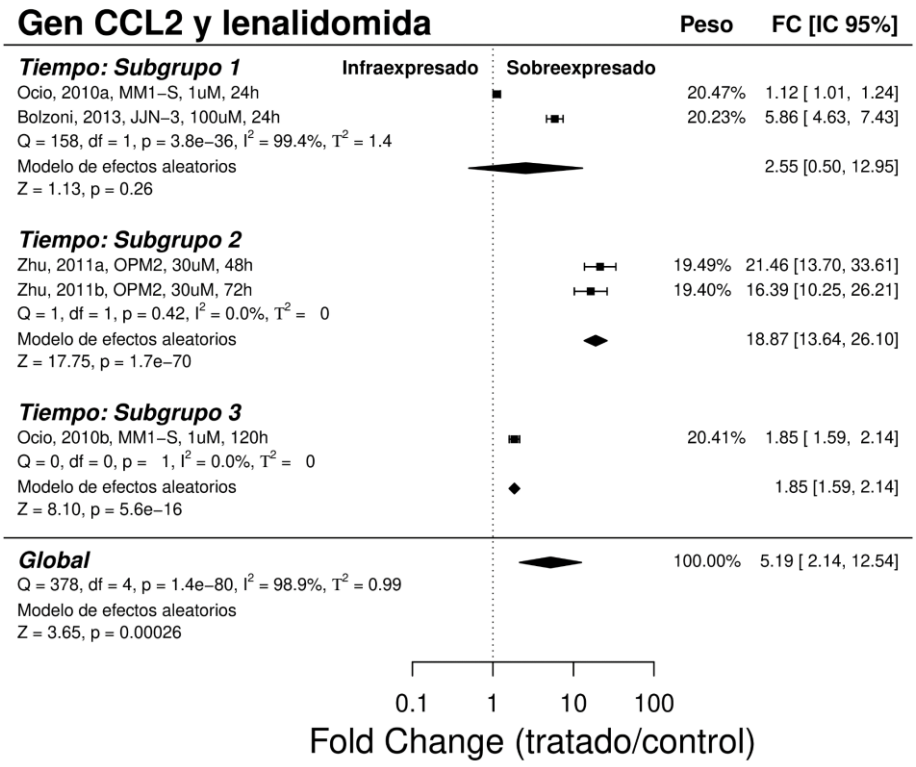



***Figura 4.69.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con lenalidomida.* ***a)*** *Diagrama de bosque del gen CCL2, que fue el más sobreexpresado considerando la mediana del FC de los cinco estudios seleccionados.*

**b**

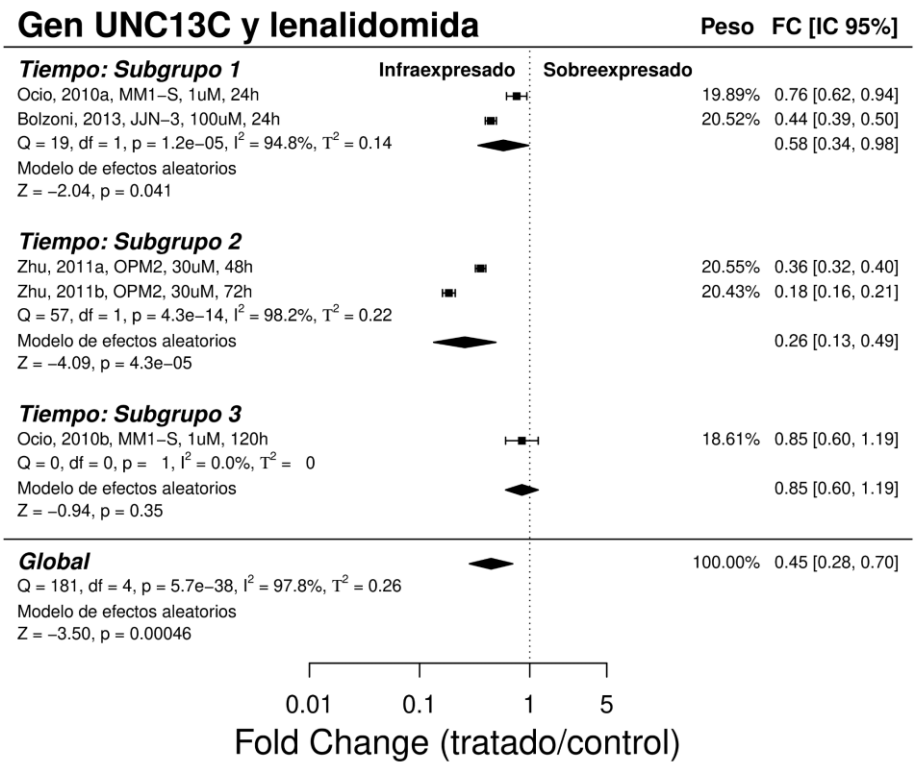



***Figura 4.69 (continuación).*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con lenalidomida.* ***b)*** *Diagrama de bosque del gen UNC13C, que fue el más infraexpresado considerando la mediana del FC de los cinco estudios seleccionados.*

Al contrastar las diferencias entre los tres subgrupos sobre los 1.164 genes analizados, se detectaron diferencias de expresión estadísticamente significativas (*p*-valor < 0,05) en 954 genes entre los subgrupos G1 y G2, en 299 genes entre los subgrupos G1 y G3, y en 1.022 genes entre los subgrupos G2 y G3 (**Anexo 18**). Las mayores diferencias se detectaron en las comparaciones realizadas frente al subgrupo G2 (**Figura 4.70**).

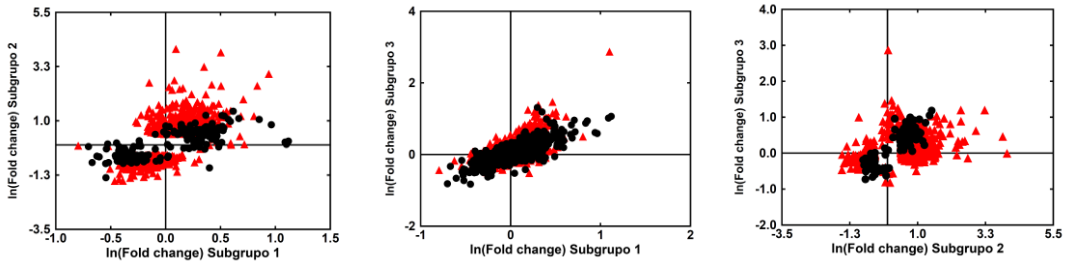



***Figura 4.70.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 1.164 genes estudiados donde se comparan los subgrupos 1, 2 y 3 del metaanálisis por subgrupos de tiempo de tratamiento con lenalidomida. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

Con los genes que presentaron diferencias de expresión estadísticamente significativas entre los tres subgrupos de tiempo de tratamiento se realizó el análisis ORA considerando las bases de vías y funciones KEGG y GO (**Figura 4.71**).

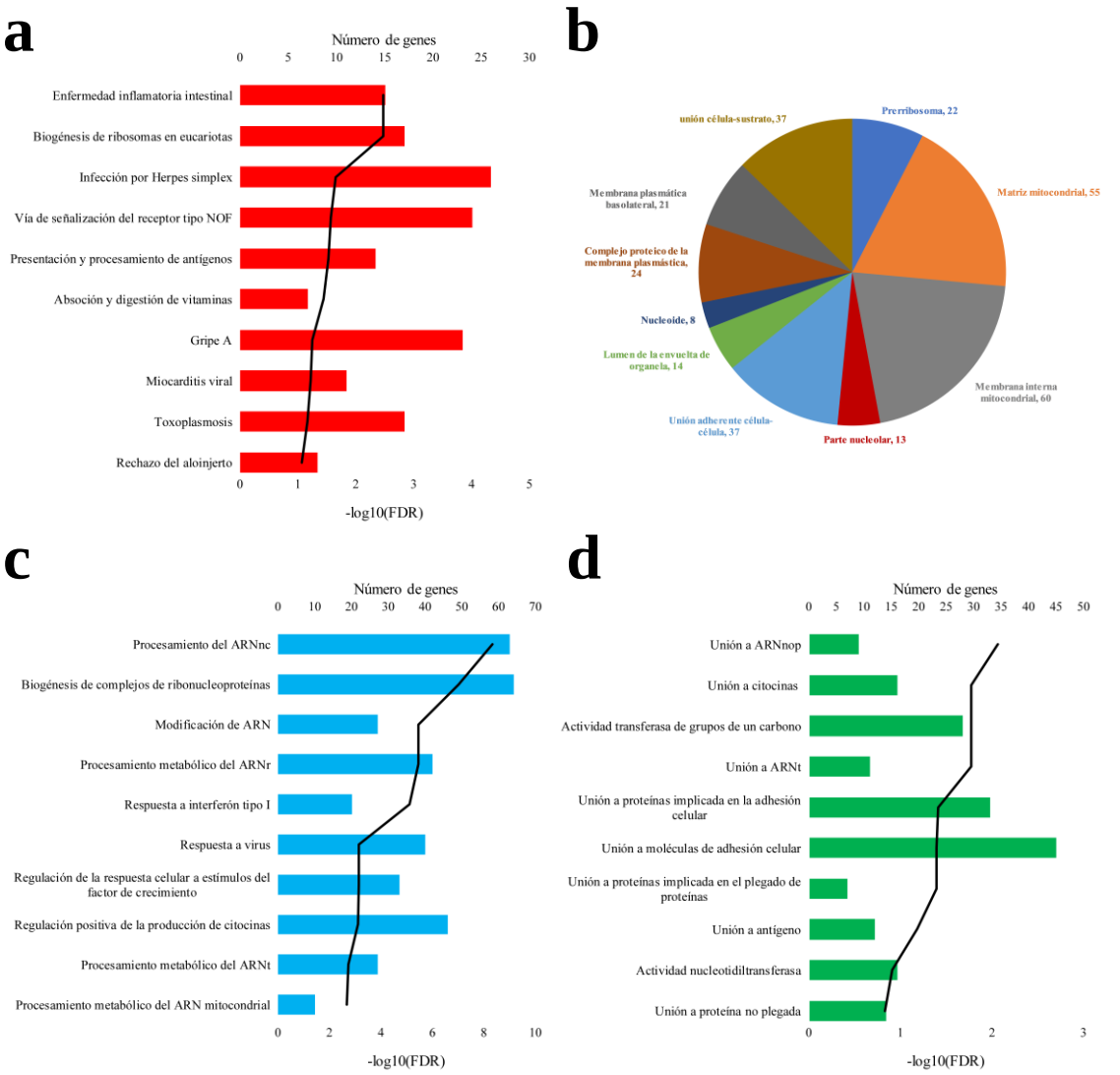



***Figura 4.71.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con lenalidomida. En esta figura se recogen las 10 rutas KEGG y los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Una de las vías KEGG entre las 10 más significativas en función del FDR en el análisis ORA fue la "biogénesis de ribosomas en eucariotas" (FDR = 0,0034). De los 82 genes que intervienen en esta ruta, 17 presentaron diferencias de expresión en al menos dos de los subgrupos de tiempo de tratamiento en nuestro trabajo. La lenalidomida ejerció una infraexpresión en estos 17 genes más acusada, en el subgrupo G2 que en los otros dos subgrupos (**Figura 4.72**). Sin embargo, no es posible achacar este efecto directamente al tiempo de tratamiento, ya que podría ser debido a la plataforma de análisis o a otro factor indeterminado, si tenemos en cuenta que los dos estudios que forman parte del

subgrupo G2 pertenecen a la misma serie (serie GSE31452). Además, el resultado de las comparaciones frente al subgrupo G3 puede no ser muy preciso al constar este subgrupo de un único estudio.

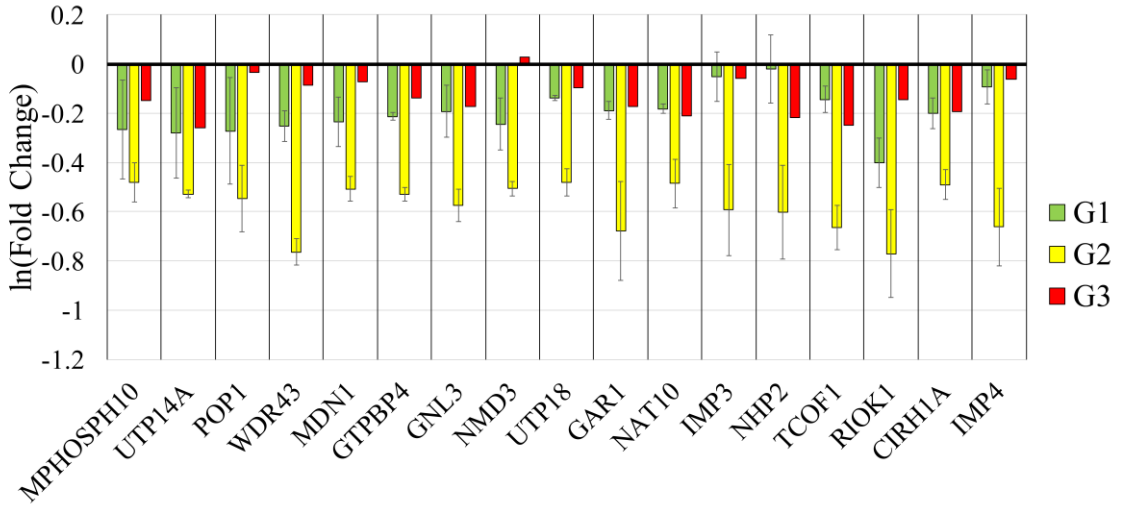
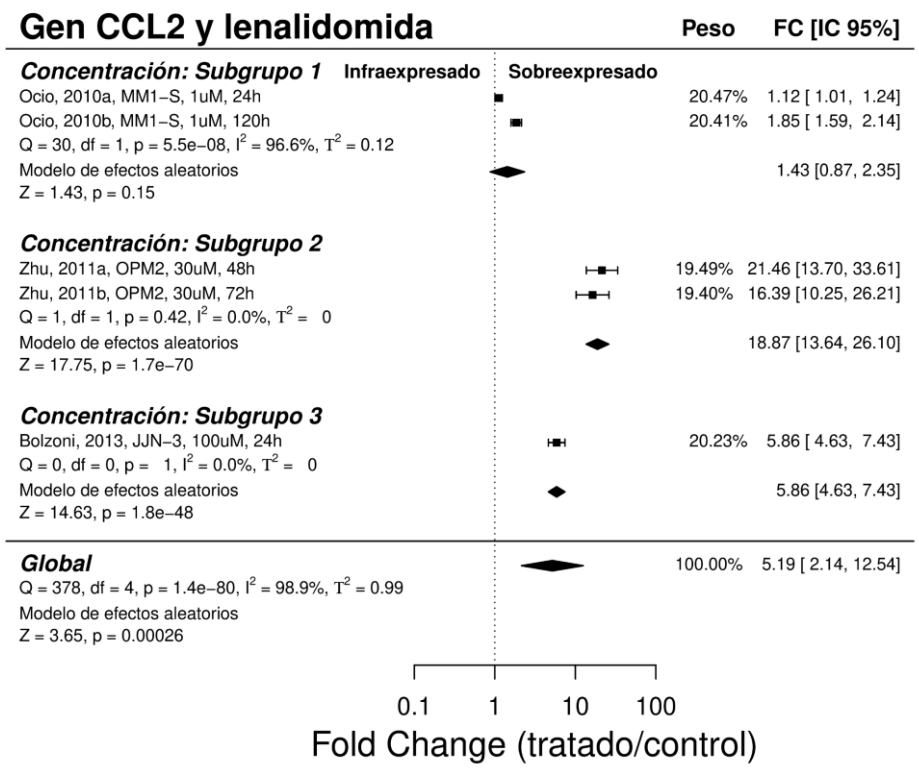



***Figura 4.72.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía de biogénesis de ribosomas en eucariotas en los tres subgrupos de tiempo de tratamiento con lenalidomida (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.4.2. Metaanálisis por subgrupos: concentración

La estratificación de los cinco estudios en función de la concentración de lenalidomida dio lugar a tres subgrupos. En el primer subgrupo (G1) se recogieron los estudios llevados a cabo con una concentración del compuesto inferior o igual a 1 µM: los estudios (a) y (b) de Ocio (2010) cumplieron este criterio. El subgrupo 2 (G2) aglutinó los estudios con concentraciones superiores a 1 µM pero inferiores o iguales a 59 µM: estudios (a) y (b) de Zhu (2011). Por último, el subgrupo 3 (G3) recopiló los estudios con una concentración superior a 59 µM, por lo que únicamente el estudio de Bolzoni (2013) satisfizo este requerimiento. El metaanálisis sobre los genes seleccionados en estos subgrupos resultó en 532 genes con expresión génica diferencial estadísticamente significativa a *p*-valor < 0,05 en el subgrupo G1, 1.107 genes en el subgrupo G2 y 776 genes en el subgrupo G3 (**Anexo 19**). Los cruces de las listas de genes diferencialmente expresados identificaron 428 genes comúnmente desregulados en los tres subgrupos, de los que 219 presentaron sobreexpresión y 188 infraexpresión al tratar con lenalidomida. En la **Figura 4.73** se recogen los diagramas de árboles de los genes *CCL2* y *UNC13C* como ejemplo de los metaanálisis por subgrupos de concentración.

**a**




**b**

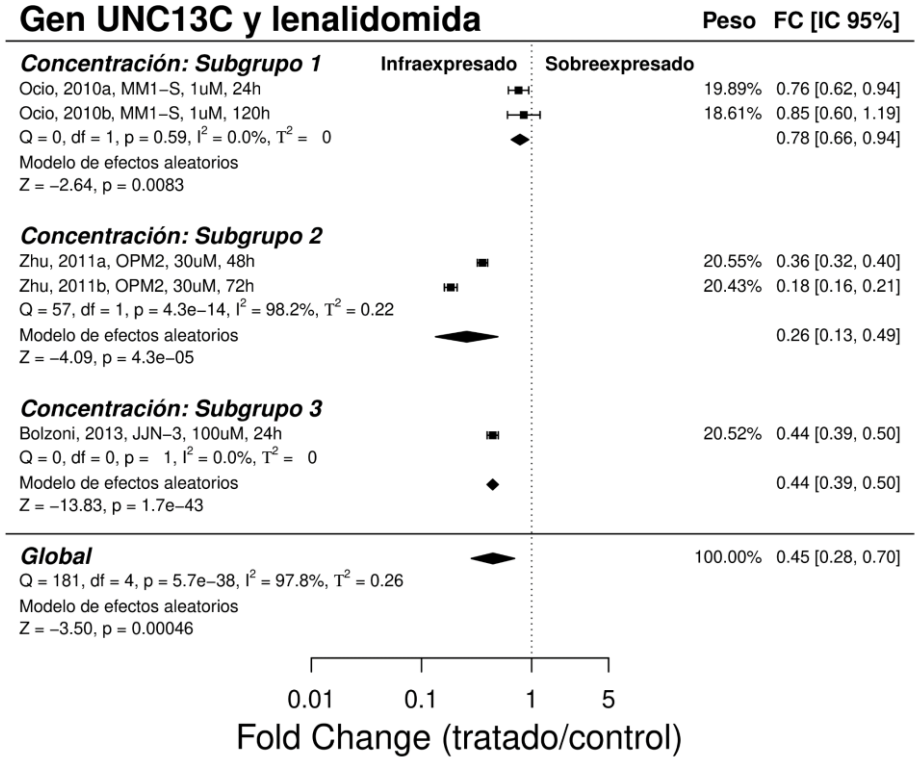



***Figura 4.73.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de lenalidomida.* ***a)*** *Diagrama de bosque del gen CCL2, que fue el más sobreexpresado considerando la mediana del FC de los cinco estudios seleccionados.* ***b)*** *Diagrama de bosque del gen UNC13C, que fue el más infraexpresado considerando la mediana del FC de los tres estudios seleccionados.*

La comparación de la expresión génica entre los tres subgrupos determinó la expresión diferencial de 960 genes entre los subgrupos G1 y G2, 341 genes entre los subgrupos G1 y G3, y 1.015 genes entre los subgrupos G2 y G3 (**Anexo 19**). Estas diferencias se muestran mediante diagrama de puntos en la **Figura 4.74**, donde se observa que al igual que en el metaanálisis por subgrupos para el tiempo de tratamiento, las diferencias de expresión con el subgrupo G2 fueron las más acusadas.

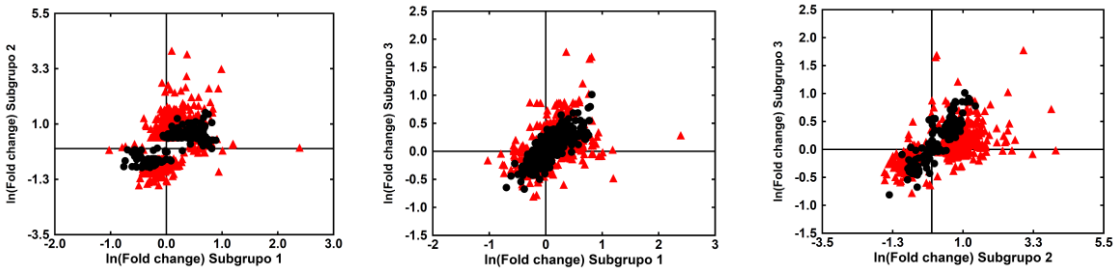



***Figura 4.74.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 1164 genes estudiados donde se comparan los subgrupos 1, 2 y 3 del metaanálisis por subgrupos de concentración de lenalidomida. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

Por último, se procedió al análisis ORA con estos genes que presentaron diferencias estadísticamente significativas entre los subgrupos de concentración de lenalidomida. Los resultados de los términos KEGG o GO que fueron más afectados por las diferencias de concentración de lenalidomida se recogen en la **Figura 4.75**.

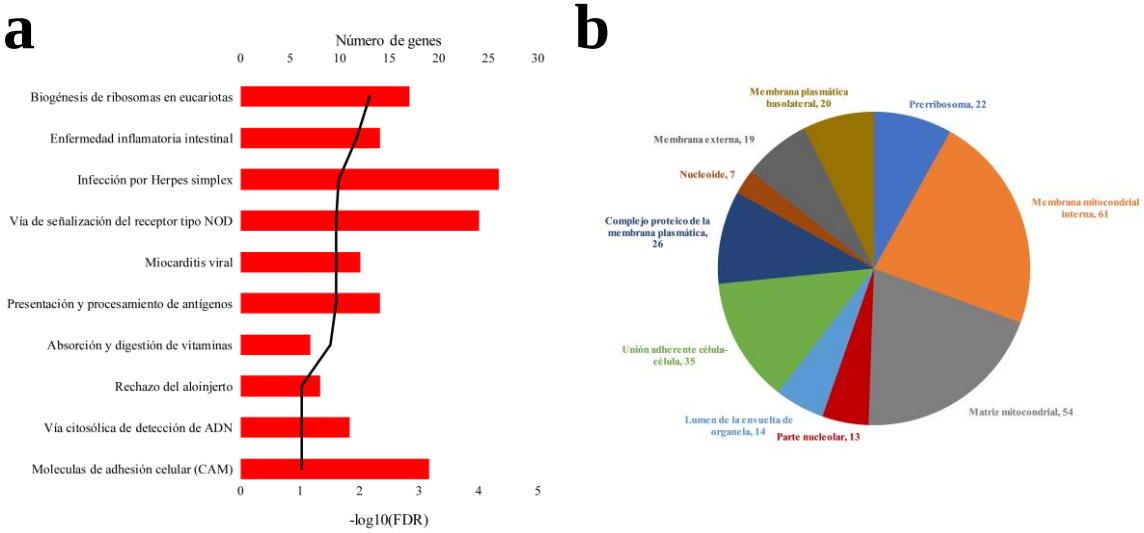



***Figura 4.75.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de lenalidomida. En esta figura se recogen las 10 vías KEGG y los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

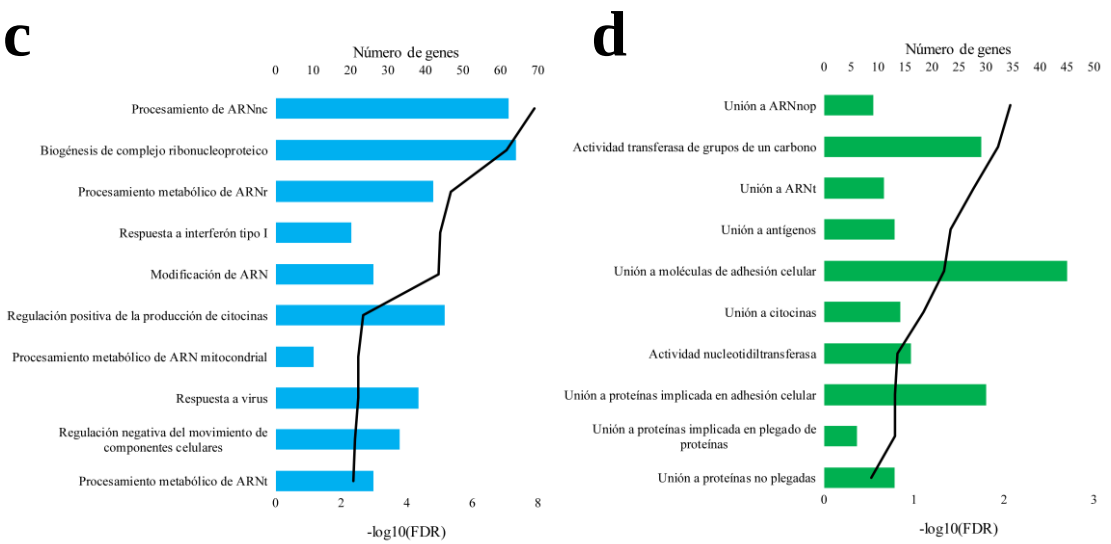



***Figura 4.75 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de lenalidomida. En esta figura se recogen las 10 vías KEGG y los10 términos GO con un menor valor de FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

De forma similar al análisis por subgrupos de tiempo de tratamiento, la vía de "biogénesis de ribosomas en eucariotas" fue una de las más significativamente sobrerrepresentadas con un FDR = 0,0068. Aunque en este caso también los genes del subgrupo G2 mostraron una desregulación mayor cuando se aplicó el tratamiento con lenalidomida (**Figura 4.76**), este efecto puede ser debido no solo a la concentración de fármaco, sino también a otros factores ya que los estudios que forman este subgrupo G2 también pertenecen a la misma serie: GSE31452.

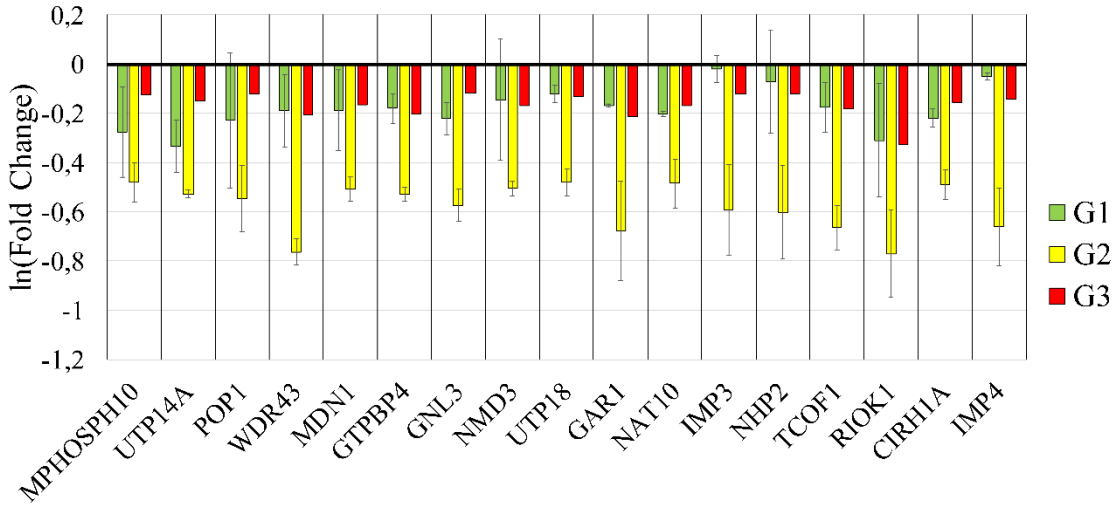



***Figura 4.76.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía de biogénesis de ribosomas en eucariotas en los tres subgrupos de concentración de lenalidomida (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

### 4.3.4.3. Metaanálisis global de lenalidomida

El metaanálisis global considerando los cinco estudios en conjunto reveló el efecto combinado estadísticamente significativo a *p*-valor < 0,05 de 948 genes, de los que 470 genes presentaron sobreexpresión y 478 infraexpresión en las muestras tratadas con lenalidomida. La lista con los 948 genes así como su sentido de expresión y significancia estadística puede ser consultadas en el **Anexo 20**. En la **Figura 4.77** se recoge el análisis ORA realizado sobre estos genes recogiéndose las 10 vías de señalización KEGG más representativas, así como las funciones GO más relevantes.

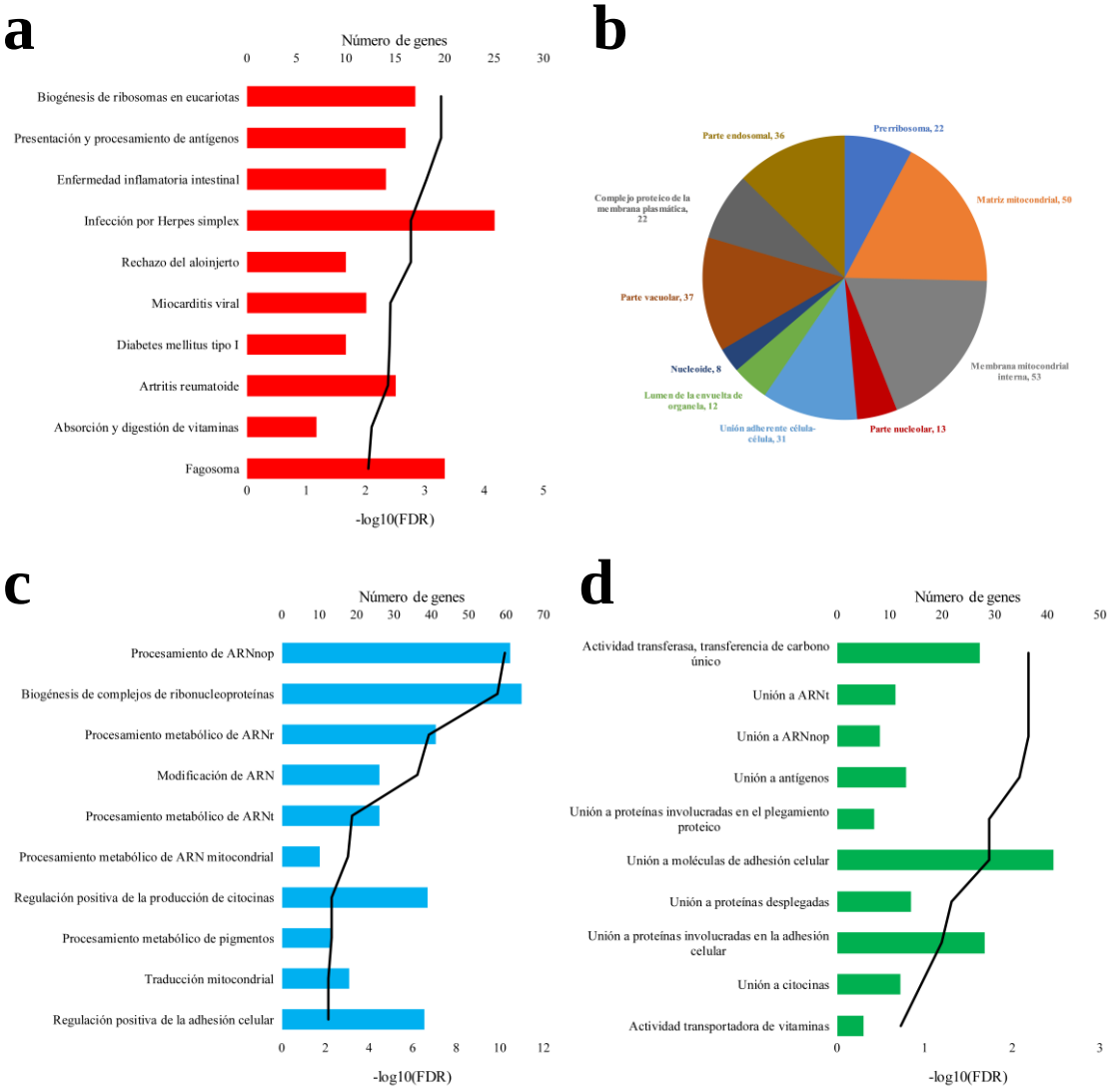



***Figura 4.77.*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 948 genes con un tamaño del efecto estadísticamente significativos en el metaanálisis de la expresión génica del tratamiento con lenalidomida. En cada panel se recogen los 10 términos con mayor significancia estadística en función del FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

La vía KEGG con un menor valor de FDR fue la "biogénesis de ribosomas en eucariotas" (FDR = 0,0005) (**Figura 4.78**). Esta vía consta de 17 genes desregulados,

todos ellos infraexpresados en las muestras tratadas con lenalidomida. La supresión de esta vía a través del tratamiento con lenalidomida ya se había descrito previamente en MM[457]. Su desregulación podría estar desencadenada por la infraexpresión del protooncogén *MYC* (*z*-valor = -2,85, *p*-valor = 0,0043), que coordina la transcripción de los genes necesarios para el procesamiento de los precursores del ARNr, que contribuyen al ensamblaje del ribosoma como *NOP56* (*z*-valor = -2,77, *p*-valor = 0,0056), *BOP1* (*z*-valor = -3,11, *p*-valor = 0,0019), *DCK1* (*z*-valor = -3,40, *p*-valor = 0,0007) y *NPM1* (*z*-valor = -2,43, *p*-valor = 0,0151)[458]. Este último gen, *NPM1*, es además responsable de la regulación de múltiples pasos en la biogénesis de ribosomas, y su pérdida conlleva la reducción de la síntesis proteica y el bloqueo de la proliferación celular[459].

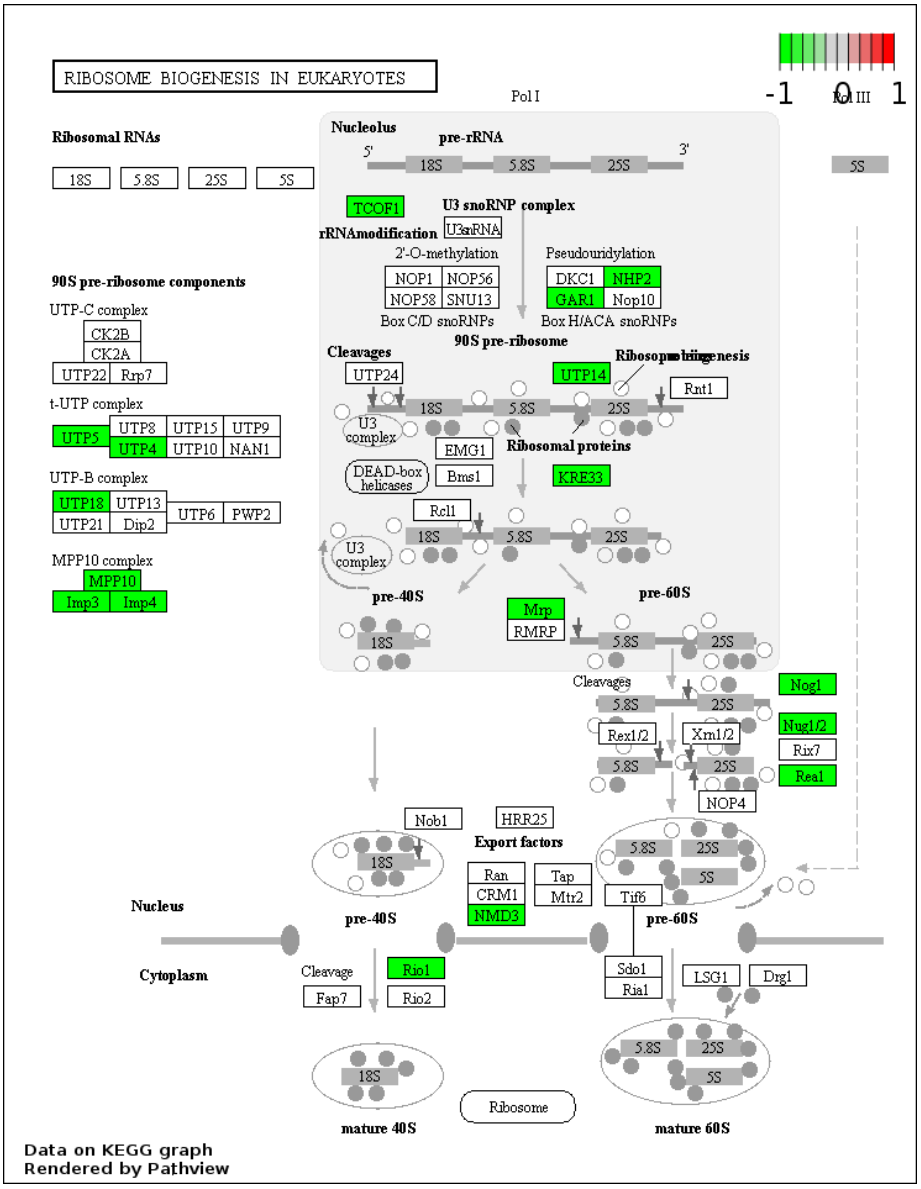



***Figura 4.78.*** *Vía de la biogénesis de ribosomas en eucariotas según la base KEGG. En verde se representan los genes infraexpresados de forma estadísticamente significativa en el metaanálisis global de la lenalidomida.*

Otra de las vías KEGG sobrerrepresentadas de manera estadísticamente significativa fue la de "presentación y procesamiento de antígenos" (FDR = 0,0005) (**Figura 4.79**). De los 77 genes que integran de forma canónica esta vía, 16 se encontraron desregulados en nuestro análisis, de los que 14 presentaron sobreexpresión y dos infraexpresión después del tratamiento con lenalidomida. Entre los genes sobreexpresados, se encontraron genes relacionados tanto con el complejo mayor de histocompatibilidad (MHC) tipo I, como con el MHC tipo II, lo que sugiere una potenciación de ambos complejos proteicos por parte de la lenalidomida. Se ha descrito que los tumores tratan de evadir la respuesta inmune adaptativa mediante la infraexpresión de MHC tipo I en la superficie celular[460]. Un mecanismo similar ocurre MHC tipo II, complejo esencial para la presentación de antígenos que producen la activación de los linfocitos T CD4+ y por consiguiente la respuesta antitumoral, y que se ha comprobado su infraexpresión en células mielomatosas[461]. Por tanto, la sobreexpresión de ambos complejos por la lenalidomida sería uno de los mecanismos que conduciría a la modulación de la respuesta inmune. El tratamiento con lenalidomida conduciría a la activación y proliferación de células T, lo que llevaría a la activación de las células NK produciendo la muerte de las células mielomatosas[462].

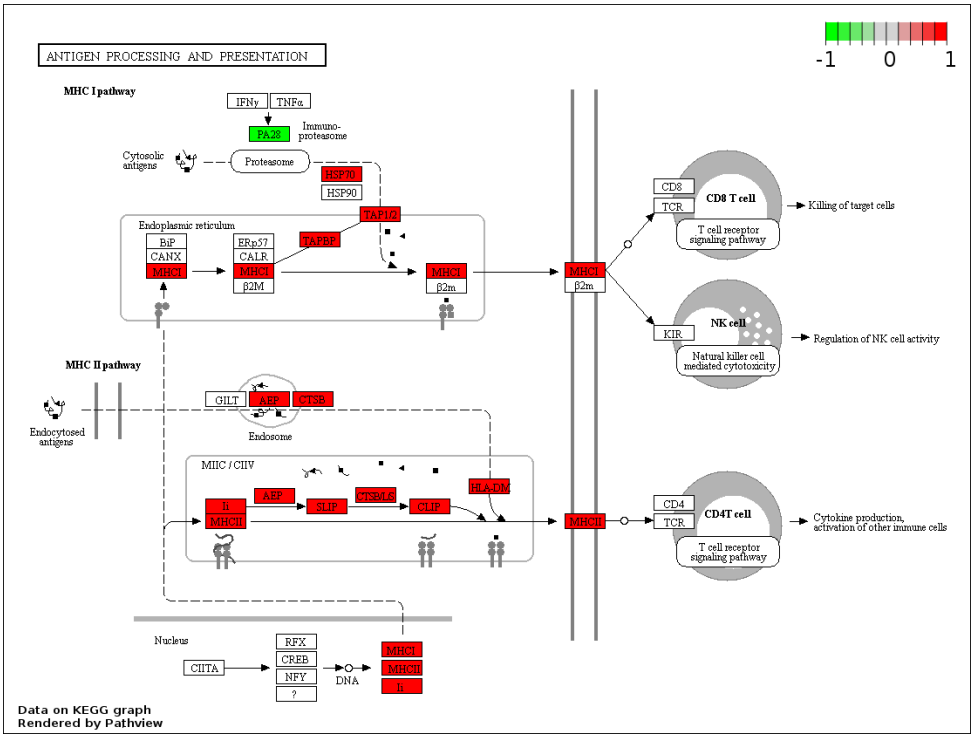



***Figura 4.79.*** *Vía de la presentación y procesamiento de antígenos según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global de la lenalidomida.*

### 4.3.5. Pomalidomida

La pomalidomida es el segundo de los agentes inmunomoduladores para el que se realizó una revisión sistemática con metaanálisis. Para ello, se realizó una búsqueda sistemática de estudios de HMCLs tratados con pomalidomida en los repositorios *online*. El resultado de esta búsqueda fue la detección de seis series en GEO, cinco series en ArrayExpress, y, en cuanto al repositorio SRA, no se detectó ninguna serie tratada con pomalidomida. Del total de 11 series detectadas, solamente seis fueron seleccionadas para su revisión detallada tras eliminar los elementos duplicados. Únicamente una de las seis series cumplió con los criterios de inclusión y exclusión necesarios para ser introducida en el metaanálisis. Finalmente, fueron añadidas dos series adicionales disponibles en el laboratorio de Hematología de Salamanca. Se comprobó la posible subdivisión de las tres series en diferentes estudios en función de las concentraciones de fármaco empleadas, el tiempo de tratamiento o la utilización de varias líneas celulares, no detectándose ningún tipo de división. Por tanto, el número final de estudios considerados para el metaanálisis fue de tres. El diagrama de flujo que se muestra en la **Figura 4.80**, detalla el esquema de selección de estudios para el metaanálisis de pomalidomida en función de los diferentes criterios de inclusión y exclusión.

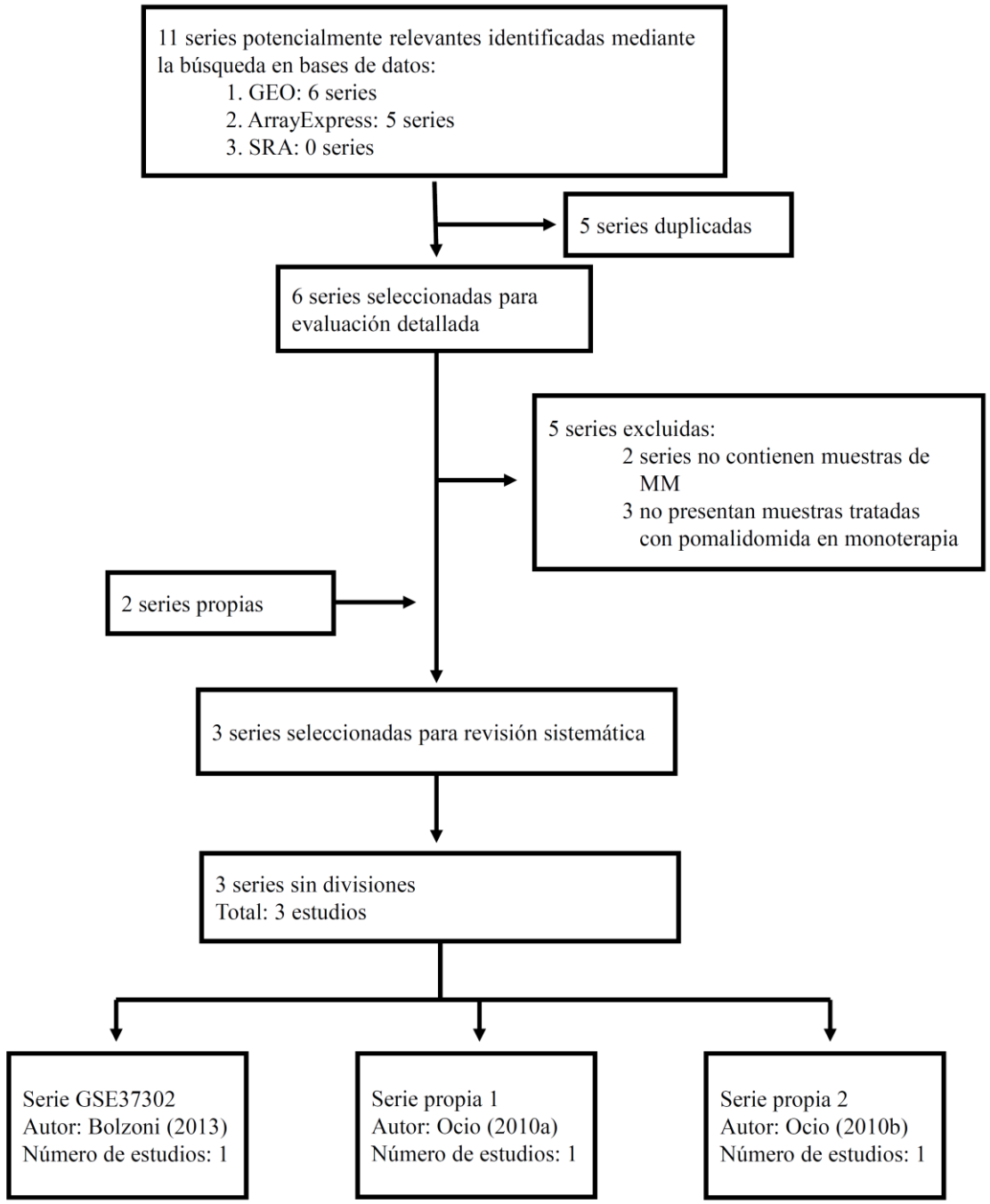



***Figura 4.80.*** *Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con pomalidomida.*

Para la determinación de los subgrupos en función de las medianas y la MAD de los tiempos de tratamiento y de la concentración aplicada se estableció un único punto de corte tanto para para el tiempo de tratamiento (24 horas) como para la concentración (1 µM). Los resultados del agrupamiento en subgrupos se recogen en la **Tabla 4.8**.

***Tabla 4.8.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con pomalidomida.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (uM) |
|---|---|---|---|---|---|---|
| GSE37302 | Bolzoni (2013)[97] | JJN-3 | Affymetrix Human Genome U133 Plus 2.0 Array | 6 | 24 | 100 |
| Salamanca | Ocio (2010a) | MM1-S | Affymetrix Human Gene 1.0 ST Array | 4 | 24 | 1 |
| Salamanca | Ocio (2010b) | MM1-S | Affymetrix Human Gene 1.0 ST Array | 4 | 120 | 1 |

*En amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedias; en rojo, estudios seleccionados para el subgrupo de tiempos o concentraciones altas.*

En cada uno de estos tres estudios se seleccionaron los genes con un valor absoluto del FC > 1,5. Finalmente se seleccionaron 212 genes para realizar el metaanálisis. La distribución de los ln(FC) de estos 212 genes se muestra en la **Figura 4.81**, donde puede observarse una ligera tendencia a la sobreexpresión (ln[FC] > 0) en todos los estudios, acentuada en los estudios Ocio (2010a) y Ocio (2010b).

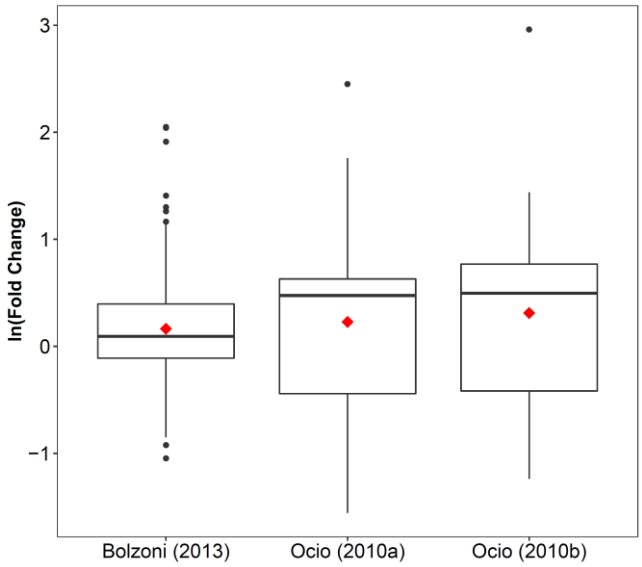



***Figura 4.81.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC]) de los 212 genes seleccionados para el metaanálisis de pomalidomida en líneas celulares de mieloma múltiple. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.5.1. Metaanálisis por subgrupos: tiempo de tratamiento

Al ser únicamente tres los estudios seleccionados para este metaanálisis, se estableció un único punto de corte a las 24 horas. No hubo estudios con un tiempo de tratamiento inferior a las 24 horas, por lo que el análisis solo se llevó a cabo en dos subgrupos. El subgrupo 1 (G1) comprendió los estudios realizados en un tiempo inferior a 24 horas: Bolzoni (2013) y Ocio (2010a). En el segundo subgrupo (G2) incluyó el único

estudio llevado a cabo a un tiempo superior a las 24 horas: Ocio (2010b). El análisis de la expresión génica diferencial mediante metaanálisis por subgrupos de tiempo detectó 123 genes estadísticamente significativos a *p*-valor < 0,05 en el subgrupo G1 y 204 genes en el subgrupo G2 (**Anexo 21**). Se encontraron 116 genes comunes a los dos subgrupos, de los que 82 presentaron sobreexpresión y 34 infraexpresión al tratar con pomalidomida. En la **Figura 4.82** se muestran, a modo de ejemplo, los dos genes que obtuvieron un mayor y un menor valor mediano del FC en el análisis del tratamiento con pomalidomida.

**a**

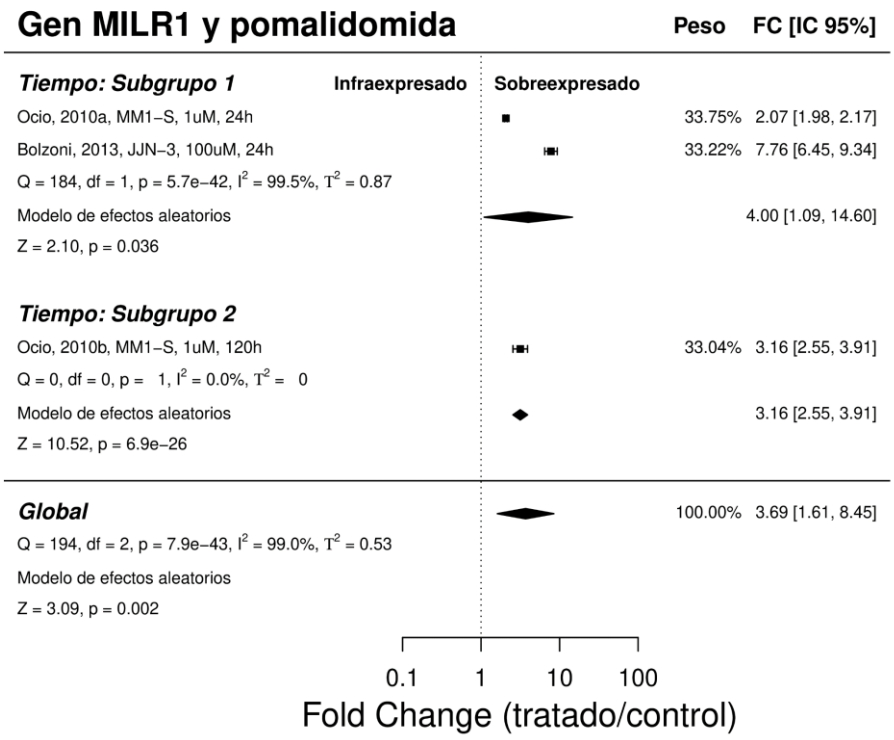



***Figura 4.82.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con pomalidomida.* ***a)*** *Diagrama de bosque del gen MILR1, que fue el más sobreexpresado considerando la mediana del FC de los tres estudios seleccionados.*

**b**

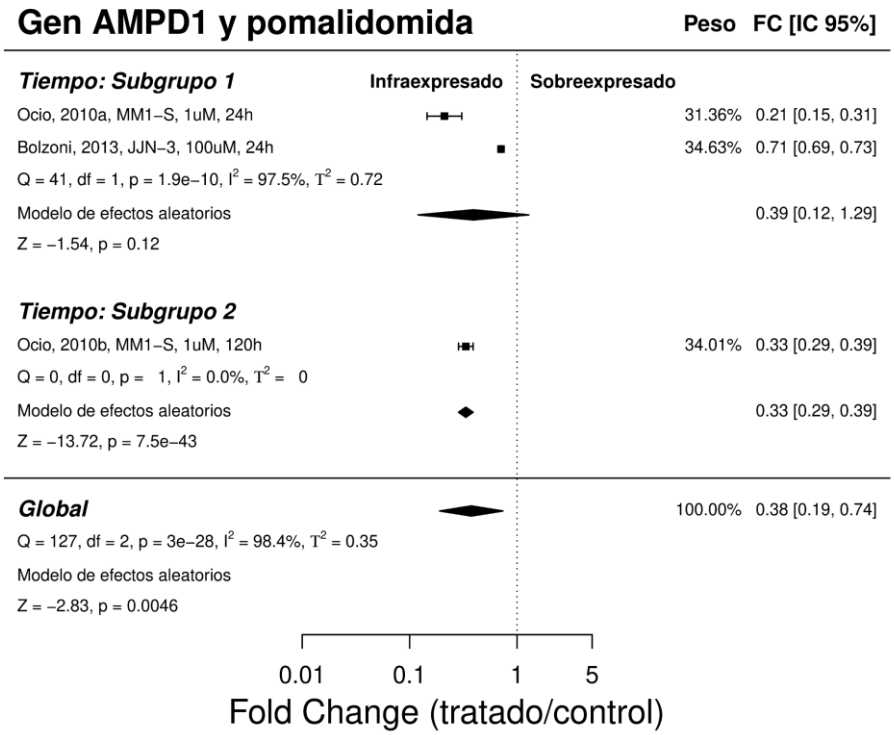



***Figura 4.82 (continuación).*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con pomalidomida.* ***b)*** *Diagrama de bosque del gen AMPD1, que fue el más infraexpresado considerando la mediana del FC de los tres estudios seleccionados.*

La comparación de los resultados obtenidos para la expresión génica en los dos subgrupos de tiempo de tratamiento, realizada mediante la prueba de Wald, detectó 71 genes que presentaron diferencias estadísticamente significativas a *p*-valor < 0,05 entre ambos subgrupos (**Anexo 21**). Estas diferencias a nivel de FC pueden observarse en la **Figura 4.83**. Todos los genes diferencialmente expresados en los dos subgrupos presentaron diferencias en cuanto a la cuantía del cambio en el mismo sentido de la expresión en los dos grupos, y no hubo genes que presentasen un cambio de sentido.

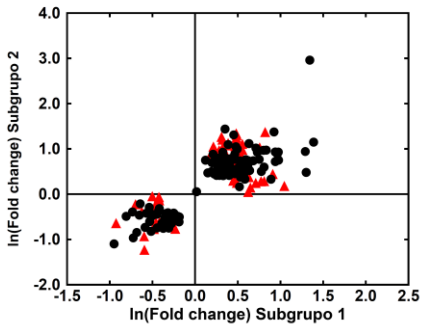



***Figura 4.83.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 212 genes estudiados donde se comparan los subgrupos 1 y 2 del metaanálisis por subgrupos de tiempo de tratamiento con pomalidomida. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre ambos subgrupos.*

Sobre estos 71 genes diferencialmente expresados se realizó el análisis ORA de genes en vías biológicas KEGG y términos GO, con el fin de determinar qué procesos podrían estar afectados por el tiempo de tratamiento. Este análisis aparece recogido en la **Figura 4.84**.

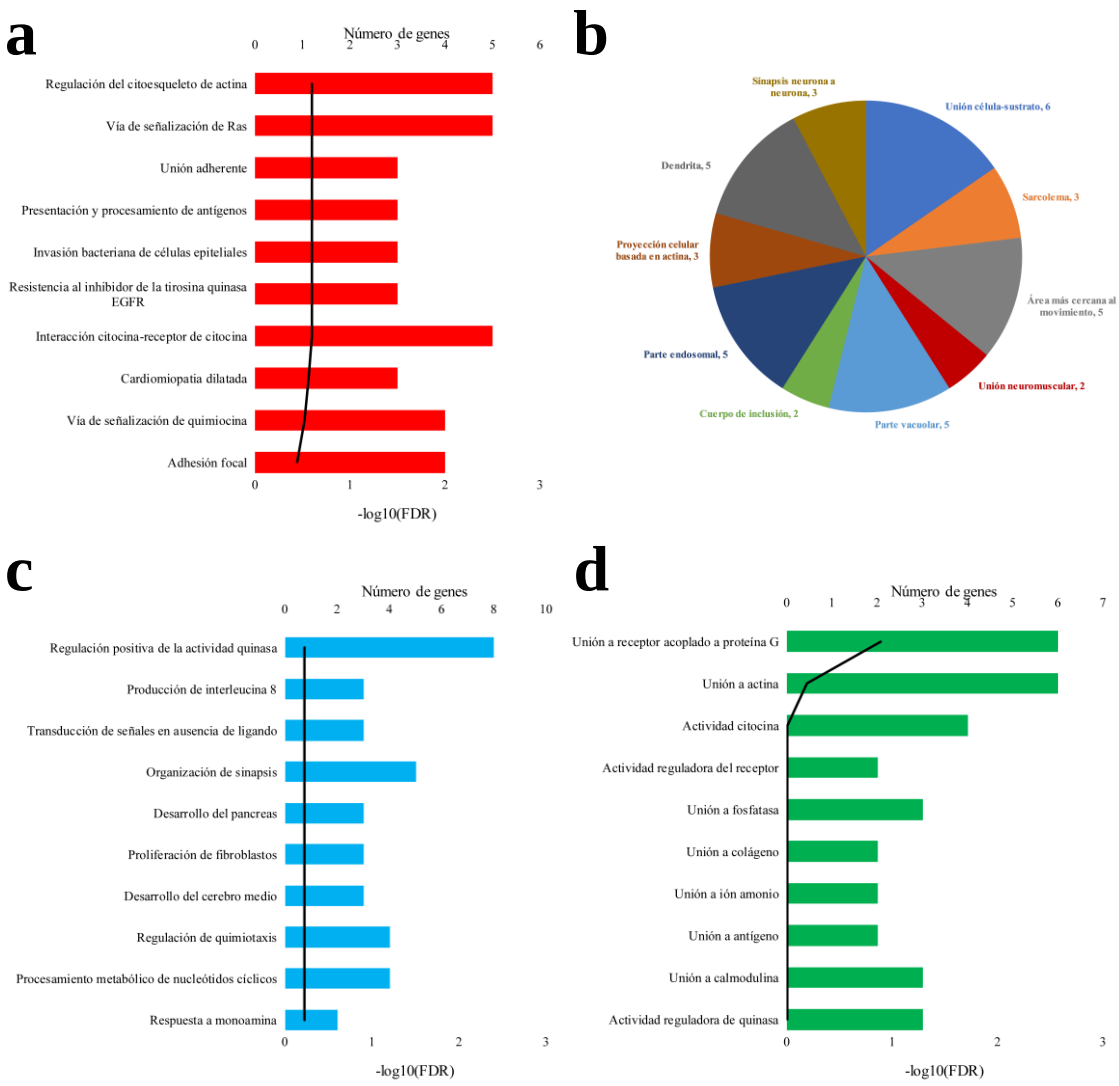



***Figura 4.84.*** *Análisis de sobrerrepresentación de los 71 genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con pomalidomida. En cada panel se recogen las 10 rutas KEGG o los 10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Ninguna vía KEGG o término GO resultó estadísticamente significativo considerando un FDR < 0,05. Por este motivo se procedió al análisis de la influencia del tiempo de tratamiento sobre la expresión de los genes, aunque para la selección de los genes que fueron analizados se utilizó como apoyo el resultado del análisis ORA. Así, se procedió al análisis de los genes recogidos la vía KEGG "regulación del citoesqueleto de actina", ya que fue la primera en significancia estadística, aunque no alcanzó valores estadísticamente significativos (FDR = 0,2535). Hay que destacar que, a pesar de no

alcanzar valores significativos, esta vía representa uno de los mecanismos de acción primarios de la pomalidomida[463]. De este modo, se detectaron cinco genes desregulados pertenecientes a esta vía, cuatro de ellos sobreexpresados y uno infraexpresado. En el caso de los genes sobreexpresados se observaron valores del FC mayores a tiempos de tratamiento más prolongados, mientras que en el caso del gen infraexpresado se observó que la pomalidomida tendría menor efecto a tiempos de tratamiento mayores (**Figura 4.85a**). Para confirmar esta tendencia, se seleccionaron los genes de la vía KEGG o término GO con un mayor número de genes desregulados, que en este caso fue la "regulación positiva de la actividad quinasa" (FDR = 0,5976), con 8 genes, ratificando los resultados obtenidos anteriormente (**Figura 4.85b**). En cualquier caso, estos resultados deben ser tratados con cautela al disponer ambos subgrupos de un número bajo de estudios.

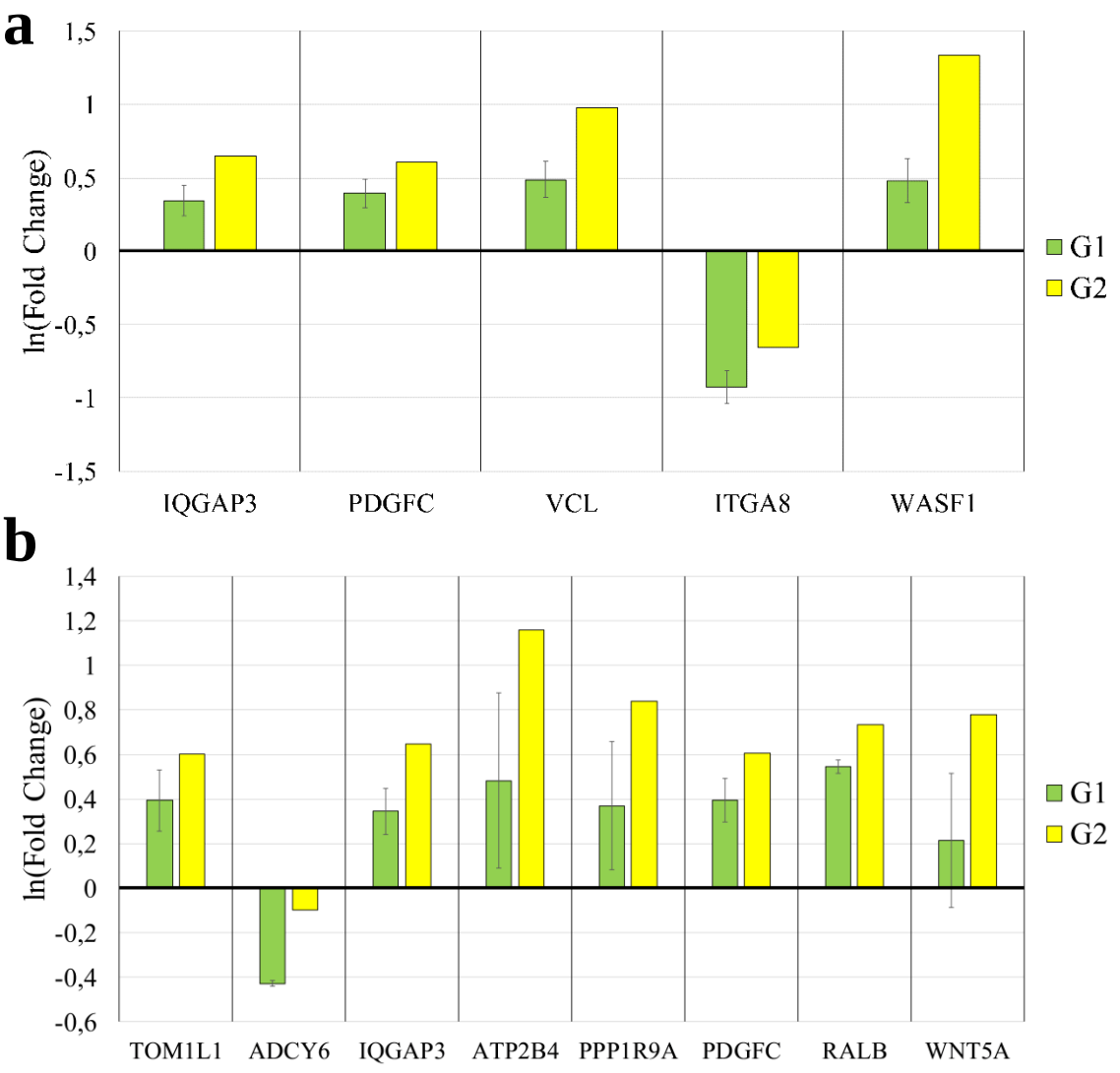



***Figura 4.85.*** *Valores promedio del ln(Fold Change) de los genes desregulados en* ***a)*** *la ruta KEGG "vía de regulación del citoesqueleto de actina" y* ***b)*** *la función GO "regulación positiva de la actividad quinasa", en los dos subgrupos de tiempo de tratamiento con pomalidomida (G1 y G2). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.5.2. Metaanálisis por subgrupos: concentración

De manera similar al análisis por subgrupos de tiempo, en el análisis por subgrupos de concentración de pomalidomida aparece un único punto de corte a 1 µM debido a la inclusión de únicamente tres estudios. Ninguno de los estudios presentó concentraciones inferiores a 1 µM. El primer subgrupo (G1), recogió los estudios a concentraciones iguales a 1 µM: estudios (a) y (b) de Ocio (2010); y el subgrupo 2 (G2) estuvo formado por el estudio de Bolzoni (2013), que fue el único realizado con una concentración superior a 1 µM. Mediante el metaanálisis en estos dos subgrupos fueron determinados 203 genes estadísticamente significativos a *p*-valor < 0,05 en el subgrupo G1 y 165 genes en el caso del G2 (**Anexo 22**). El cruce de estas dos listas determinó que 156 de estos genes fueron comunes a ambas, de los que 101 genes presentaron sobreexpresión y 45 infraexpresión al tratamiento con pomalidomida. Dos ejemplos de genes para este metaanálisis por subgrupos aparecen recogidos en la **Figura 4.86**.

**a**

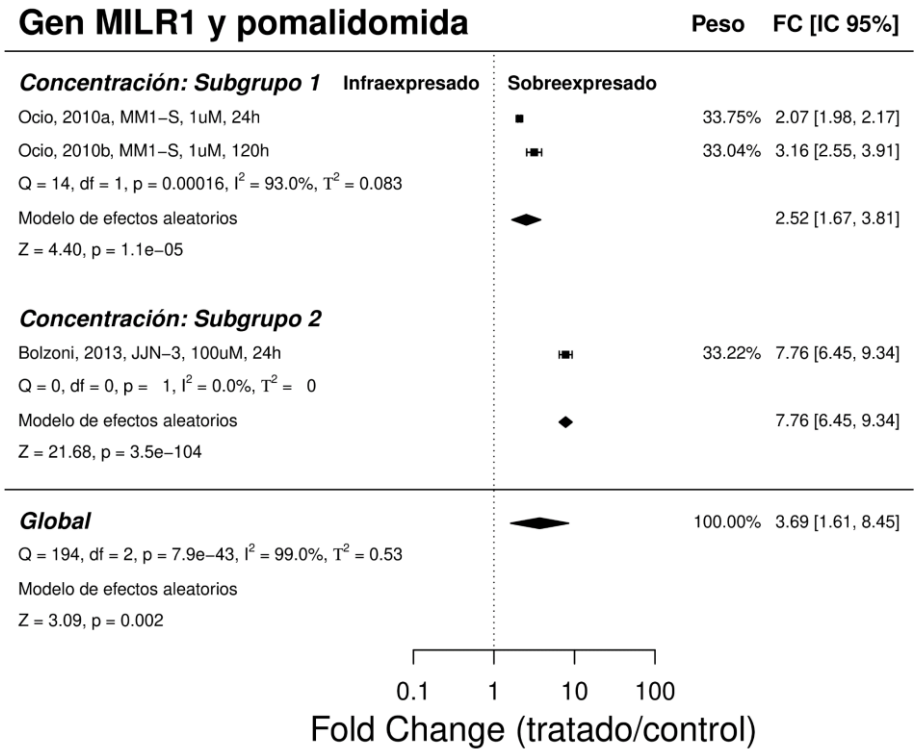



***Figura 4.86**. Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos concentración de pomalidomida. **a)** Diagrama de bosque del gen MILR1, que fue el más sobreexpresado considerando la mediana del FC de los tres estudios seleccionados.*

**b**

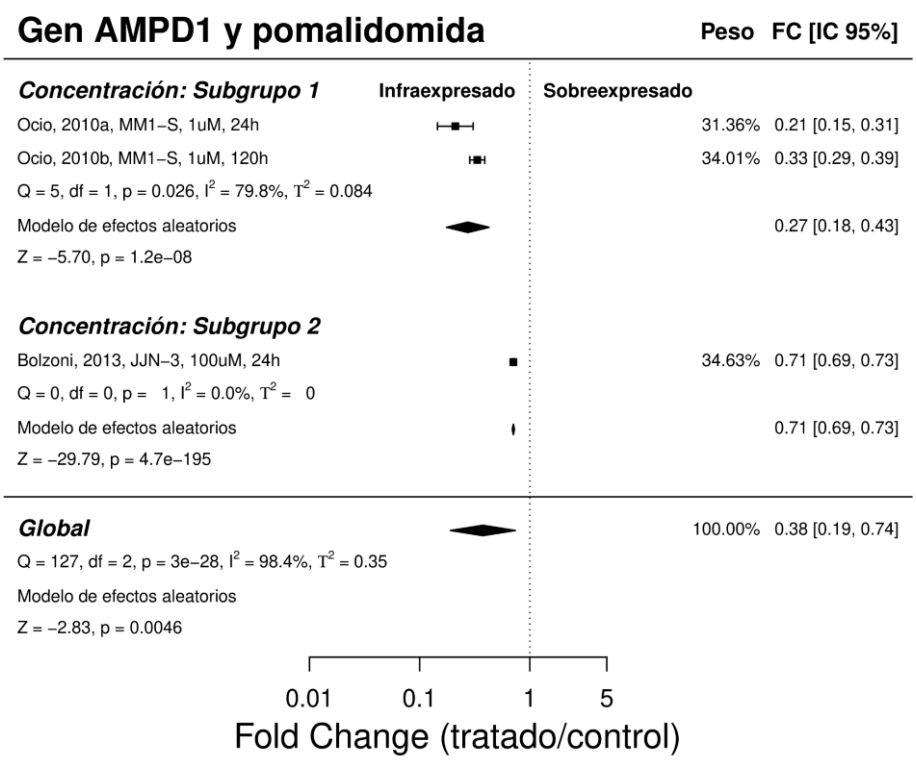



***Figura 4.86 (continuación)**. Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos concentración de pomalidomida. **b)** Diagrama de bosque del gen AMPD1, que fue el más infraexpresado considerando la mediana del FC de los tres estudios seleccionados.*

La comparación de los resultados de los metaanálisis de los dos subgrupos determinó que 161 genes de los 212 genes analizados presentaron diferencias estadísticamente significativas entre ambos subgrupos (**Anexo 22**), lo que supone un porcentaje del 76,1% del total de genes analizado (**Figura 4.87**), implicando que la mayor parte de los genes muestra diferencias de expresión en función de la concentración aplicada de pomalidomida.

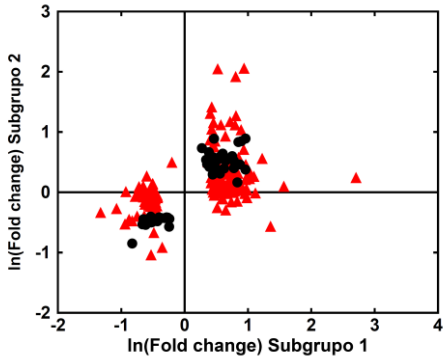



***Figura 4.87.** Diagrama de puntos de los valores de ln(FC) obtenidos para los 212 genes estudiados donde se comparan los subgrupos 1 y 2 del metaanálisis por subgrupos concentración de pomalidomida. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre los subgrupos.*

Esto podría ser indicador de una alta heterogeneidad de los dos grupos dependiente de la concentración de fármaco aplicada, por lo que se trató de elucidar qué vías podrían verse afectadas por este cambio y de qué manera las diferentes concentraciones afectan a la expresión génica. Para ello se procedió al análisis de rutas biológicas KEGG y funciones GO sobre estos genes (**Figura 4.88**).

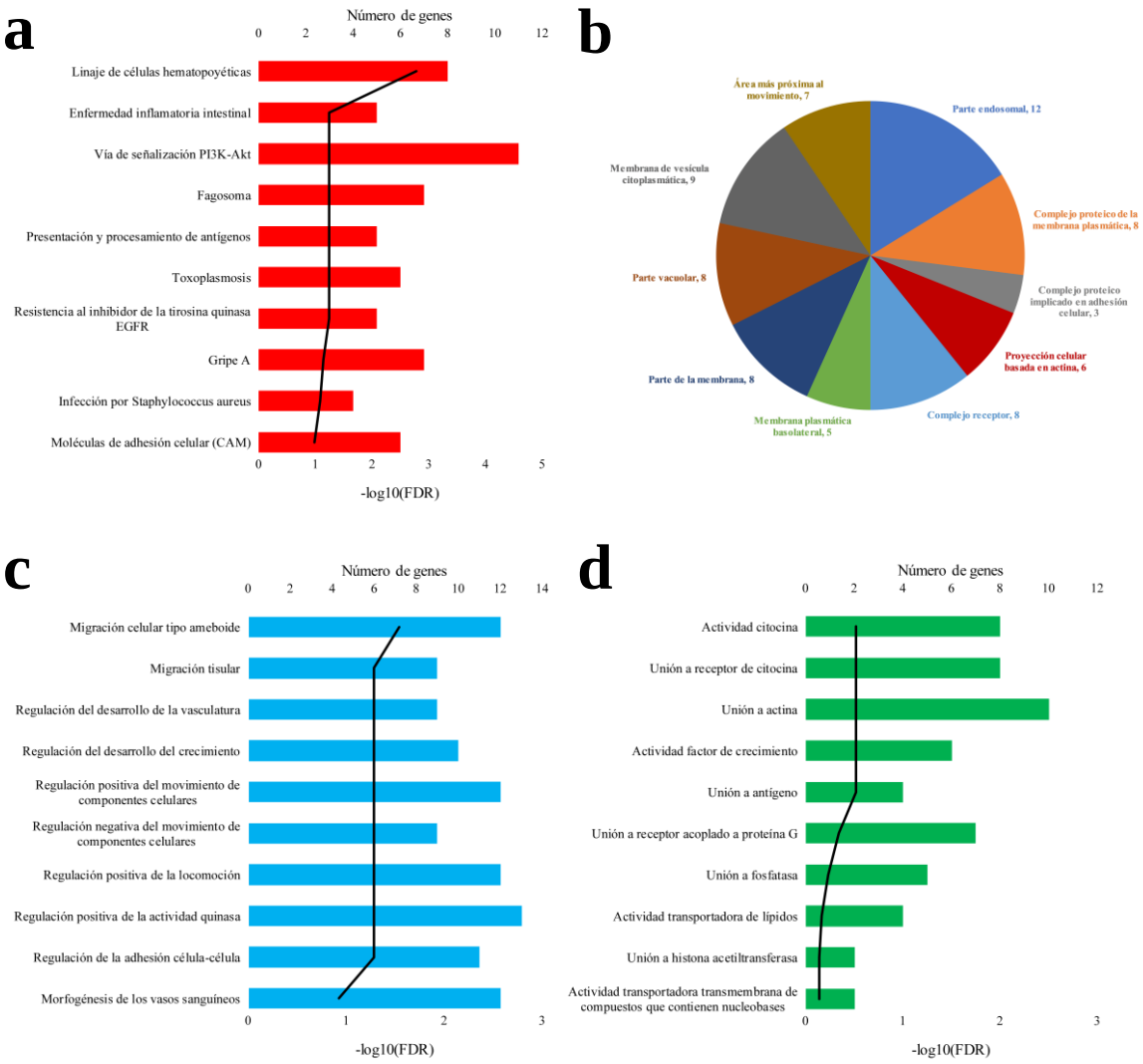



***Figura 4.88.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos concentración de pomalidomida. En cada panel se recogen las 10 rutas KEGG o los 10 términos GO estadísticamente más significativos en función de su FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Entre las vías KEGG, solamente la vía de "linaje de células hematopoyéticas" resultó estadísticamente significativa con un FDR = 0,0017. Esta vía está representada por 8 genes desregulados, de los que seis presentaron mayor incremento del FC en el subgrupo G1. Otros dos genes, *IL7* y *KITLG*, presentaron mayor desregulación a mayores concentraciones (**Figura 4.89a**). El mismo análisis sobre los genes correspondientes al único término GO significativo, esto es, el PB "migración celular tipo ameboide" (FDR

= 0,0288), obtuvo resultados similares (**Figura 4.89b**). Por tanto, en ningún caso podría establecerse una tendencia general a que uno u otro grupo de concentración establezca un patrón concreto de expresión génica, pudiendo ser un evento particular de cada gen o un evento totalmente aleatorio. En cualquier caso, este resultado no podría ser concluyente debido al bajo número de estudios en ambos subgrupos.

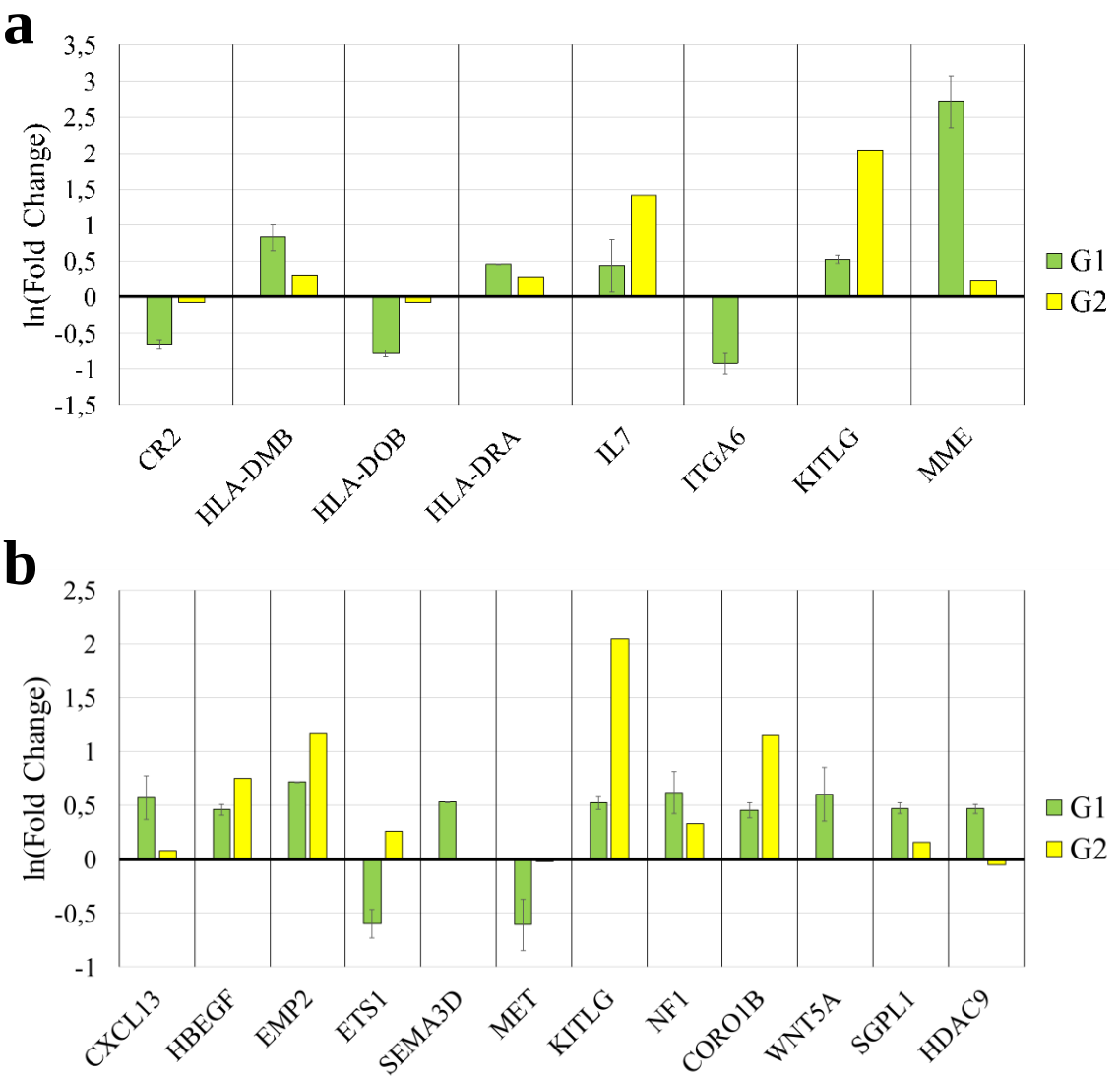



***Figura 4.89.*** *Valores promedio del ln(Fold Change) de los genes desregulados en* ***a)*** *la vía KEGG del "linaje de células hematopoyéticas" y* ***b)*** *el término GO de "migración celular tipo ameboide", en los dos subgrupos de concentración de pomalidomida (G1 y G2). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.5.3. Metaanálisis global de la pomalidomida

El análisis global para la pomalidomida considerando los tres estudios seleccionados en conjunto, reveló un tamaño del efecto estadísticamente significativo a *p*-valor < 0,05 en la expresión de 165 genes, de los que 110 genes presentaron sobreexpresión y 55 genes infraexpresión al tratamiento con pomalidomida (**Anexo 23**).

En comparación con la lenalidomida, que es el otro IMiD estudiado en este trabajo, la pomalidomida indujo una cantidad muy inferior de cambios en los niveles de expresión de los genes analizados (948 genes desregulados por la lenalidomida, frente a los 165 genes de la pomalidomida). Sin embargo, a pesar de estas disimilitudes, el análisis ORA en rutas biológicas y funciones GO sobre los 165 genes (**Figura 4.90**) reveló vías comunes desreguladas por ambos fármacos.

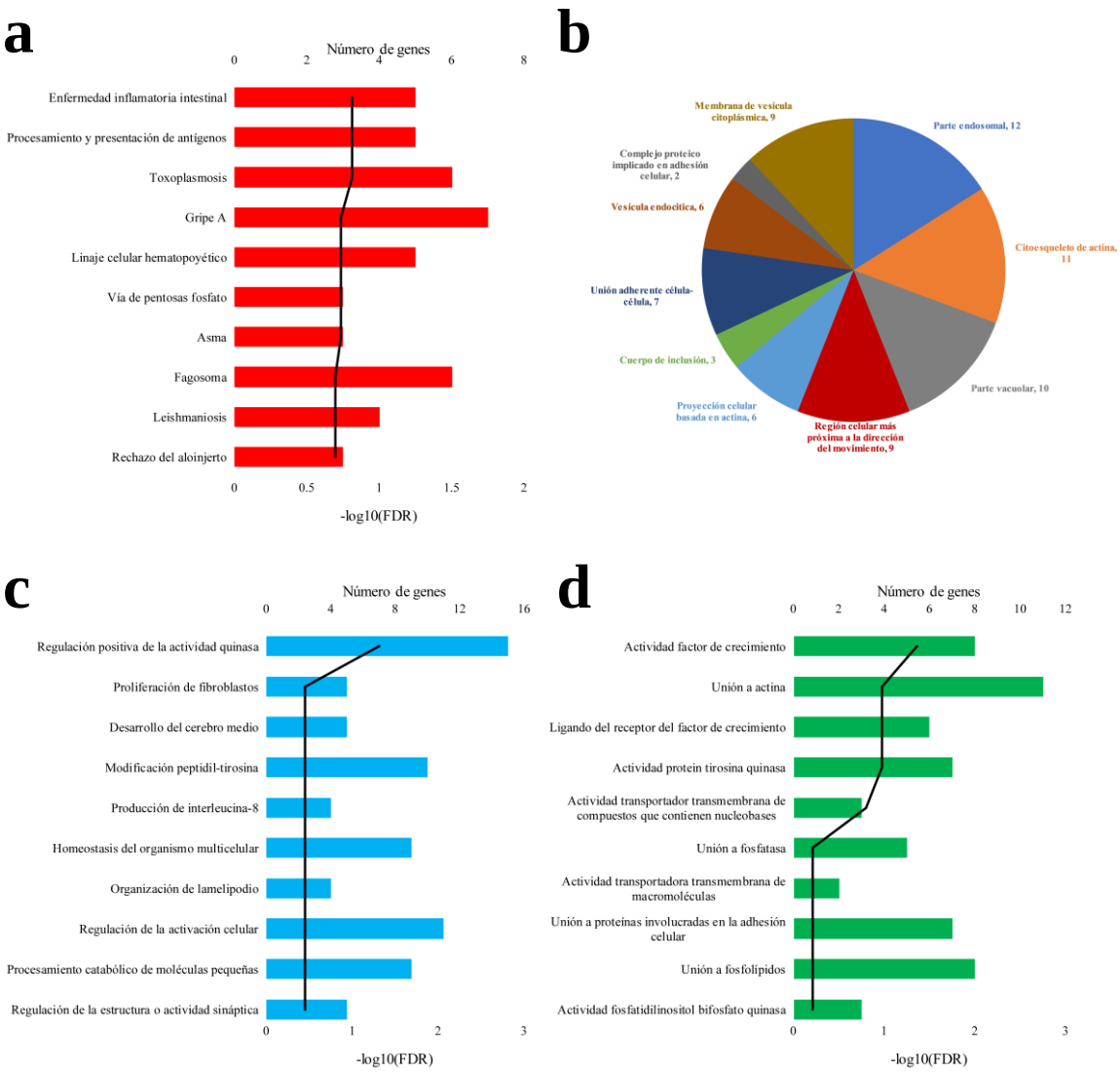



***Figura 4.90.*** *Análisis de sobrerrepresentación en vías biológicas KEGG y términos GO considerando los 165 genes estadísticamente significativos en estudio mediante metaanálisis de la expresión génica en líneas celulares tratadas con pomalidomida. En cada panel se recogen las 10 rutas KEGG o los 10 términos GO estadísticamente más significativos en función del FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

De esta manera, aunque no hubo vías KEGG significativas a FDR < 0,05, hay que destacar que una de las vías con mayor sobrerrepresentación fue, al igual que en el caso de la lenalidomida, la vía de "procesamiento y presentación de antígenos" (FDR = 0,1539) (**Figura 4.91**). En general, se observa que la pomalidomida actúa

sobreexpresando componentes del MHC tipo II, además del componente de la vía del MHC de tipo I *HSPA1A* (*z*-valor = 2,15, *p*-valor = 0,0318), miembro de la familia de proteínas Hspa70. Por tanto, el mecanismo de modulación de la respuesta inmune sería muy similar al llevado a cabo por la lenalidomida, actuando en el caso de la lenalidomida sobre un mayor número de componentes tanto del MHC tipo I como del MHC tipo II.

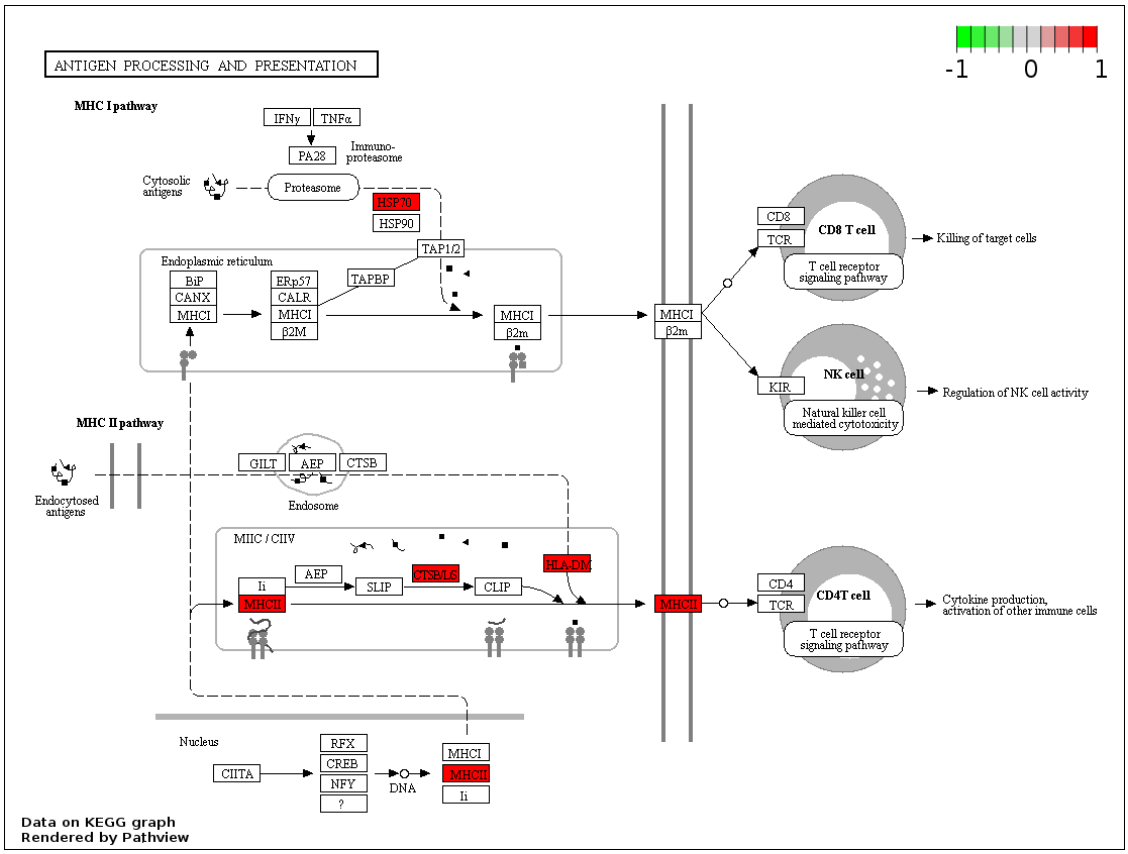



***Figura 4.91.*** *Vía de la presentación y procesamiento de antígenos según la base KEGG. En rojo se representan los genes infraexpresados de forma estadísticamente significativa en el metaanálisis global de la pomalidomida.*

Por otro lado, la pomalidomida podría además estar modulando la proliferación y diferenciación de ciertos tipos celulares a través de su acción sobre el "linaje hematopoyético" (FDR = 0,183) (**Figura 4.92**). Esta acción la llevaría a cabo mediante la sobreexpresión de la interleucina 7 (*IL7*, *z*-valor = 2,14, *p*-valor = 0,0323), que es una molécula que estimula la proliferación de las células del linaje linfoide como las células B y T, y las células NK[464]. La inducción de estos tipos celulares por la pomalidomida conduciría a una respuesta inmune antimieloma[465], que produciría la lisis de las células cancerígenas por mecanismos como los mediados por las células NK.

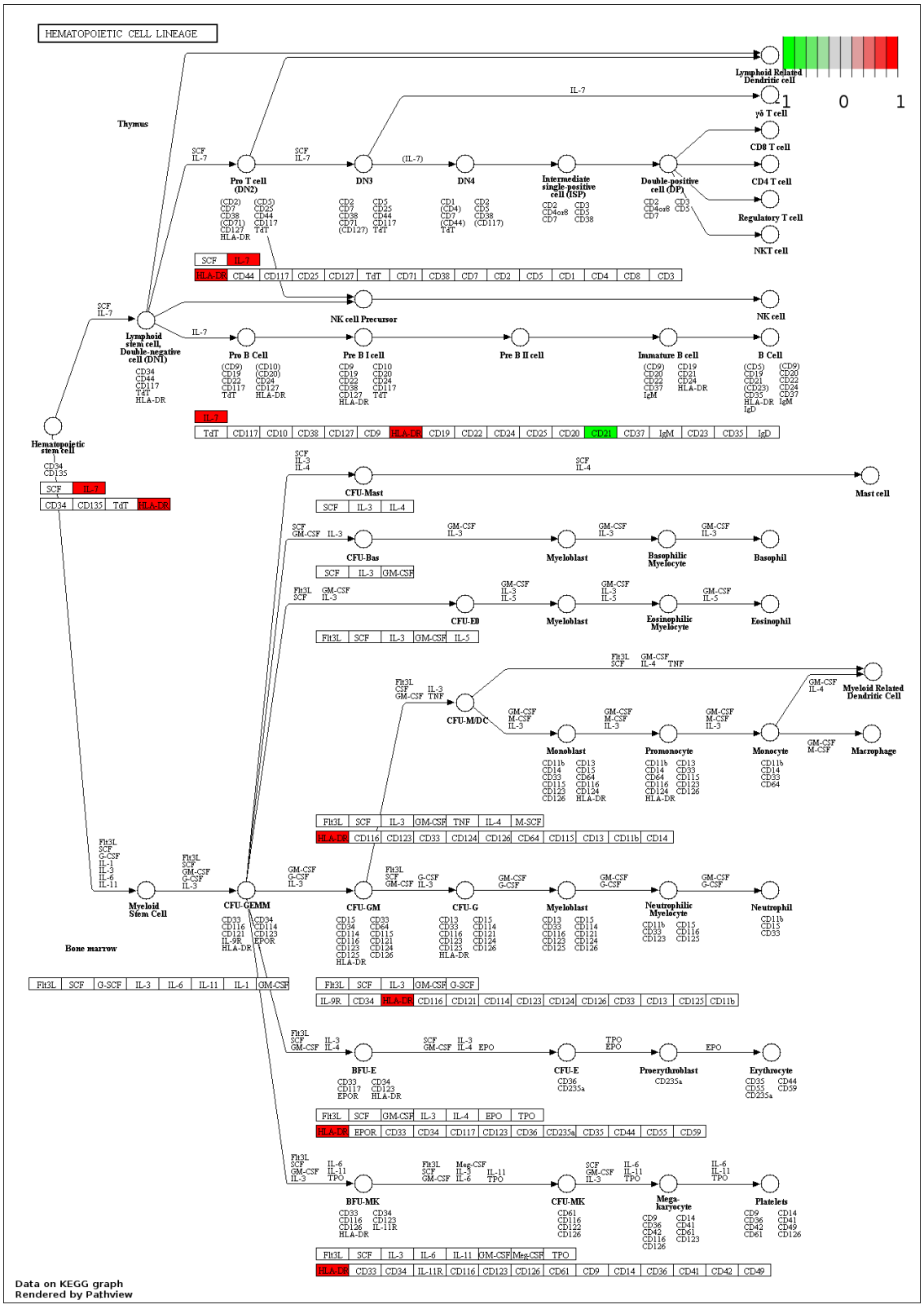



***Figura 4.92.*** *Linaje hematopoyético según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global de la pomalidomida.*

### 4.3.6. Panobinostat

Se realizó en un primer paso una búsqueda sistemática de estudios llevados a cabo en líneas celulares de MM tratadas con panobinostat depositados en repositorios online de datos genómicos. Esta búsqueda obtuvo 30 series en GEO, 23 series en ArrayExpress y 72 muestras en SRA correspondientes a dos series. Treinta y cinco series fueron seleccionadas para su revisión detallada tras ser eliminados los elementos duplicados en los tres repositorios. Dos de las 35 series cumplieron los criterios de inclusión y exclusión necesarios para ser incluidas en el metaanálisis. Se añadió además una serie adicional que se disponía en el laboratorio de Hematología de Salamanca. Dos de las series seleccionadas, GSE30524 y GSE56623, fueron divididas en dos y cuatro estudios, respectivamente. El motivo de la subdivisión fue que en ambas series se aplicó una misma concentración de panobinostat a tiempos diferentes. En el caso de la serie GSE56623, se realizó una segunda subdivisión ya que se aplicaron estos regímenes de tratamiento sobre dos líneas celulares diferentes. Como resultado, el número final de estudios considerados para el metaanálisis fue de 7. El diagrama de flujo que se muestra en la **Figura 4.93**, detalla el esquema de selección de estudios para el metaanálisis de la panobinostat en función de los diferentes criterios de inclusión y exclusión.

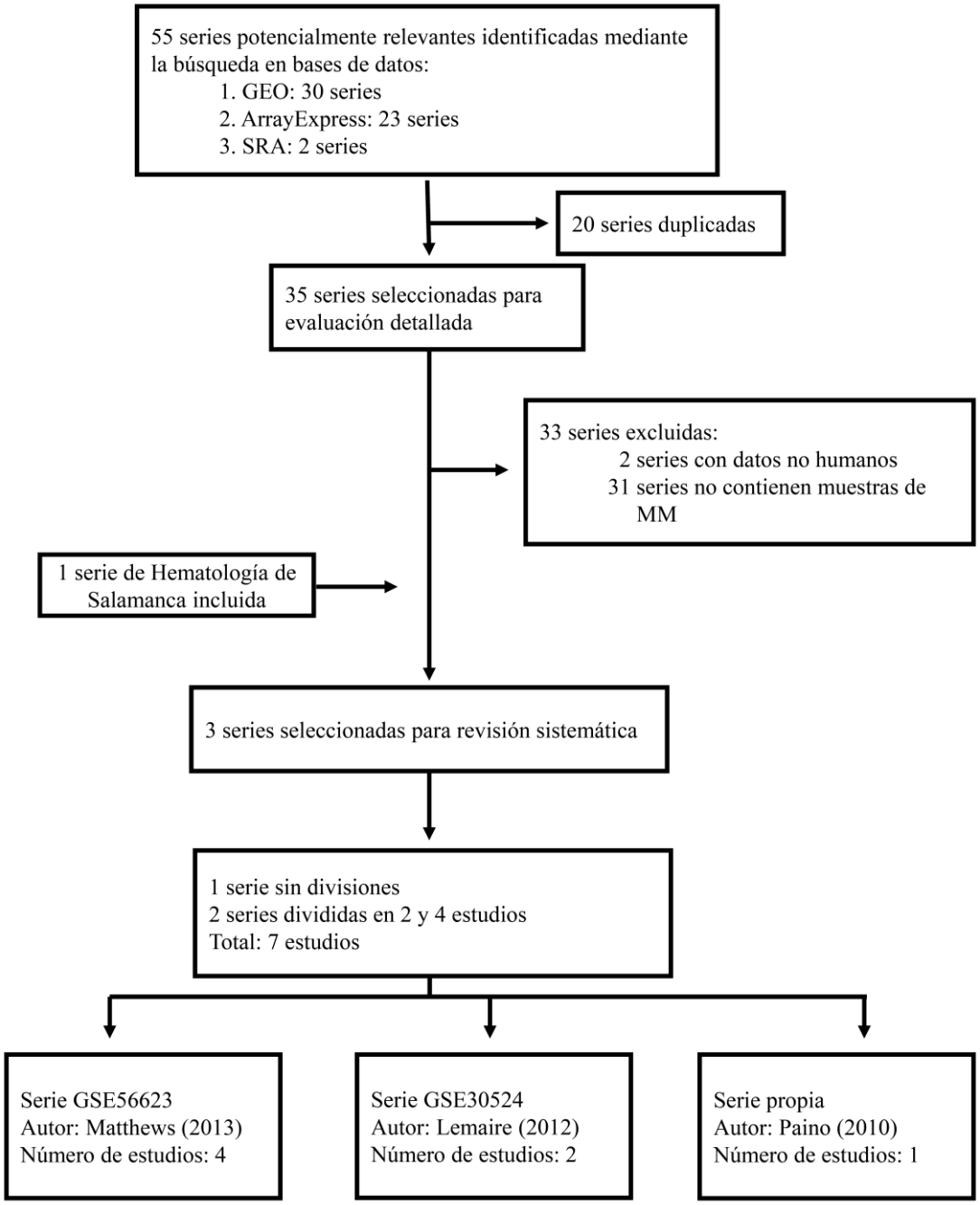



***Figura 4.93**. Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con panobinostat.*

Los 7 estudios seleccionados se clasificaron en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de panobinostat. Se establecieron como puntos de corte los tiempos de tratamiento a seis y a 42 horas (mediana = 24 horas), mientras que para la concentración se estableció un único punto de corte a 10 nM al ser la MAD = 0 (mediana = 10 nM). Los resultados del agrupamiento en subgrupos pueden observarse en la **Tabla 4.9**.

***Tabla 4.9.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con panobinostat.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (nM) |
|---|---|---|---|---|---|---|
| GSE56623 | Matthews (2013a)[466] | JJN-3 | Illumina HiSeq 2500 | 6 | 4 | 10 |
| GSE56623 | Matthews (2013b)[466] | JJN-3 | Illumina HiSeq 2500 | 6 | 24 | 10 |
| GSE56623 | Matthews (2013c)[466] | U266 | Illumina HiSeq 2500 | 6 | 4 | 10 |
| GSE56623 | Matthews (2013d)[466] | U266 | Illumina HiSeq 2500 | 6 | 24 | 10 |
| GSE30524 | Lemaire (2012a)[467] | RPMI-8266 | Affymetrix Human Genome U133 Plus 2.0 Array | 2 | 6 | 20 |
| GSE30524 | Lemaire (2012b)[467] | RPMI-8267 | Affymetrix Human Genome U133 Plus 2.0 Array | 2 | 24 | 20 |
| Salamanca | Paino (2010) | MM1-S | Affymetrix Human Genome U133 Plus 2.0 Array | 4 | 48 | 7 |

*En verde, estudios seleccionados para el subgrupo G1, en amarillo, para el subgrupo G2 y en rojo, para el subgrupo G3, de tiempo de tratamiento o concentración de panobinostat. NA: no aplicable debido a que es una serie del Servicio de Hematología de Salamanca.*

Una vez establecidos los subgrupos de tiempo y concentración fueron determinados los genes candidatos para el metaanálisis. Se consideraron los genes cuyo valor absoluto del FC fue mayor a 1,5 en los tres estudios o, al menos, en todos los estudios de uno de los subgrupos de tiempo o concentración, excluyendo los subgrupos que solamente constasen de un estudio. De esta manera, fueron seleccionados 2.056 genes para realizar su estudio mediante metaanálisis. En la **Figura 4.94**, se muestra la distribución del ln(FC) de estos 2.056 genes, donde puede observarse una mayor dispersión del ln(FC) en los estudios (a-b) de Lemaire (2012) y (a-d) de Matthews (2013), probablemente debido al hecho de que fueron realizados a tiempos más largos de tratamiento. Sin embargo, esta tendencia no se aprecia en el estudio de Paino (2010), quizá porque fue el estudio en el que, a pesar de contar con el mayor tiempo de tratamiento (48 h), se aplicó también la menor concentración del fármaco (7 nM). La influencia sobre la expresión génica de estos dos factores será evaluada en los siguientes apartados mediante metaanálisis por subgrupos.

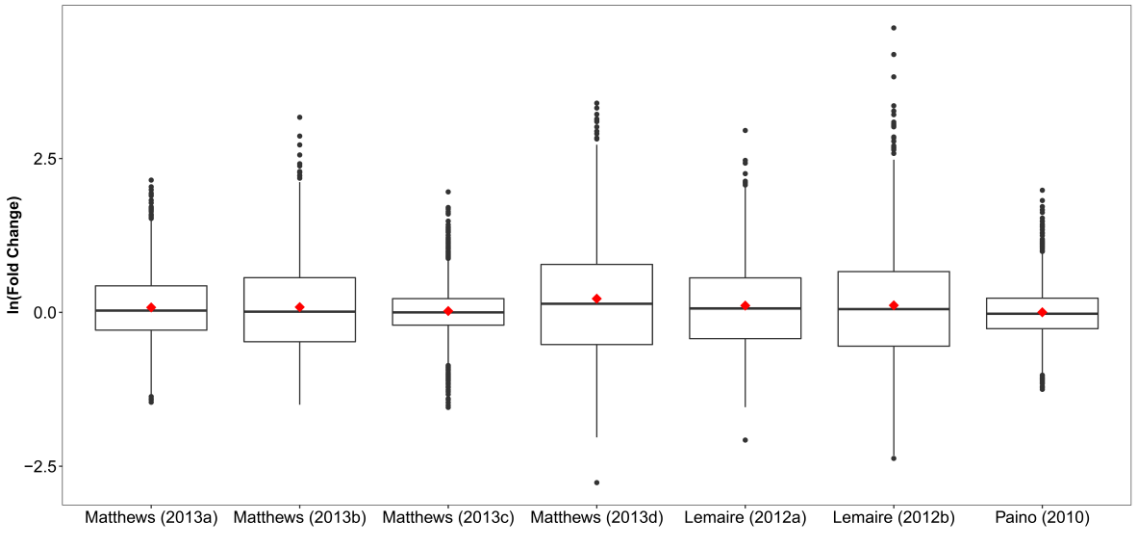
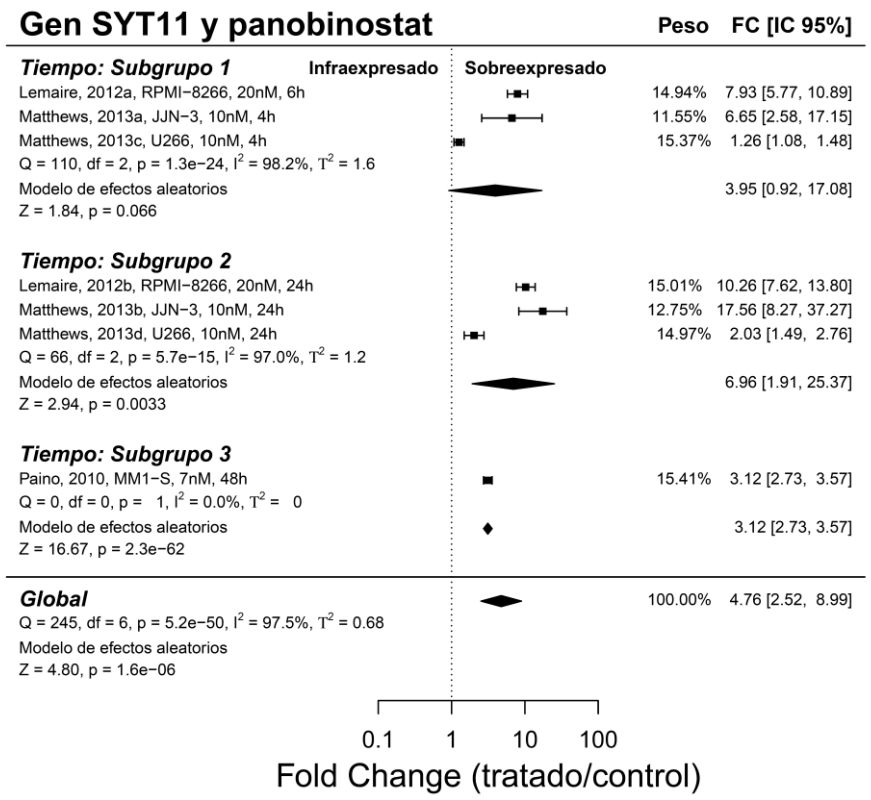



***Figura 4.94.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC]) de los 2.056 genes seleccionados para el metaanálisis de panobinostat en monoterapia en líneas celulares de mieloma múltiple. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.6.1. Metaanálisis por subgrupos: tiempo de tratamiento

Los 7 estudios de panobinostat se clasificaron en tres subgrupos de tiempo de tratamiento. En un primer subgrupo (G1) se recogieron los estudios de Lemaire (2012a) y (a) y (c) de Matthews (2013), que fueron realizados a un tiempo inferior o igual a las 6 horas. El segundo subgrupo (G2) comprendió los estudios con tiempos superiores a las 6 horas, pero inferiores o iguales a las 42 horas, de manera que fueron tres los estudios que cumplieron estos requisitos: Lemaire (2012b) y los estudios (b) y (d) de Matthews (2013). Por último, el tercer subgrupo (G3) estuvo formado por un único estudio cuyo tiempo de tratamiento fue superior a las 42 horas (Paino [2010]). El metaanálisis por subgrupos de tiempo detectó 1.266 genes con diferencias de expresión estadísticamente significativas a *p*-valor < 0,05 en el subgrupo G1, 1.498 genes en el caso del subgrupo G2 y 1.269 genes en el subgrupo G3 (**Anexo 24**). El cruce de las listas de genes diferencialmente expresados reveló 665 genes comunes a los tres subgrupos, de los que 232 presentaron sobreexpresión y 356 infraexpresión al tratar con panobinostat, el resto de los genes presentaron sentidos de expresión opuestos entre alguno de los subgrupos. En la **Figura 4.95** se muestran dos ejemplos de diagramas de bosque de los genes con mayor valor absoluto de la mediana de FC considerando los 7 estudios.

**a**




**b**

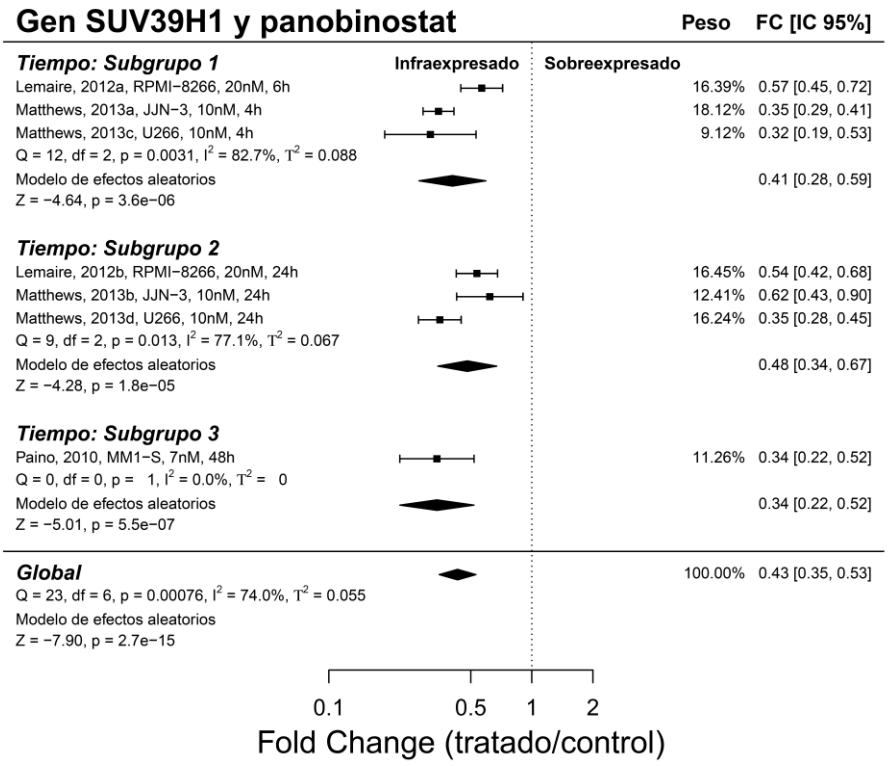



***Figura 4.95.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con panobinostat.* ***a)*** *Diagrama de bosque del gen SYT11, que fue el más sobreexpresado considerando la mediana del FC de los 7 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen SUV39H1, que fue el más infraexpresado considerando la mediana del FC de los 7 estudios seleccionados.*

Se compararon las diferencias de expresión génica entre los tres subgrupos mediante una prueba estadística tipo Wald. Esta prueba detectó que 837 genes presentaron diferencias estadísticamente significativas (*p*-valor < 0,05) entre los subgrupos G1 y G2, 942 genes entre los subgrupos G1 y G3, y 936 genes entre los subgrupos G2 y G3 (**Anexo 24**). En la **Figura 4.96** se muestra mediante un diagrama de puntos las diferencias entre los valores de ln(FC) obtenidos en los tres subgrupos.

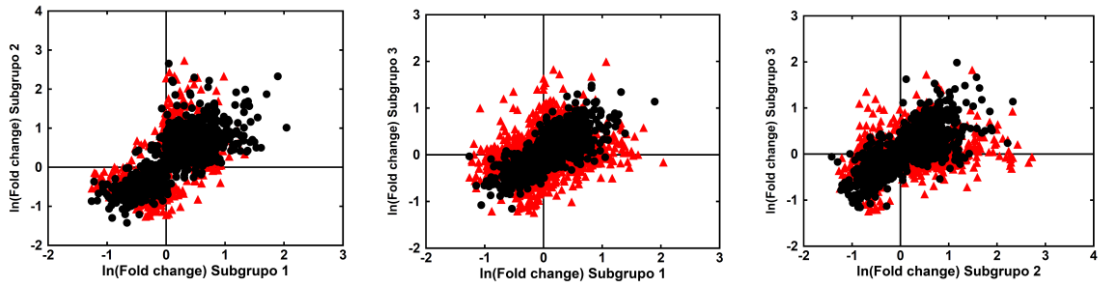



***Figura 4.96.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 2.056 genes estudiados donde se comparan los subgrupos 1, 2 y 3. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

Sobre los genes desregulados entre los tres subgrupos se realizó el análisis de sobrerrepresentación de genes en rutas biológicas KEGG y términos GO. Las 10 vías KEGG y términos GO a los niveles PB, FM y CC se recogen en la **Figura 4.97**.

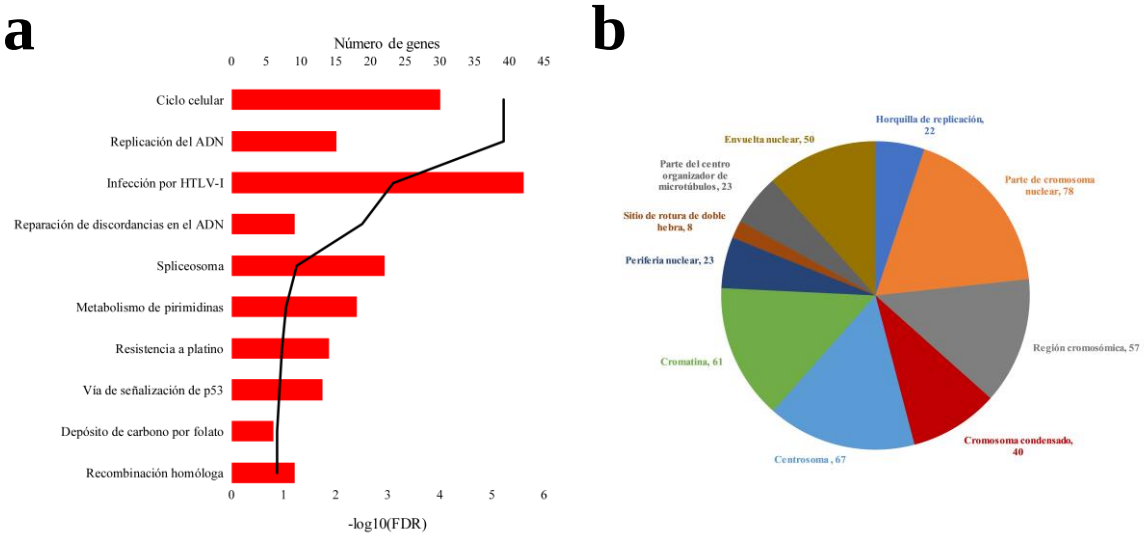



***Figura 4.97.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con panobinostat. En cada panel se recogen las 10 rutas KEGG o los 10 términos GO con mayor significancia estadística en función del FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

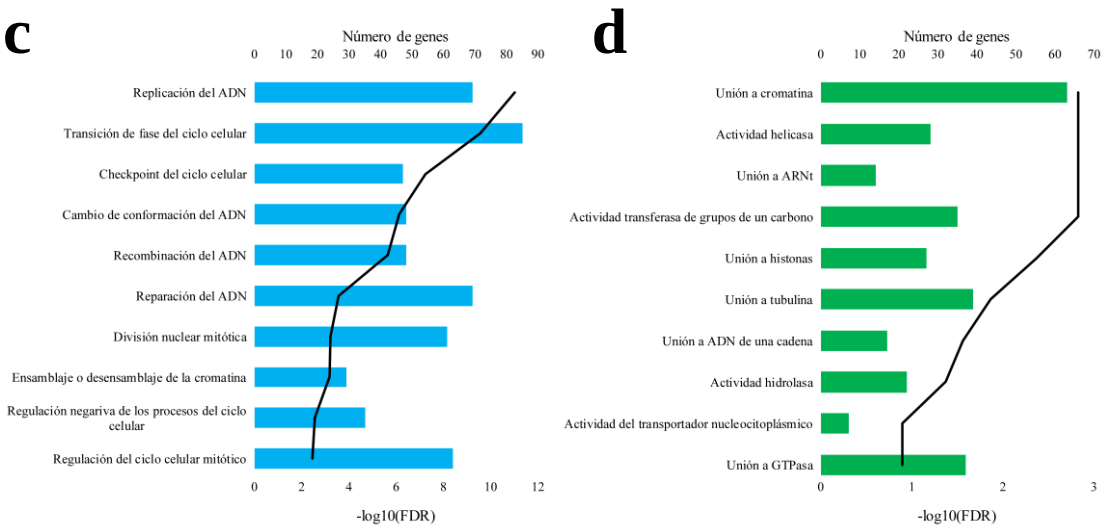



***Figura 4.97 (continuación)**: Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con panobinostat. En cada panel se recogen las 10 rutas KEGG o los 10 términos GO con mayor significancia estadística en función del FDR. **c)** TOP 10 procesos biológicos GO y **d)** TOP 10 funciones moleculares GO.*

El análisis ORA muestra numerosos procesos de gran importancia en el mecanismo de acción del panobinostat, como son el ciclo celular[468], la replicación[469], la recombinación[470] y la reparación del ADN[471], que podrían estar afectados por el tiempo de tratamiento. Para dilucidar la influencia temporal sobre la expresión génica, seleccionamos la vía KEGG con un menor valor de FDR, que fue el "ciclo celular" (FDR < 0,0001), contando con un total de 30 genes desregulados. El estudio de la influencia del tiempo de tratamiento sobre estos 30 genes no arrojó resultados concluyentes, ya que no se detectó ningún patrón de expresión génica dependiente del tiempo, como puede observarse en la **Figura 4.98**. Los cambios de expresión al tratamiento parecieron depender exclusivamente del gen analizado, ya que, por ejemplo, mientras los genes de la familia de inhibidores de ciclinas dependientes de quinasas (*CDKN*) presentaron una mayor desregulación a tiempos de tratamiento bajos, otros genes, como los de la familia del complejo de mantenimiento de minicromosomas (*MCM*), presentaron la mayor desregulación a tiempos medios de tratamiento.

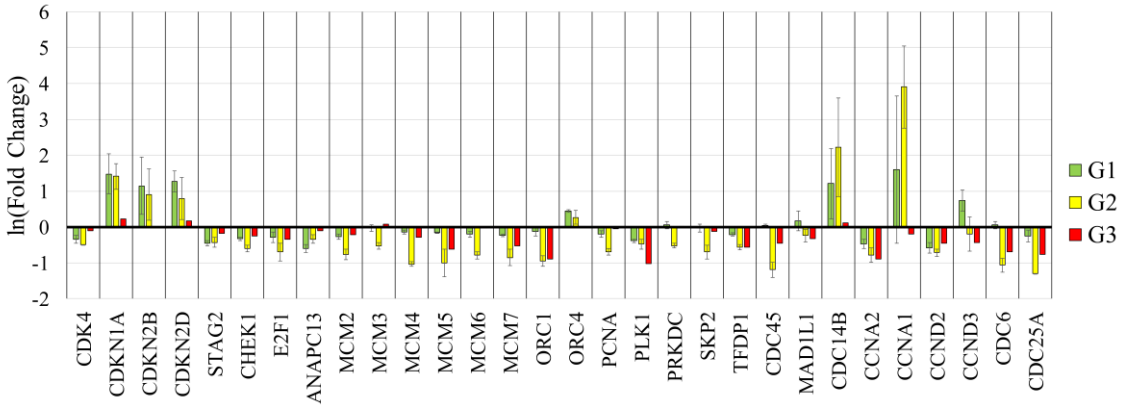
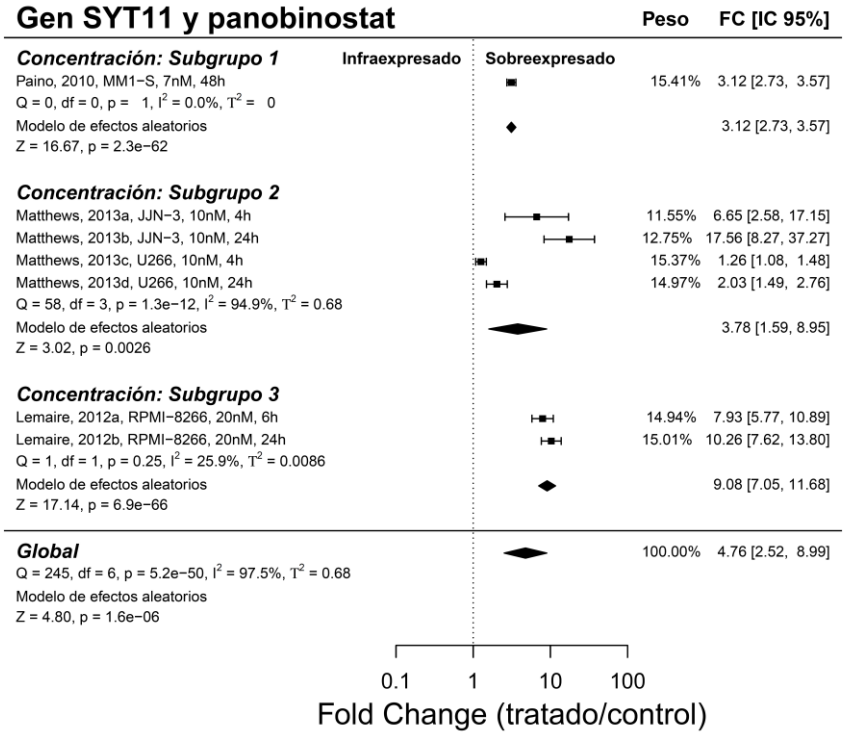



***Figura 4.98.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía del ciclo celular en los tres subgrupos de tiempo de tratamiento con panobinostat (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.6.2. Metaanálisis por subgrupos: concentración

El punto de corte para determinar los subgrupos de concentración de panobinostat fue de 10 nM, de manera que fueron establecidos tres subgrupos considerando los 7 estudios seleccionados. En el primer grupo (G1), que agrupó los estudios a una concentración menor de 10 nM, solamente se incluyó el estudio de Paino (2010). En el segundo subgrupo (G2) se recogieron los cuatro estudios (a-d) de Matthews (2013) cuya concentración de panobinostat fue de 10 nM. Finalmente, en el tercer subgrupo (G3) se encuadraron los dos estudios, (a) y (b) de Lemaire (2012), cuyas concentraciones fueron superiores a 10 nM. Mediante el metaanálisis por subgrupos se determinaron 1.269 genes con expresión diferencial estadísticamente significativa a *p*-valor < 0,05 en el subgrupo G1, 1.400 genes en el caso del subgrupo G2 y 1.430 genes en el subgrupo G3 (**Anexo 25**). El cruce de estas tres listas determinó que 662 de estos genes fueron comunes a los tres subgrupos, de los que 264 presentaron sobreexpresión y 313 infraexpresión al tratar con panobinostat. En la **Figura 4.99** se muestran dos ejemplos de diagrama de bosque para el metaanálisis por subgrupos de concentración de panobinostat.

**a**




**b**

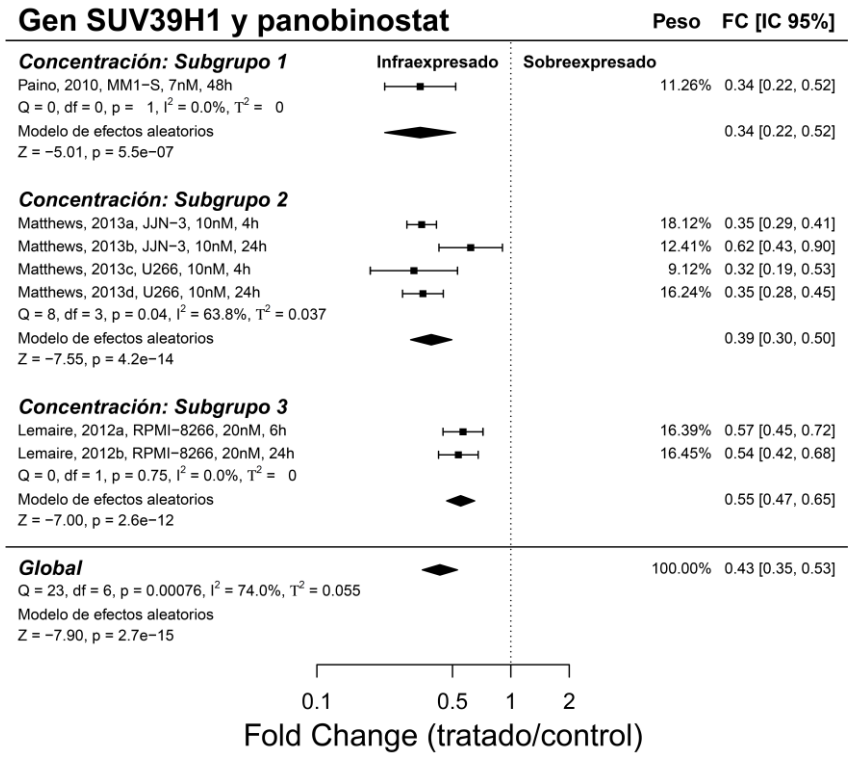



***Figura 4.99.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de panobinostat.* ***a)*** *Diagrama de bosque del gen SYT11, que fue el más sobreexpresado considerando la mediana del FC de los 7 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen SUV39H1, que fue el más infraexpresado considerando la mediana del FC de los 7 estudios seleccionados.*

La comparación de la expresión génica entre los tres subgrupos reveló que 835 genes presentaron diferencias estadísticamente significativas entre los subgrupos G1 y G2, 909 genes entre los subgrupos G1 y G3, y 513 genes entre los subgrupos G2 y G3 (**Anexo 25**). En la **Figura 4.100** pueden observarse los valores de ln(FC) de cada uno de los 2.056 genes estudiados en los tres subgrupos de concentración de bortezomib en una representación mediante diagrama de puntos.

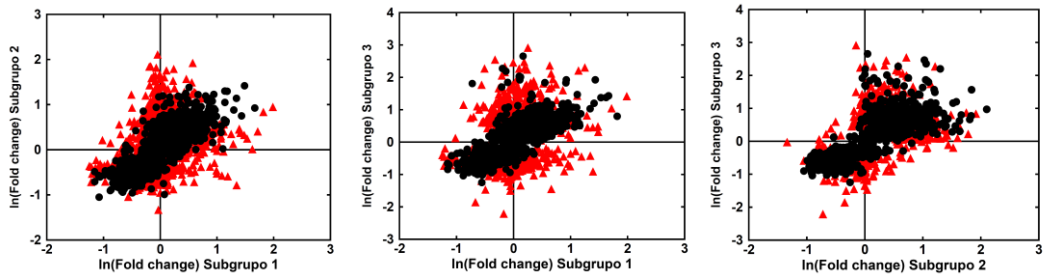



***Figura 4.100.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 2.056 genes estudiados donde se comparan los subgrupos 1, 2 y 3. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

Sobre los genes desregulados entre al menos dos de los subgrupos se llevó a cabo un análisis ORA en vías biológicas KEGG y términos GO para determinar los procesos afectados por la aplicación de diferentes concentraciones de panobinostat. Los resultados de este análisis se recogen en la **Figura 4.101**.

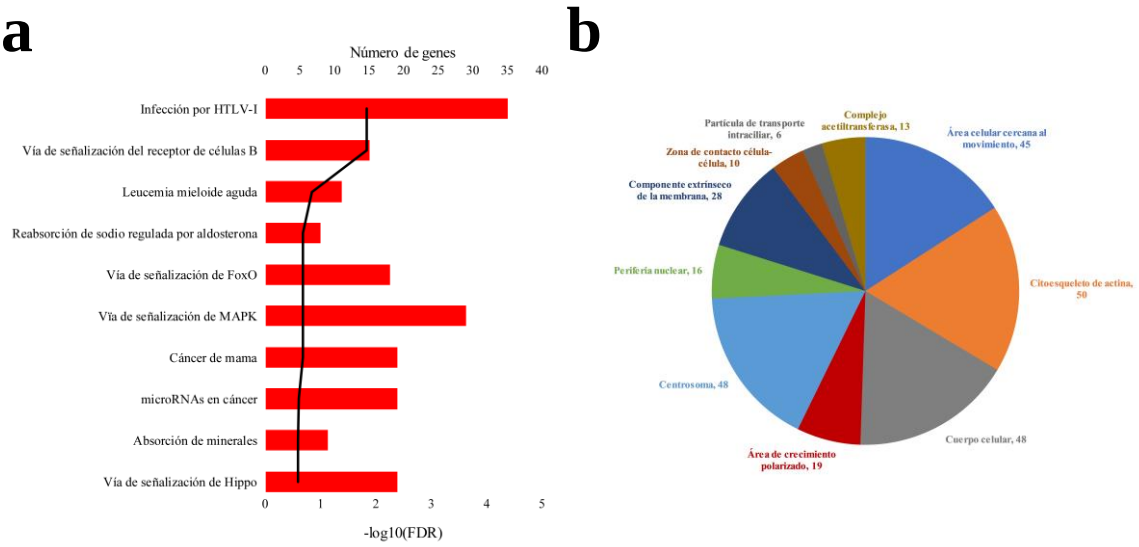



***Figura 4.101.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de panobinostat. En cada uno de los paneles se recogen las 10 rutas KEGG y los 10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localizaciones celulares GO.*

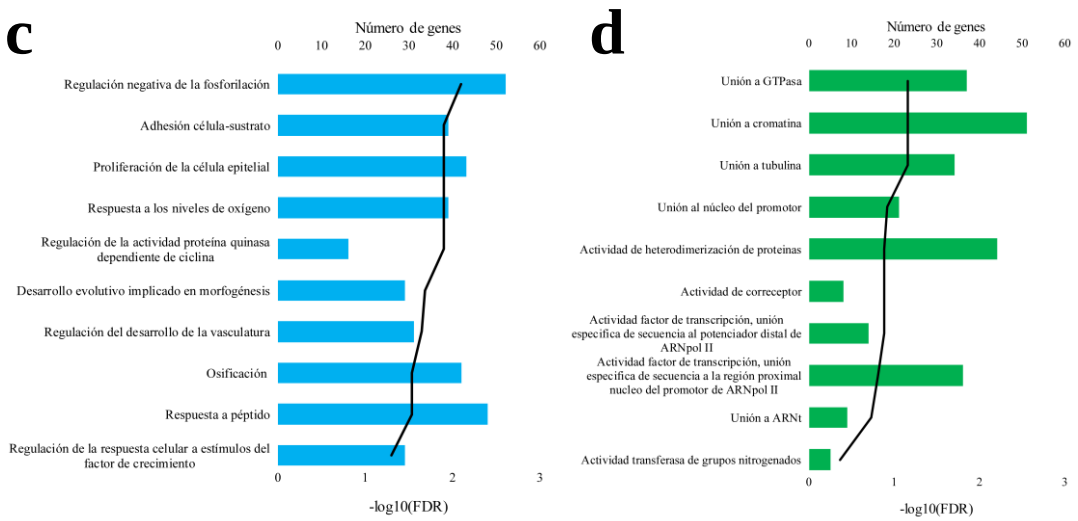



***Figura 4.101 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de panobinostat. En cada uno de los paneles se recogen las 10 rutas KEGG y los 10 términos GO con un menor valor de FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Entre los procesos que podrían verse afectados por aplicar distintas concentraciones de panobinostat se encontraba la "vía de señalización del receptor de células B" (FDR = 0,0149), que implica de forma secundaria a otras vías relevantes en el mecanismo de acción del panobinostat como la "vía MAPK"[472] y la "vía de señalización NF-κB"[473], o vías asociadas a mecanismos de resistencia frente a este fármaco como la "vía de regulación del citoesqueleto de actina"[474]. En nuestro trabajo encontramos 15 genes desregulados entre al menos dos de los subgrupos de concentración asociados a la "vía de señalización del receptor de células B" (**Figura 4.102**), pero no fue posible asociar ninguna tendencia en el cambio de expresión génica a la diferencia de concentración por dos motivos: 1) los estudios del subgrupo G2 por un lado, y del subgrupo G3 por otro correspondieron a una única serie, GSE56623 y GSE30524, respectivamente, por lo que la causa de las diferencias entre ambos subgrupos podría ser el factor microarray. 2) el subgrupo G1, solamente recogió un estudio (Paino [2010]).

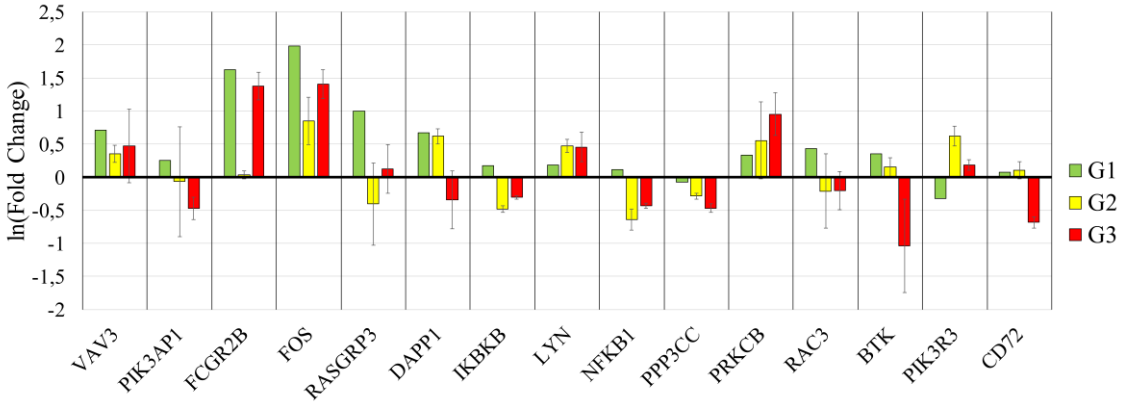



***Figura 4.102.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía de señalización del receptor de células B en los tres subgrupos de concentración de panobinostat (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.6.3. Metaanálisis global de panobinostat

El metaanálisis global sobre la expresión génica diferencial considerando los valores de FC de los 2.056 genes seleccionados en los 7 estudios reveló el efecto combinado estadísticamente significativo a *p*-valor < 0,05 de 1.706 genes, de los que 912 presentaron sobreexpresión y 794 genes infraexpresión en las muestras tratadas con panobinostat. La lista con los 1.706 genes estadísticamente significativos puede ser consultada en el **Anexo 26**. En la **Figura 4.103** se recogen las 10 rutas biológicas KEGG, así como los 10 PB, FM y CC GO más relevantes relacionados con estos 1.706 genes estadísticamente significativos.

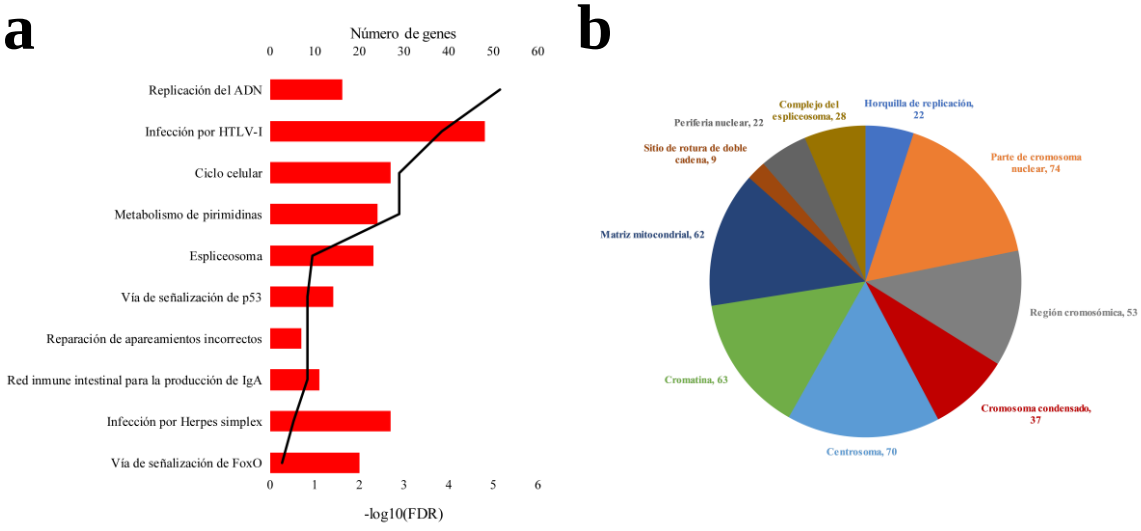



***Figura 4.103.*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 1.706 genes con un tamaño del efecto estadísticamente significativo en el metaanálisis de la expresión génica para el tratamiento con panobinostat. En cada panel se recogen las 10 vías KEGG o los 10 términos GO con menor valor del FDR.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

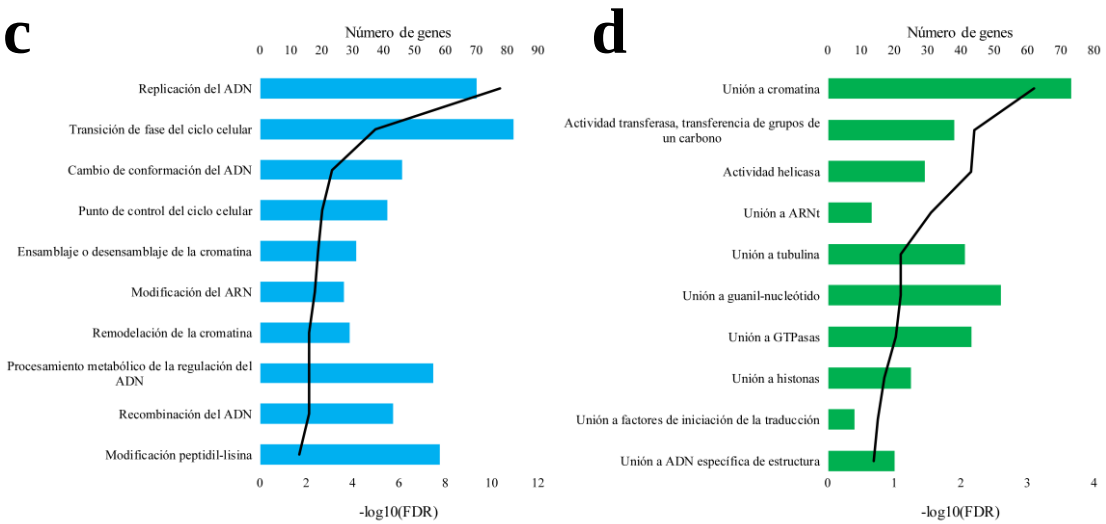



***Figura 4.103 (continuación).*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 1.706 genes con un tamaño del efecto estadísticamente significativo en el metaanálisis de la expresión génica para el tratamiento con panobinostat. En cada panel se recogen las 10 vías KEGG o los 10 términos GO con menor valor del FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Tanto la vía KEGG como la FM GO de "replicación del ADN" fueron los términos con un mayor valor de significancia en sus respectivos análisis con un FDR < 0,0001 en ambos casos. Como ya se indicó anteriormente, la regulación de la replicación del ADN es uno de los mecanismos de acción conocidos de los HDACi como panobinostat[469]. De los 36 genes en que consiste esta vía KEGG, en nuestro trabajo hemos detectado una infraexpresión estadísticamente significativa al tratamiento en 16 de ellos (**Figura 4.104**). La regulación negativa de los genes implicados en replicación del ADN ya ha sido reportada de forma particular para panobinostat en patologías como el carcinoma de células escamosas de cabeza y cuello (CCECC)[475]. Uno de los mecanismos que regula este proceso es dependiente de la expresión del gen *PCNA*, que actúa de abrazadera para la unión de la ADN polimerasa. La aplicación de panobinostat sobre las células habría producido la inhibición del gen *PCNA* (*z*-valor = -4,46, *p*-valor < 0,0001) a través de la activación del gen *CDKN1A* (*z*-valor = 6,97, *p*-valor < 0,0001)[476]. Sin embargo, según se ha descrito en CCECC, el hecho de que se haya observado desregulación negativa de una buena parte de los genes implicados en este proceso, y de la existencia de un posible mecanismo de represión de la replicación como es la inhibición de *PCNA*, parece que podría no ser suficiente para producir la inhibición de esta vía, ya que funcionaría de manera normal en presencia del fármaco, y por consiguiente, no se llegaría a producir la detención del crecimiento celular que se ha descrito que provoca el panobinostat[477].

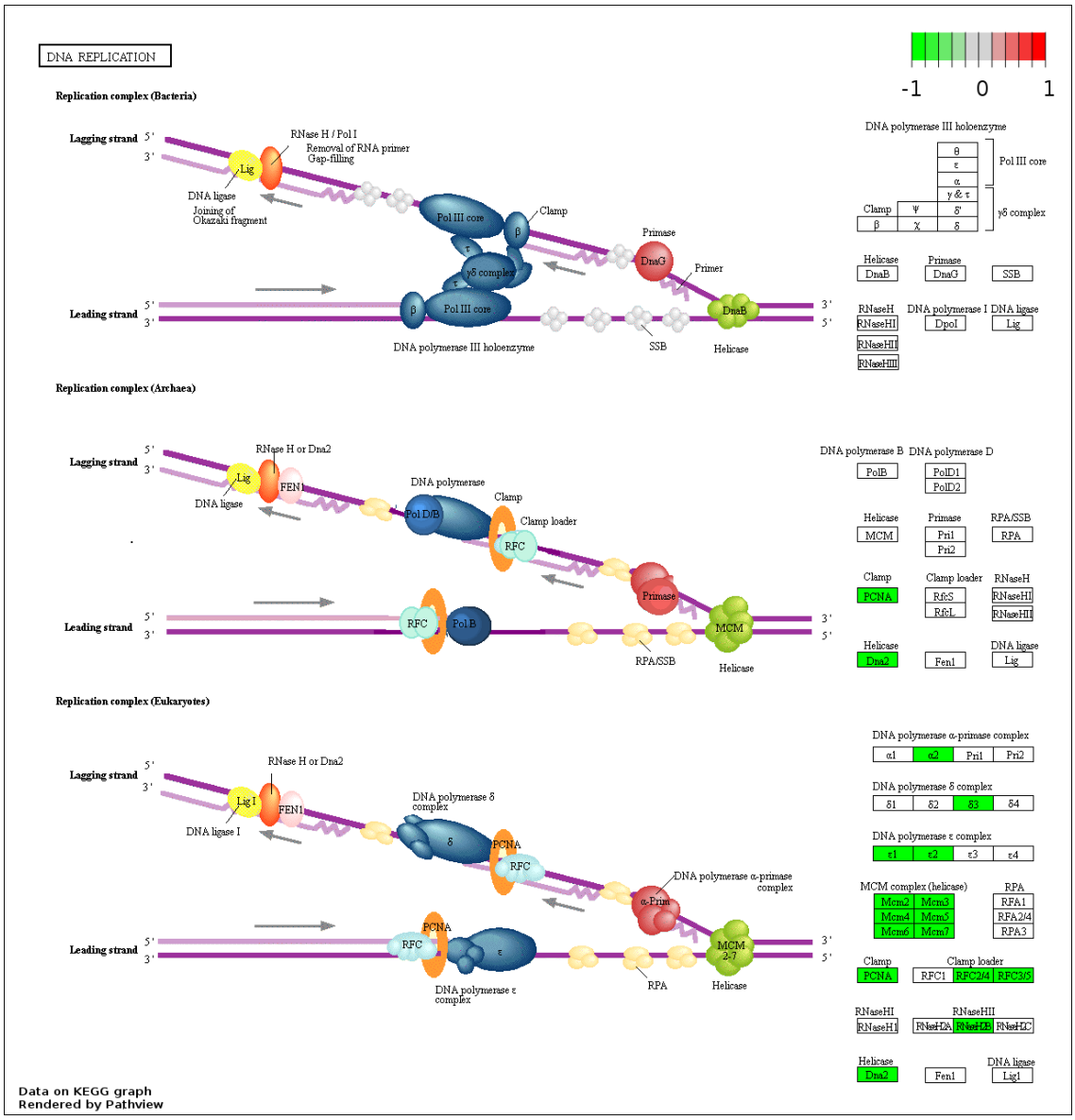



***Figura 4.104.*** *Vía de la replicación del ADN según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del panobinostat.*

Esto conduce a la búsqueda de otras vías alternativas para dilucidar el mecanismo de acción implicado en la parada del crecimiento producida por el panobinostat. En el análisis ORA resultaron también tener gran relevancia las vías y funciones asociadas al ciclo celular, de manera que la propia vía KEGG del "ciclo celular", así como los PB asociados a esta, como la "transición de fase del ciclo celular" o el "punto de control del ciclo celular", mostraron una sobrerrepresentación en genes altamente significativa (FDR < 0,01). Centrándonos en la vía KEGG "ciclo celular", detectamos 27 genes desregulados de los que seis presentaron sobreexpresión y 21 infraexpresión con el tratamiento (**Figura 4.105**). Entre estos genes nos encontramos de nuevo con *PCNA*, pero también con el gen *CDKN1A*. Respecto a este último, se ha demostrado que *CDKN1A*, reprime el promotor del gen *PLK1*[478], que a su vez está regulado positivamente por los factores de

transcripción E2F[479], controlados asimismo por *CDKN1A*. También se ha demostrado que la pérdida de expresión de *PLK1* induce la entrada en apoptosis de la célula tumoral[480], así como la inhibición del crecimiento tumoral[481]. De este modo, parece que el mecanismo de inhibición del crecimiento podría estar en relación con la sobreexpresión de *CDKN1A*, que llevaría a la represión de los genes *E2F1* (*z*-valor = -3,88, *p*-valor = 0,0001) y *E2F2* (*z*-valor = -3,33, *p*-valor = 0,0009), conduciendo también a la inhibición de *PLK1* (*z*-valor = -7,86, *p*-valor < 0,0001) dependiente de E2F[477].

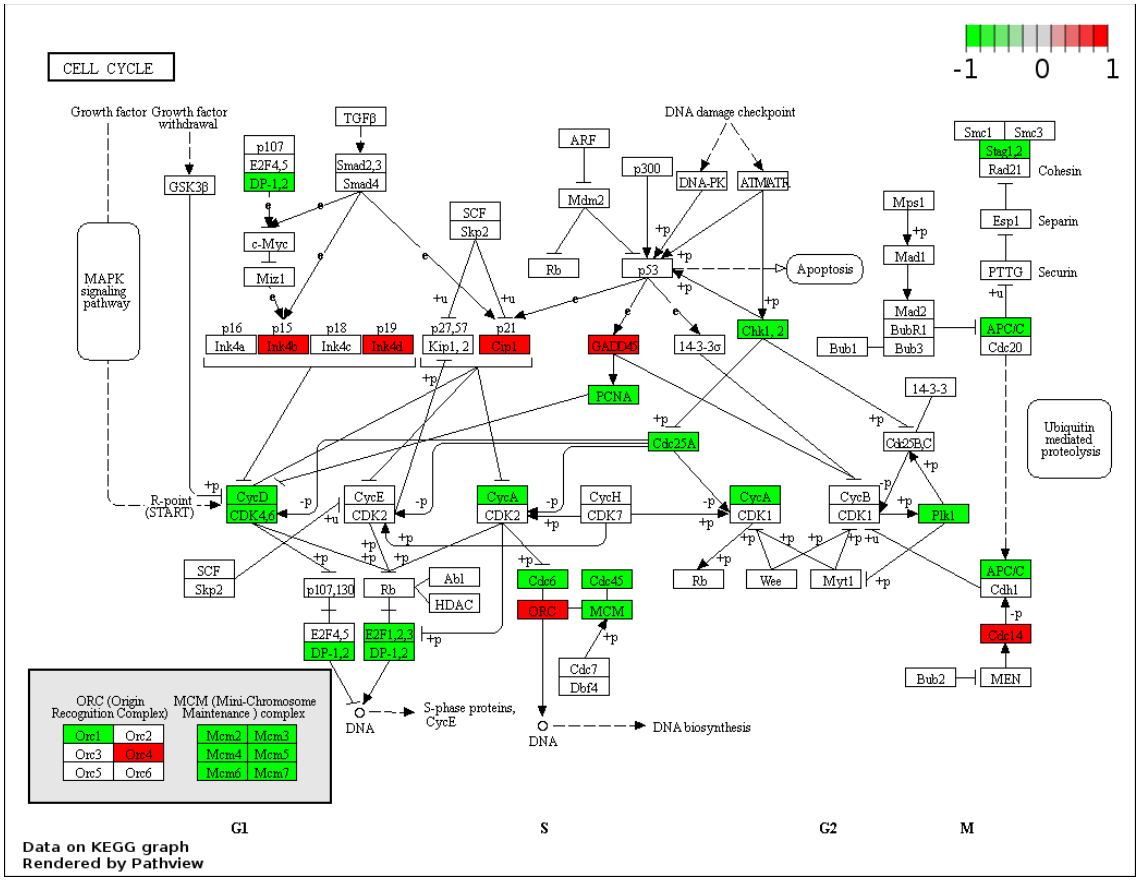



***Figura 4.105.*** *Vía del ciclo celular según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del panobinostat.*

### 4.3.7. Azacitidina

Se llevó a cabo una búsqueda sistemática de estudios de expresión génica de muestras de HMCLs tratadas con este fármaco en monoterapia. Como resultado de esta búsqueda sistemática, llevada a cabo en tres repositorios, se identificaron 268 series en GEO, 108 series en ArrayExpress y 64 muestras en SRA correspondientes a tres series. Tras la eliminación de las series duplicadas se seleccionaron 377 series para ser revisadas con mayor profundidad. De estas 377 series revisadas, únicamente una serie cumplió los criterios de inclusión y exclusión necesarios para ser considerada en el metaanálisis. Esta se trata de la serie GSE56623 que adicionalmente tuvo que ser dividida en cuatro estudios

al emplear dos líneas celulares y dos tiempos de tratamiento. En los cuatro estudios la concentración aplicada de azacitidina fue de 10 nM. El diagrama de flujo que se muestra en la **Figura 4.106**, detalla el esquema de selección de estudios para el metaanálisis de la azacitidina en función de los diferentes criterios de inclusión y exclusión.

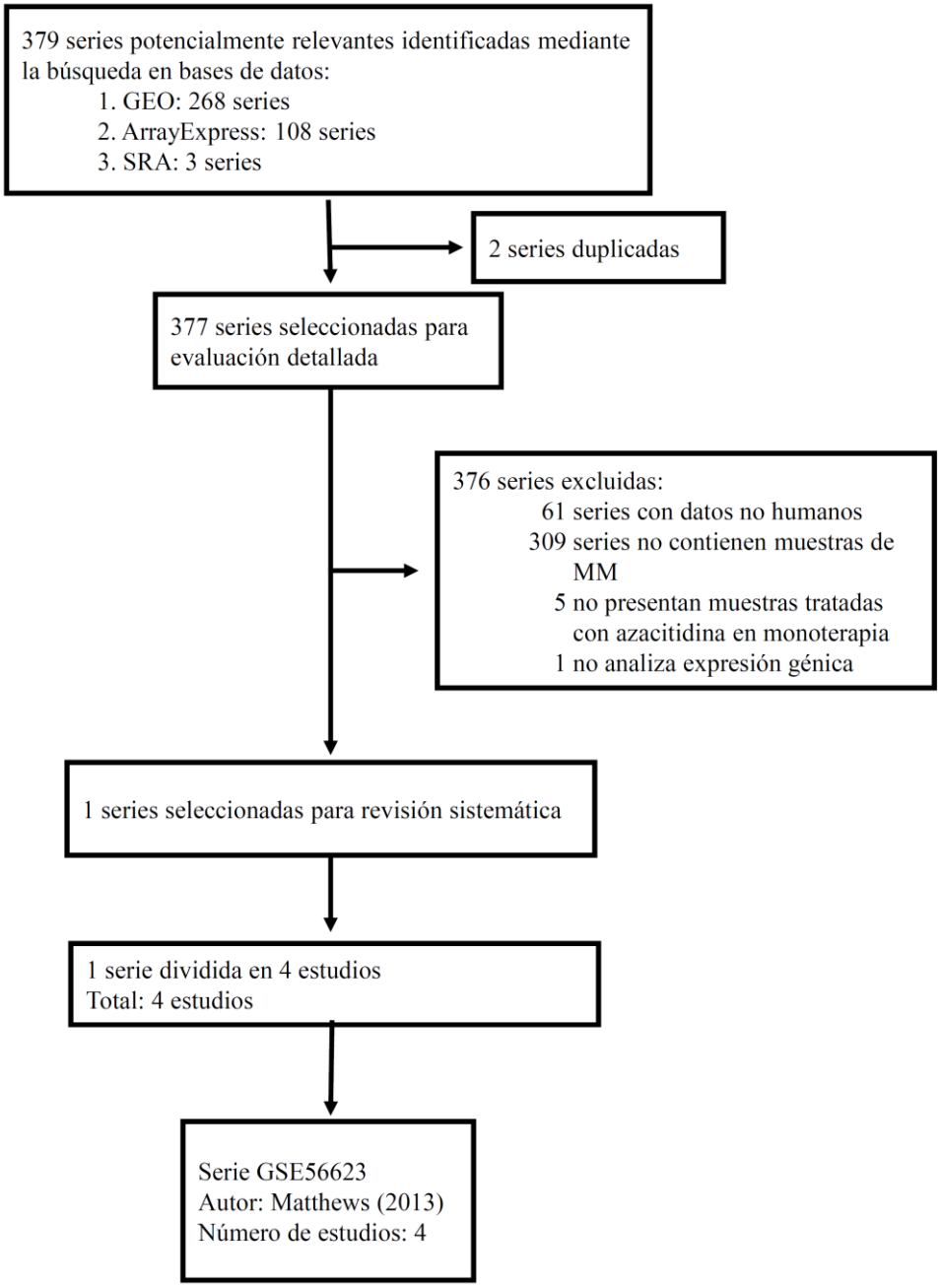



***Figura 4.106**. Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con azacitidina.*

Los cuatro estudios resultantes fueron clasificados en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento, estableciendo los tiempos a 4 y a 24 horas (mediana de 9 horas) como puntos de corte. En este caso no se fijaron subgrupos en

función de la concentración de azacitidina ya que los cuatro estudios presentaron los mismos valores de concentración de fármaco. Los resultados del agrupamiento pueden observarse en la **Tabla 4.10**.

***Tabla 4.10.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con azacitidina.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (uM) |
|---|---|---|---|---|---|---|
| GSE56623 | Matthews (2013a)[466] | JJN-3 | Illumina HiSeq 2500 | 6 | 4 | 10 |
| GSE56623 | Matthews (2013b)[466] | JJN-3 | Illumina HiSeq 2500 | 6 | 24 | 10 |
| GSE56623 | Matthews (2013c)[466] | U266 | Illumina HiSeq 2500 | 6 | 4 | 10 |
| GSE56623 | Matthews (2013d)[466] | U266 | Illumina HiSeq 2500 | 5 | 24 | 10 |

*En verde estudios seleccionados para el subgrupo G1 y en amarillo para el subgrupo G2, de tiempo de tratamiento con azacitidina.*

El cálculo de la expresión diferencial en cada uno de los estudios, considerando el valor absoluto del FC, supuso la selección de 145 genes candidatos para el metaanálisis. La representación gráfica del ln(FC) de estos genes se muestra en la **Figura 4.107** donde se puede observar una tendencia a la infraexpresión de los genes de los estudios con tiempos de tratamiento menores (Matthews [2013a] y Matthews [2013c]). La influencia sobre la expresión de los distintos tiempos de aplicación de la azacitidina será evaluada mediante metaanálisis por subgrupos en los siguientes apartados.

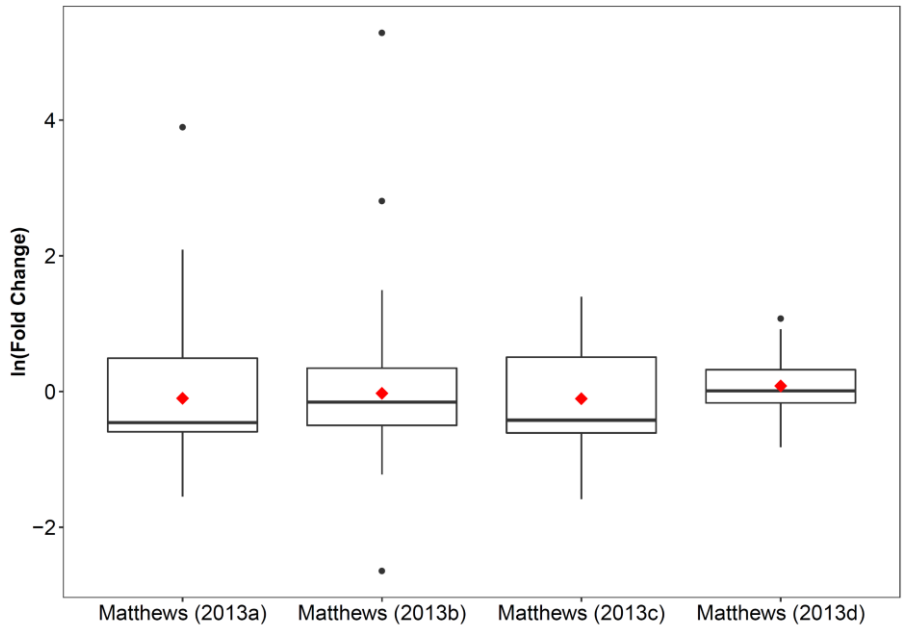



***Figura 4.107.*** *Diagrama de caja del ln(Fold Change) de los 145 genes seleccionados para el metaanálisis de azacitidina en monoterapia en líneas celulares de mieloma múltiple. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.7.1. Metaanálisis por subgrupos: tiempo de tratamiento

El estudio de la expresión génica diferencial mediante metaanálisis en los subgrupos de tiempo se realizó considerando los 145 genes seleccionados como se indicó anteriormente. Para llevar a cabo el metaanálisis, los cuatro estudios fueron clasificados en dos subgrupos en función del tiempo de tratamiento con azacitidina, ya que solamente se dispuso de estudios con tiempos de tratamiento a las cuatro y 24 horas. Por tanto, los estudios (a) y (c) de Matthews (2013) fueron asignados al subgrupo 1 (G1), mientras que el segundo subgrupo (G2) recogió los estudios (b) y (d) también de Matthews (2013). A través de la técnica de metaanálisis, se detectaron 111 genes con un valor combinado de la expresión diferencial estadísticamente significativo a *p*-valor < 0,05 en el subgrupo G1, y 57 genes en el caso de G2 (**Anexo 27**). Tras cruzar las dos listas de genes, se determinaron 48 genes comunes a los dos subgrupos, de los que 22 genes presentaron sobreexpresión y 24 genes infraexpresión en las muestras tratadas con azacitidina. En la **Figura 4.108** se muestran dos ejemplos de diagramas de bosque de los genes con mayor valor absoluto de la mediana de FC considerando los cuatro estudios.

**a**

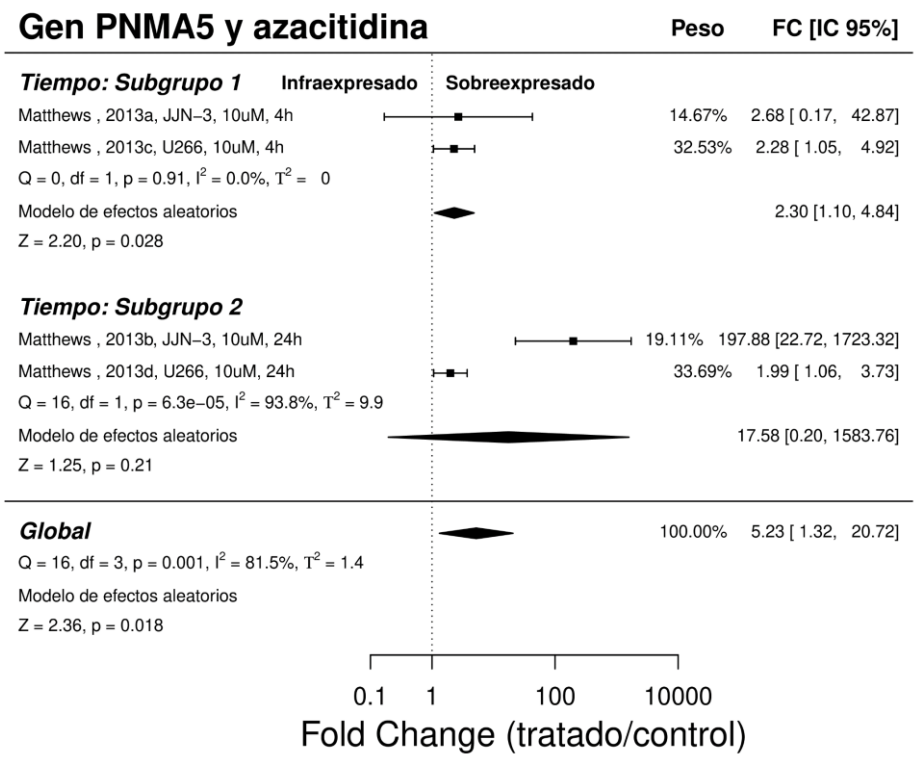



***Figura 4.108.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con azacitidina.* ***a)*** *Diagrama de bosque del gen PNMA5, que fue el más sobreexpresado considerando la mediana del FC de los cuatro estudios seleccionados.*

**b**

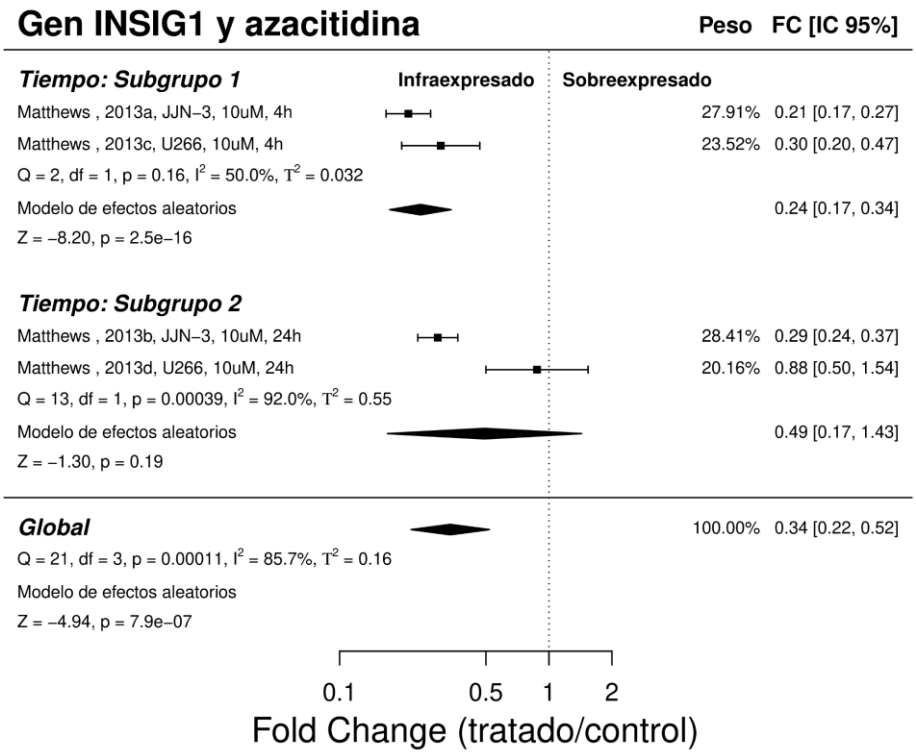



***Figura 4.108 (continuación).*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con azacitidina.* ***b)*** *Diagrama de bosque del gen INSIG1, que fue el más infraexpresado considerando la mediana del FC de los cuatro estudios seleccionados.*

La comparación de los resultados obtenidos en los dos subgrupos mediante una prueba estadística tipo Wald detectó que 61 genes de los 145 sometidos a metaanálisis presentaron diferencias estadísticamente significativas entre G1 y G2 (**Anexo 27**). En la **Figura 4.109** se muestra mediante un diagrama de puntos las diferencias entre los valores de ln(FC) obtenidos en los dos subgrupos.

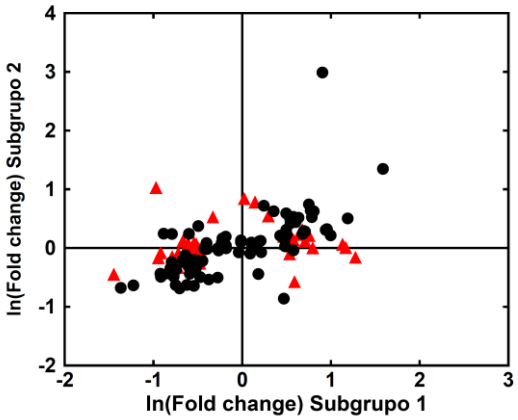



***Figura 4.109.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 145 genes estudiados donde se comparan los subgrupos 1 y 2. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

En un último paso se realizó el análisis ORA de genes en rutas biológicas KEGG y términos GO utilizando los 61 genes diferencialmente expresados entre los dos subgrupos. Mediante este análisis se buscó determinar qué vías o términos se vieron más afectados por la aplicación de diferentes tiempos de tratamiento. Los resultados de este análisis se detallan en la **Figura 4.110**.

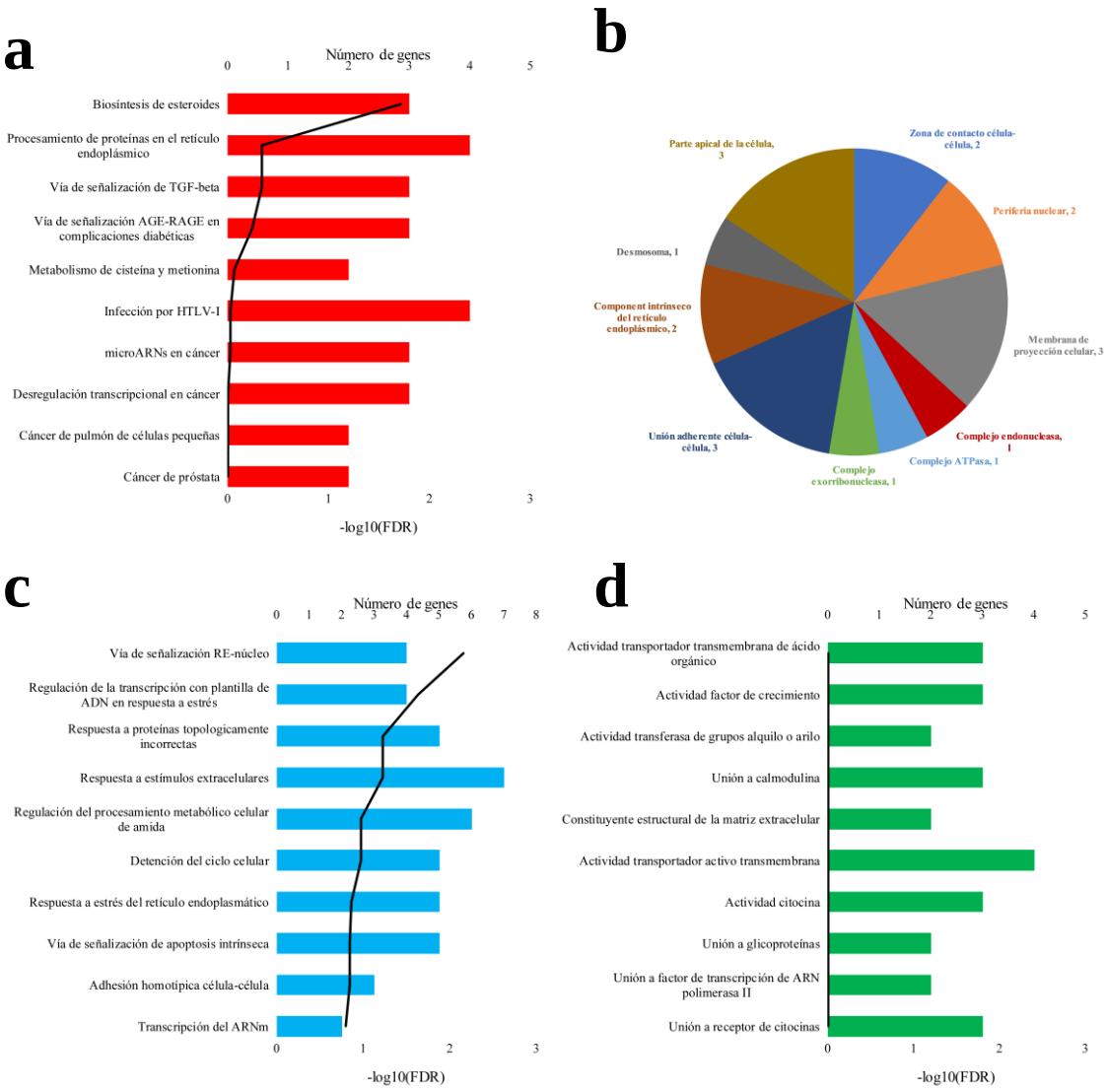



***Figura 4.110.**: Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con azacitidina. En cada panel se recogen las 10 rutas KEGG o los10 términos GO con un menor valor de FDR. **a)** TOP 10 rutas biológicas KEGG, **b)** TOP 10 componentes celulares GO, **c)** TOP 10 procesos biológicos GO y **d)** TOP 10 funciones moleculares GO.*

Considerando las vías KEGG, la influencia del tiempo de tratamiento fue estadísticamente significativa en la vía de "biosíntesis de esteroides" con un FDR = 0,0195. El efecto de la azacitidina en esta vía ha sido previamente reportado en relación a la modulación de la biosíntesis de colesterol[482]. En este estudio se detectaron tres genes desregulados por azacitidina cuya expresión podría ser dependiente del tiempo de tratamiento: *FDFT1*, cuya desregulación ha sido descrita en trabajos con decitabina[483],

*DHCR7* y *LSS*, todos implicados en la biosíntesis de colesterol. El efecto del tiempo de tratamiento que se detectó en este análisis fue un mayor cambio de expresión génica a los tiempos del subgrupo G1, es decir, a las 4 horas. El efecto del compuesto a los tiempos del subgrupo G2 (24 horas) apareció muy atenuado en relación a G1 (**Figura 4.111a**). Para confirmar este resultado con una lista más amplia de genes, se procedió al análisis de los genes del PB más significativo, que en este caso fue la "vía de señalización RE-núcleo" (FDR = 0,0071), observándose el mismo efecto que en el análisis anterior, tanto en genes sobreexpresados como infraexpresados (**Figura 4.111b**). Por tanto, parece que los tiempos cortos de tratamiento podrían favorecer el mecanismo de desregulación de la expresión génica de la azacitidina.

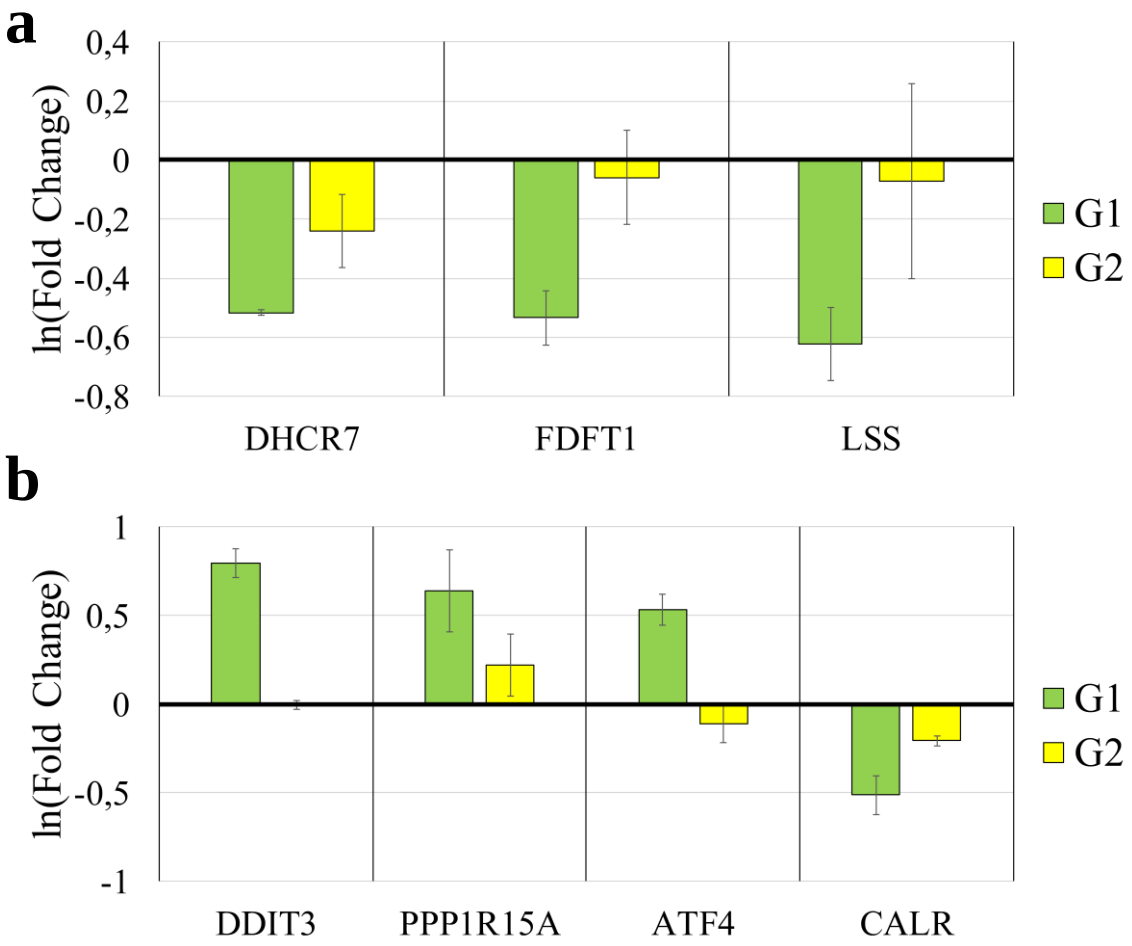



***Figura 4.111.*** *Valores promedio del ln(Fold Change) de los genes desregulados en* ***a)*** *la vía KEGG de "biosíntesis de esteroides" y* ***b)*** *la FM de GO "vía de señalización retículo endoplásmico (RE)-núcleo", en los dos subgrupos de tiempo de tratamiento con azacitidina (G1 y G2). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.7.2. Metaanálisis por subgrupos: concentración

Como se indicó en el **Apartado 4.3.7.** no se llevó a cabo el metaanálisis por subgrupos de concentración ya que los cuatro estudios seleccionados presentaron la misma concentración de azacitidina.

### 4.3.7.3. Metaanálisis global de la azacitidina

El metaanálisis global sobre los cuatro estudios seleccionados para el metaanálisis de la azacitidina reveló el efecto combinado estadísticamente significativo (*p*-valor < 0,05) de 91 genes, de los que 38 presentaron sobreexpresión y 53 genes infraexpresión en las muestras tratadas con azacitidina (**Anexo 28**). En la **Figura 4.112** se recogen las 10 rutas biológicas KEGG y los términos GO a los niveles PB, FM y CC más relevantes asociados con estos 91 genes.

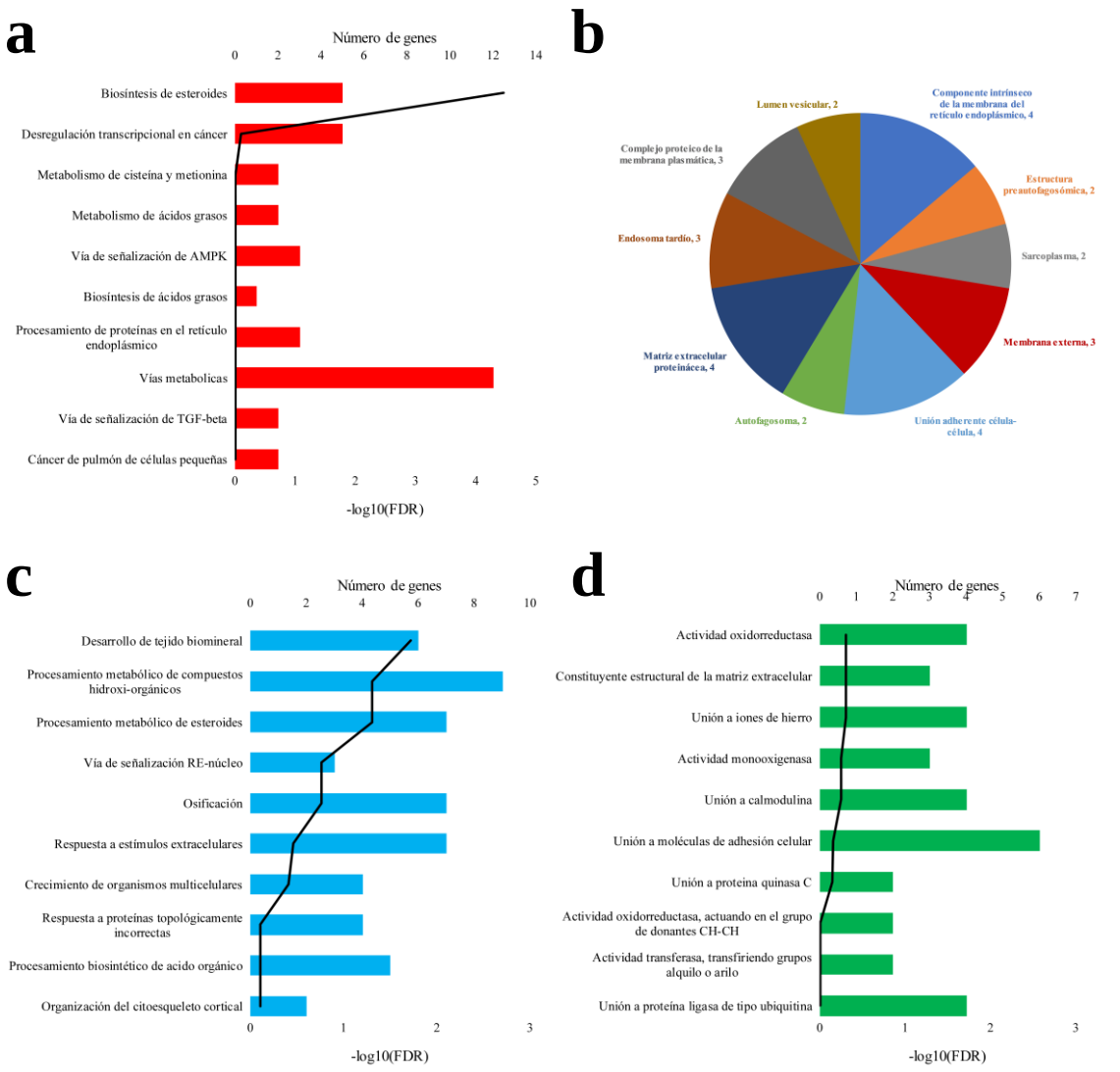



***Figura 4.112.*** *Análisis de sobrerrepresentación sobre vías biológicas KEGG y términos GO considerando los 91 genes que presentaron una diferencia del tamaño del efecto estadísticamente significativa en el metaanálisis de la expresión génica para el tratamiento con azacitidina. En cada panel se recogen las 10 vías KEGG y los 10 términos GO con menor valor del FDR.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

En el análisis de vías KEGG solamente una vía obtuvo resultados estadísticamente significativos: la "vía de biosíntesis de esteroides" (FDR < 0,0001). Como ya se apuntó

anteriormente, estudios previos han reportado la acción de la azacitidina sobre uno de los brazos de esta vía como es la biosíntesis de colesterol[482]. En nuestro trabajo se detectaron cinco genes desregulados en esta vía, de los que cuatro presentaron infraexpresión, *DHCR7*, *FDFT1*, *LSS1* y *MSMO1*, mientras que un gen, el citocromo *CYP27B1*, presentó sobreexpresión. Aunque la detección de infraexpresión podría no considerarse *a priori* como un efecto esperado en el mecanismo de acción de un agente hipometilante, la azacitidina, así como la decitabina, cuentan con mecanismos de acción duales, que en el caso de la azacitidina son debidos a la dosis aplicada del compuesto, produciendo hipometilación a dosis bajas y citotoxicidad a dosis más altas[139]. Volviendo a la "vía de biosíntesis de esteroides", los cinco genes desregulados están implicados en la biosíntesis de colesterol (**Figura 4.113**) y uno de ellos, *FDFT1*, codifica una enzima, la escualeno sintasa (SQS), esencial para la biosíntesis de colesterol en humanos[484] y responsable de la regulación del gen *LSS1*, implicado en el transporte inverso de colesterol del hígado al plasma[485]. La actividad de la enzima SQS, junto con el proceso de síntesis de colesterol, se han demostrado esenciales para la formación de balsas lipídicas asociadas al colesterol en las células cancerígenas[486], proponiéndose que la sobreexpresión de este gen podría estar involucrado en el desarrollo de procesos carcinogénicos[487]. La infraexpresión del gen *FDFT1* (*z*-valor = -2,21, *p*-valor = 0,0272) se ha observado en tratamientos con otros agentes hipometilantes como la decitabina[483], y su desregulación junto con la de los genes *LSS1* (*z*-valor = -2,55, *p*-valor = 0,0107) y *MSMO1* (*z*-valor = -3,60, *p*-valor = 0,0003) también se ha descrito con otros agentes terapéuticos como el paclitaxel[488], sugiriéndose en ambos casos que provocaría una regulación negativa de la biosíntesis del colesterol mediante la cual, este fármaco ejercería su efecto citotóxico en las células cancerígenas.

En lo que respecta a la sobreexpresión del gen *CYP27B1* (*z*-valor = 6,15, *p*-valor < 0,0001) se ha visto que su promotor aparece metilado en varios tipos de cáncer, como el cáncer de mama[489] o los linfomas[490]. De hecho, el descenso de su expresión se ha asociado con el incremento de la agresividad en cáncer de ovario[491], y el incremento de su expresión se ha asociado a un papel quimiopreventivo en cáncer debido a la modulación de la producción de vitamina D[492]. Por tanto, la azacitidina podría producir la desmetilación del promotor del gen *CYP27B1*, favoreciendo su expresión y la producción de vitamina D, lo que conduciría a la reducción del riesgo de progresión del cáncer[493]. Además, este gen también está implicado en uno de los PB estadísticamente significativos asociados en este trabajo al tratamiento con azacitidina, el "desarrollo de tejido biomineral" (FDR = 0,0192). Se ha demostrado en ratones, que la ausencia de la proteína codificada por *CYP27B1*, produce una pobre mineralización ósea[494], por lo que el rescate de la expresión de este gen, junto con genes como *NBR1* (*z*-valor = 4,05, *p*-valor < 0,0001), implicado en el mantenimiento del hueso, podría conducir a una mejoría en el funcionamiento del proceso de osificación en pacientes tratados con azacitidina.

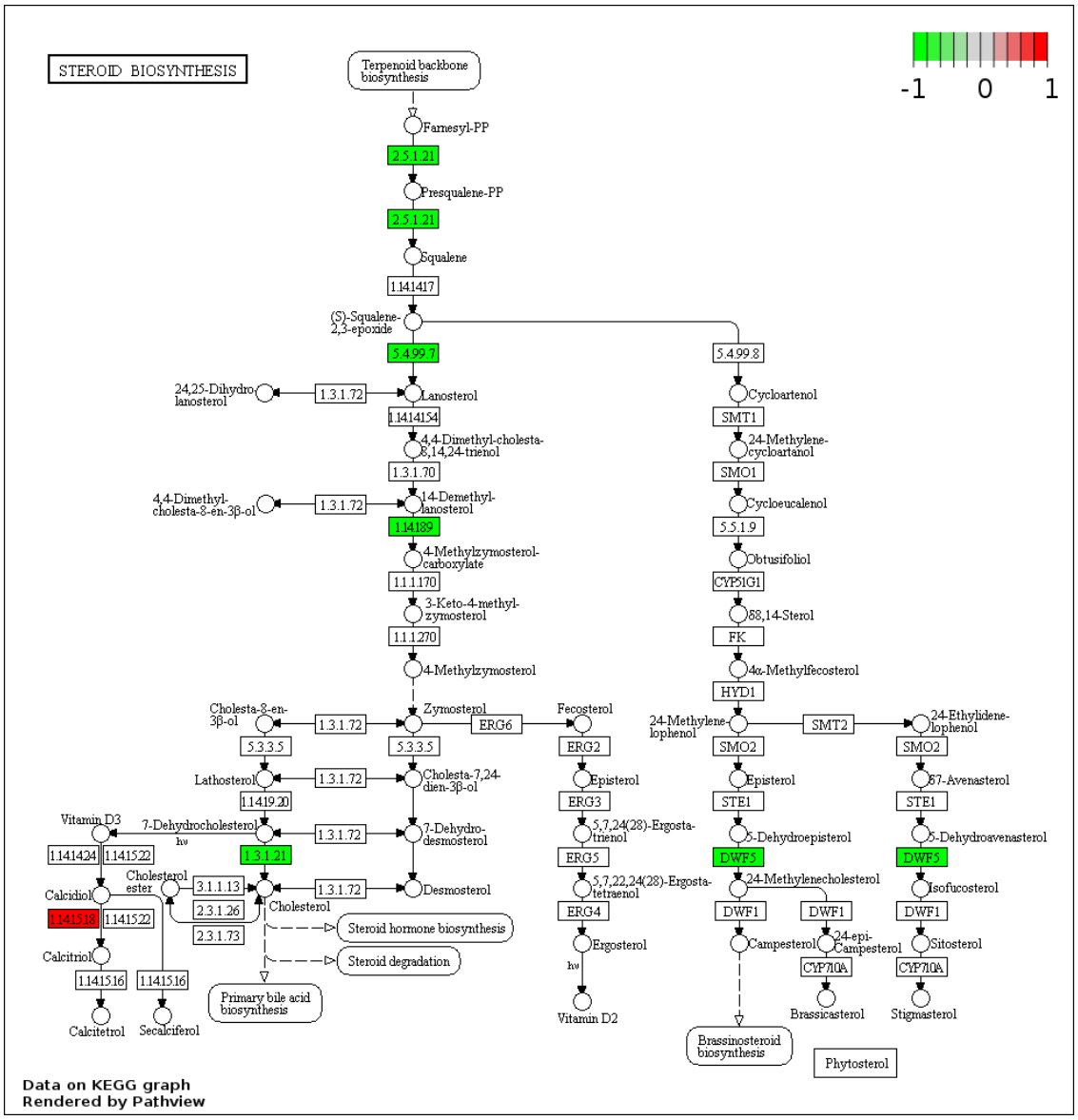



***Figura 4.113.*** *Vía de la biosíntesis de esteroides según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global de la azacitidina.*

### 4.3.8. Decitabina

Se llevó a cabo una búsqueda sistemática de estudios de HMCLs tratadas con decitabina en monoterapia en repositorios online de datos genómicos. De esta manera, se localizaron 161 series en GEO, 130 series en ArrayExpress y 45 muestras correspondientes a 7 series en SRA. Del total de 298 series, 204 series fueron seleccionadas para su revisión en detalle tras la eliminación de los elementos duplicados en los tres repositorios. Únicamente dos de las 204 series cumplieron los criterios de inclusión y exclusión necesarios para ser consideradas en el metaanálisis. Una de las series incluidas, concretamente la serie E-MEXP-1269, fue posteriormente dividida en tres estudios debido a que presentó datos procedentes de tres líneas celulares diferentes. Así, el número final de estudios considerados para el metaanálisis fue de cuatro. El diagrama de flujo que se muestra en la **Figura 4.114**, detalla el esquema de selección de estudios para el metaanálisis de la decitabina en función de los diferentes criterios de inclusión y exclusión.

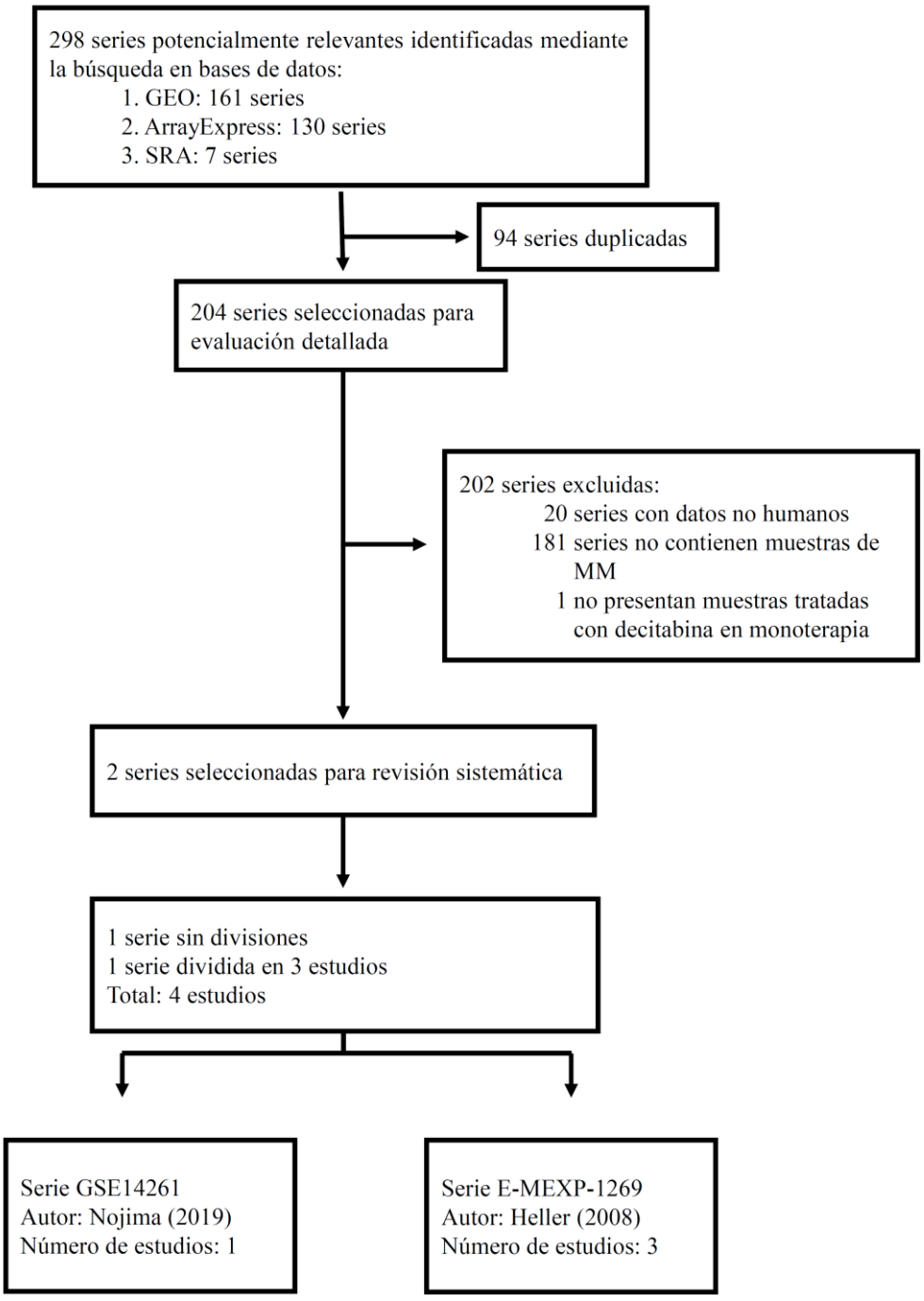



***Figura 4.114.*** *Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con decitabina.*

Estos cuatro estudios seleccionados fueron a continuación clasificados en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de decitabina. El punto de corte para el tiempo de tratamiento fue establecido a las 168 horas, mientras que el punto de corte para la concentración se determinó a 0,5 µM. Los resultados del agrupamiento en subgrupos pueden observarse en la **Tabla 4.11**.

***Tabla 4.11.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con decitabina.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (uM) |
|---|---|---|---|---|---|---|
| GSE14261 | Nojima (2009)[443] | OPM1 | Agilent-Whole Human Genome Microarray 4x44K | 2 | 72 | 0.1 |
| E-MEXP-1269 | Heller (2008a)[495] | H929 | Affymetrix Human Genome U133A | 4 | 168 | 0.5 |
| E-MEXP-1269 | Heller (2008b)[495] | MM1-S | Affymetrix Human Genome U133A | 4 | 168 | 0.5 |
| E-MEXP-1269 | Heller (2008c)[495] | U266 | Affymetrix Human Genome U133A | 4 | 168 | 0.5 |

*En verde, estudios seleccionados para el subgrupo de tiempos o concentraciones bajos; en amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedias.*

Una vez determinados los subgrupos para el metaanálisis de la decitabina, se seleccionaron 123 genes candidatos en función de su valor absoluto del FC. En la **Figura 4.115** se muestra la distribución del ln(FC) de los 123 genes seleccionados, donde puede apreciarse una tendencia en todos los estudios a la sobreexpresión de los genes seleccionados, con medianas para los ln(FC) superiores a cero en los cuatro estudios. La prevalencia de la sobreexpresión en todos los estudios podría ser el reflejo del efecto de la hipometilación producida por la decitabina.

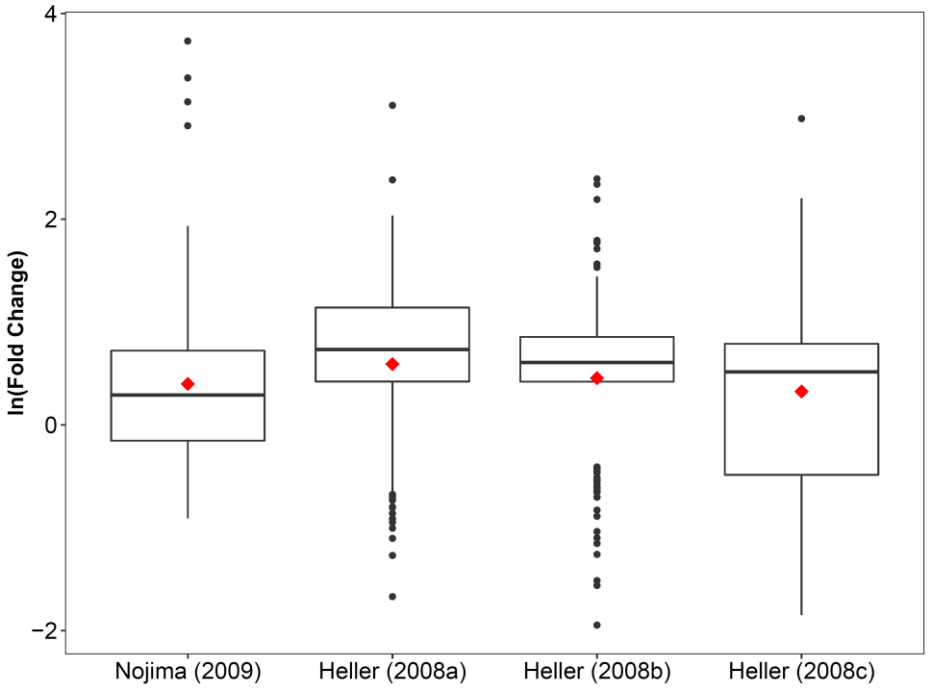



***Figura 4.115.*** *Diagrama de caja del ln(Fold Change) de los 123 genes seleccionados para el metaanálisis de decitabina en monoterapia en líneas celulares de MM. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

La posible influencia de los tiempos de tratamiento y concentraciones de decitabina sobre la expresión génica en los cuatro estudios seleccionados se evaluó mediante metaanálisis por subgrupos, tal y como se recoge a continuación.

#### 4.3.8.1. Metaanálisis por subgrupos: tiempo de tratamiento y concentración

El metaanálisis por subgrupos de tiempo y concentración para la decitabina se llevó a cabo de forma conjunta debido a que la clasificación de los estudios en subgrupos fue similar para ambos factores de variabilidad. De este modo, en el primer subgrupo (G1) únicamente se incluyó el estudio de Nojima (2009), que se realizó a un tiempo de tratamiento de 72 horas y una concentración de 0,1 µM. En el subgrupo 2 (G2) se encuentran los estudios con un tiempo de tratamiento de 168 horas y una concentración de 0,5 µM, quedando encuadrados en el subgrupo los estudios (a), (b) y (c) de Heller (2008). El metaanálisis por subgrupos reveló 60 genes con diferencias en la expresión estadísticamente significativas a *p*-valor < 0,05 en el subgrupo G1 y 87 genes en el subgrupo G2 (**Anexo 29**). El cruce de las listas de genes estadísticamente significativos de ambos subgrupos encontró que 41 de estos genes fueron comunes a ambas listas, de los que 29 estaban sobreexpresados y 7 infraexpresados al tratar con decitabina, el resto de los genes presentaron sentidos de expresión en los dos subgrupos. En la **Figura 4.116** se muestran dos ejemplos de diagramas de bosque de los genes con mayor valor absoluto de la mediana de FC considerando los cuatro estudios.

**a**

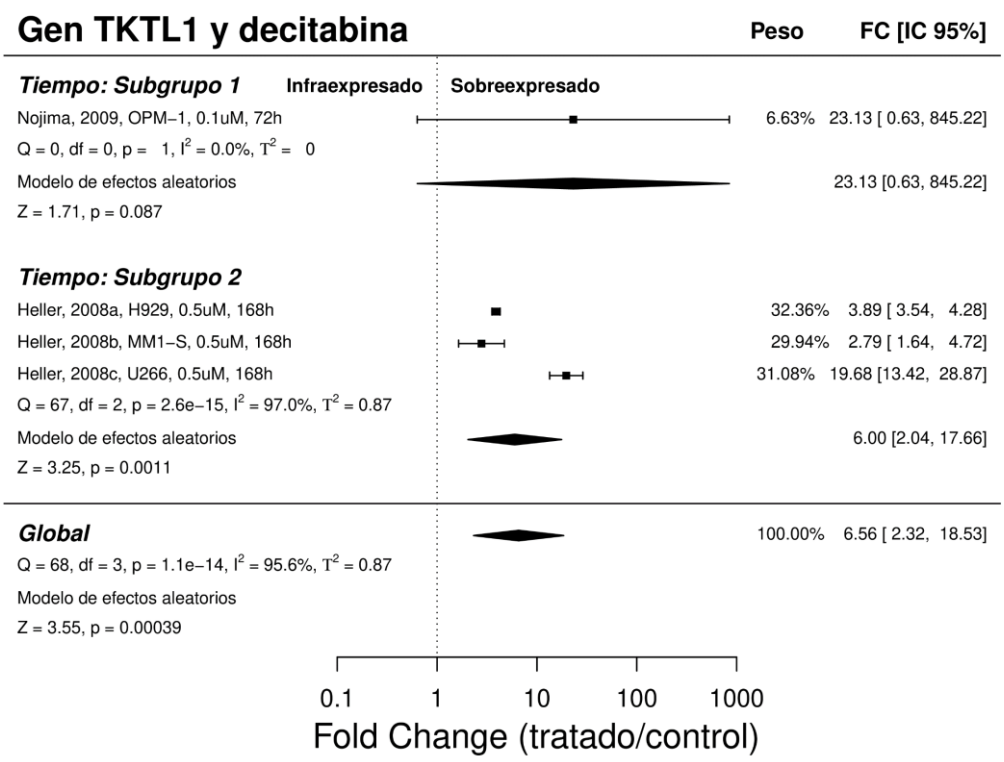



***Figura 4.116**: Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con decitabina. **a)** Diagrama de bosque del gen TKTL1 que fue el más sobreexpresado considerando la mediana del FC de los cuatro estudios seleccionados.*

**b**

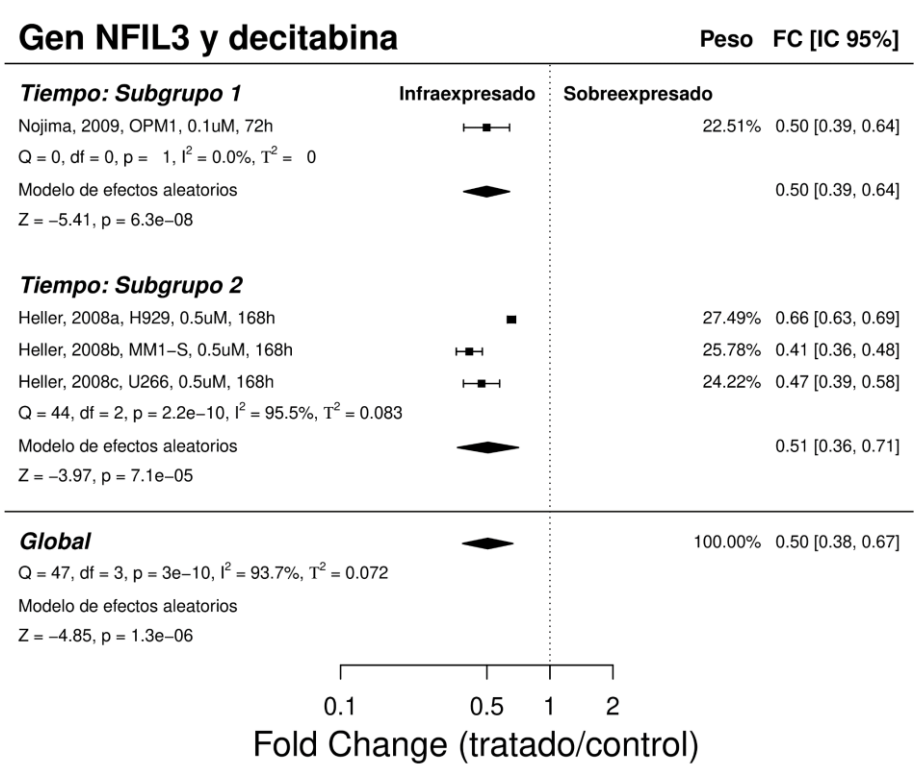



***Figura 4.116 (continuación)**. Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con decitabina. **b)** Diagrama de bosque del gen NFIL3, que fue el más infraexpresado considerando la mediana del FC de los cuatro estudios seleccionados.*

En un siguiente análisis se trató de identificar las posibles diferencias entre los dos subgrupos mediante una prueba estadística tipo Wald en los 123 genes analizados, obteniendo que 34 genes presentaron diferencias estadísticamente significativas entre los subgrupos G1 y G2 (**Anexo 29**). En la **Figura 4.117** se muestra las diferencias entre los valores de ln(FC) obtenidos en los dos subgrupos.

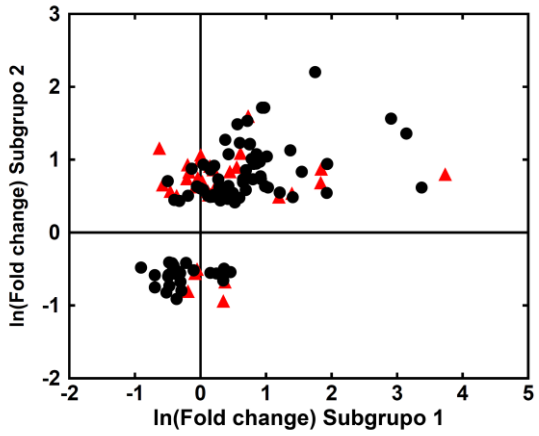



***Figura 4.117.** Diagrama de puntos de los valores de ln(FC) obtenidos para los 123 genes estudiados donde se comparan los subgrupos 1 y 2. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre los dos subgrupos.*

Mediante un análisis ORA sobre los 34 genes se trató de determinar las vías biológicas KEGG y los términos GO que podrían verse afectados por la aplicación de diferentes tiempos de tratamiento o concentración de decitabina (**Figura 4.118)**.

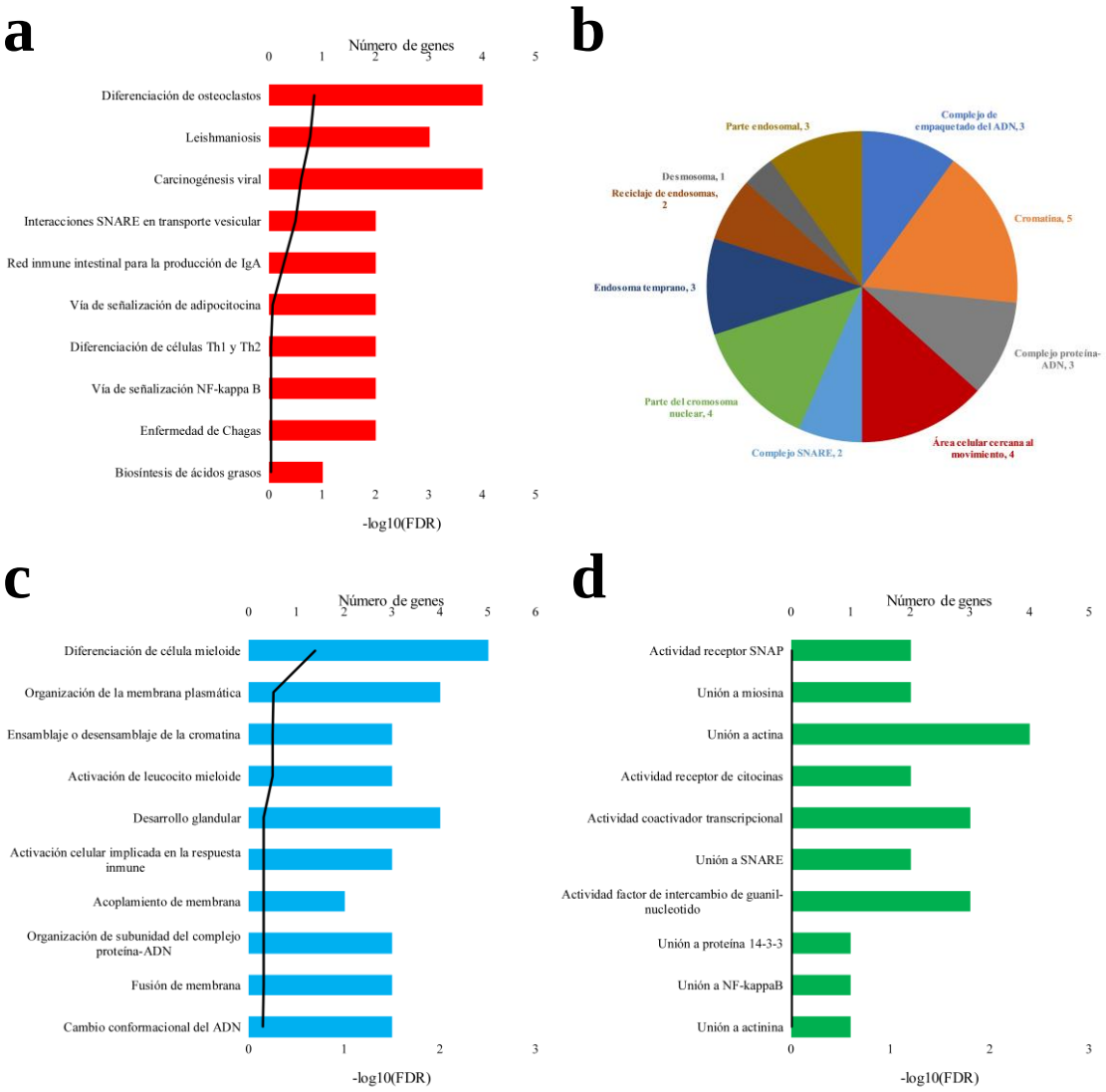



***Figura 4.118****. Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento y concentración de decitabina. En cada panel se recogen las 10 vías KEGG o los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Ninguna de las vías biológicas KEGG o términos GO resultó estadísticamente significativo a FDR < 0,05. No obstante, se procedió al estudio del efecto del tiempo de tratamiento y la concentración aplicada de decitabina sobre los genes de la vía KEGG "diferenciación de osteoclastos" que fue la que alcanzó un menor valor de FDR (FDR = 0,1447). Esta vía constó de cuatro genes desregulados en los que se observó un mayor incremento de la expresión en los estudios del subgrupo G2 que en el estudio del G1 (**Figura 4.119**), lo que podría indicar que un aumento del tiempo de tratamiento y de la concentración induciría a este incremento de la expresión. Sin embargo, este resultado

tiene que ser tratado de forma cautelosa al consistir el subgrupo G1 en un único estudio, y al estar el G2 constituido por tres estudios de la misma serie experimental.

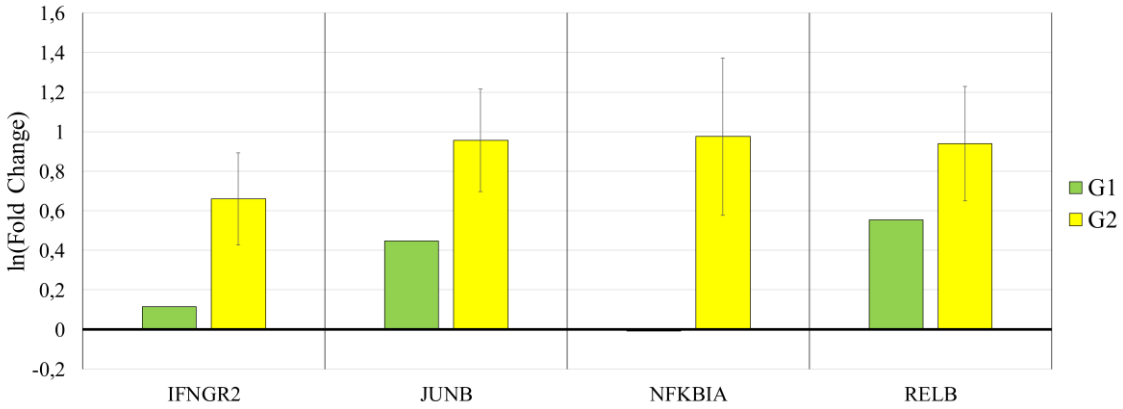



***Figura 4.119.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía de diferenciación de osteoclastos en los dos subgrupos de decitabina (G1 y G2). Las barras de error representan la desviación estándar del ln(Fold Change).*

### 4.3.8.2. Metaanálisis global de la decitabina

El metaanálisis global considerando los cuatro estudios de manera conjunta reveló el efecto combinado estadísticamente significativo ($p$-valor < 0,05) de 80 genes, de los que 66 estaban sobreexpresados y 14 infraexpresados en las muestras tratadas con decitabina (**Anexo 30**). En la **Figura 4.120** se recogen las 10 rutas biológicas KEGG, y los términos GO más relevantes asociados con estos 80 genes desregulados.

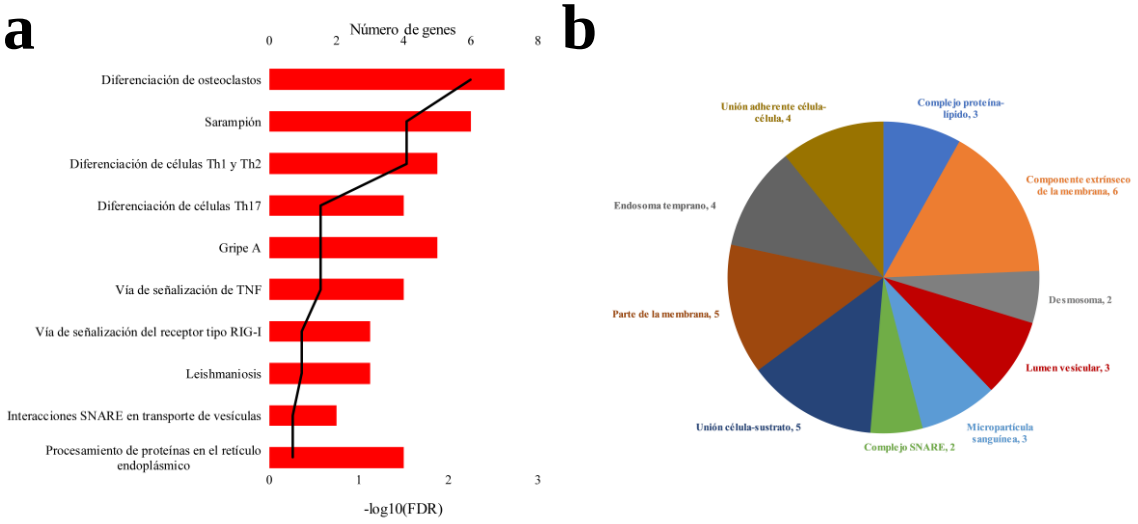



***Figura 4.120.*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 80 genes con una diferencia en el tamaño del efecto estadísticamente significativa en el metaanálisis de la expresión génica de la decitabina. En cada panel se recogen las 10 rutas KEGG o los10 términos con menor FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

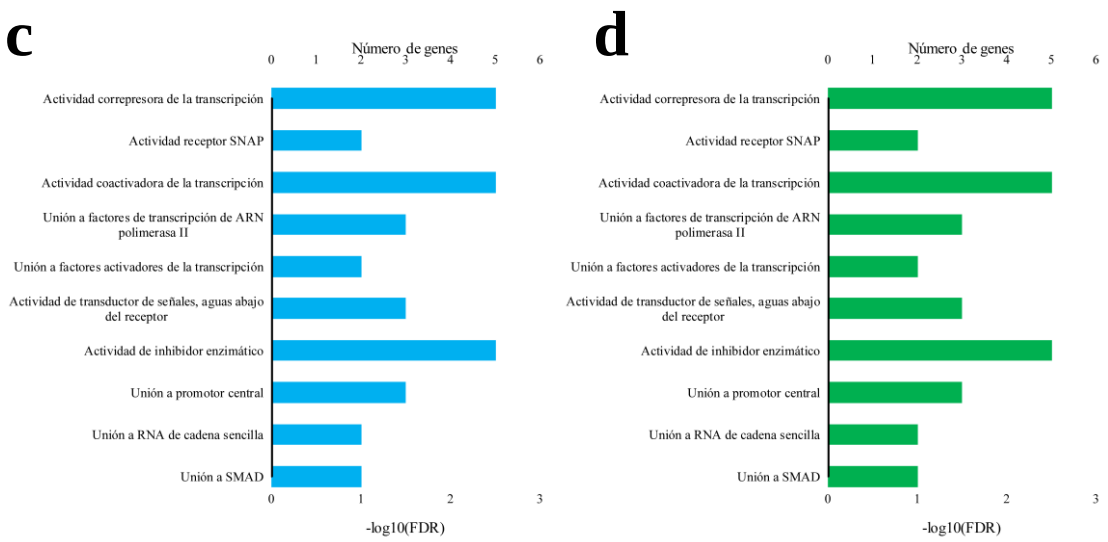



***Figura 4.120 (continuación).*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 80 genes con una diferencia en el tamaño del efecto estadísticamente significativa en el metaanálisis de la expresión génica de la decitabina. En cada panel se recogen las 10 rutas KEGG o los10 términos con menor FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

A pesar del bajo número de genes diferencialmente expresados, tres vías KEGG resultaron estadísticamente significativas a FDR < 0,05. La vía con mayor significancia estadística según el FDR fue la "vía de diferenciación de osteoclastos" (FDR = 0,0057), representada por 7 genes, todos ellos sobreexpresados después del tratamiento con decitabina (**Figura 4.121**). Uno de estos genes es *JUNB* (*z*-valor = 5,79, *p*-valor < 0,0001), cuya activación a través de la desmetilación de su promotor por la decitabina ha sido demostrada en estudios en leucemia mieloide crónica (LMC)[496]. Se ha observado además que su ausencia en leucemia linfoide B conduce a la aparición de células transformadas muy agresivas, siendo un indicador de una enfermedad de rápida progresión[497]. Sin embargo, también se ha reportado que *JUNB* puede tener un efecto proliferativo en patologías como el linfoma de Hodgkin[498]. Estudios recientes en MM, han demostrado *in vitro* y en modelos murinos que el silenciamiento de *JUNB* reduce el crecimiento y la supervivencia tumoral, si bien también se ha observado que niveles altos de *JUNB* en pacientes con MM se asociaban con mejor pronóstico. Esto ha llevado a los autores a especular con que la función de este gen depende del origen celular, el estadio del ciclo celular y las condiciones ambientales de las células a estudio[499].

Entre los seis genes restantes de esta vía, tres de ellos, *RELB*, *NFKB2* y *NFKBIA*, pertenecen a la vía de señalización NF-κB. Esta vía es de gran importancia en MM ya que está involucrada en procesos como la formación de osteoclastos[500], la expresión de reguladores de ciclo celular o la angiogénesis[501]. La sobreexpresión del gen *NFKBIA* (*z*-valor = 2,74, *p*-valor < 0,0062) por parte de la decitabina llevaría a la inhibición de la vía clásica NF-κB, que está mediada por el dímero p50-RelA, conduciendo a uno de los efectos previamente observados para este compuesto que es la represión de la osteoclastogénesis[502]. Sin embargo, el tratamiento con decitabina también produce la

sobreexpresión de los genes *RELB* (*z*-valor = 6,71, *p*-valor < 0,0001) y *NFKB2* (*z*-valor = 6,84, *p*-valor < 0,0001) que son mediadores no canónicos de la vía NF-κB, que codifican las proteínas RelB y p52, respectivamente. Mientras que *NFKBIA* produce una inhibición de la vía clásica de NF-κB, el efecto inhibitorio sobre la vía alternativa mediada por p52-RelB es pobre[503], con lo que estos datos sugieren que se estaría produciendo la activación de la vía alternativa no canónica de NF-κB, quizá en respuesta a la inhibición de la vía clásica, lo que podría estar asociado a un mecanismo de quimiorresistencia, tal y como se ha sugerido previamente en linfoma de células T/NK[504].

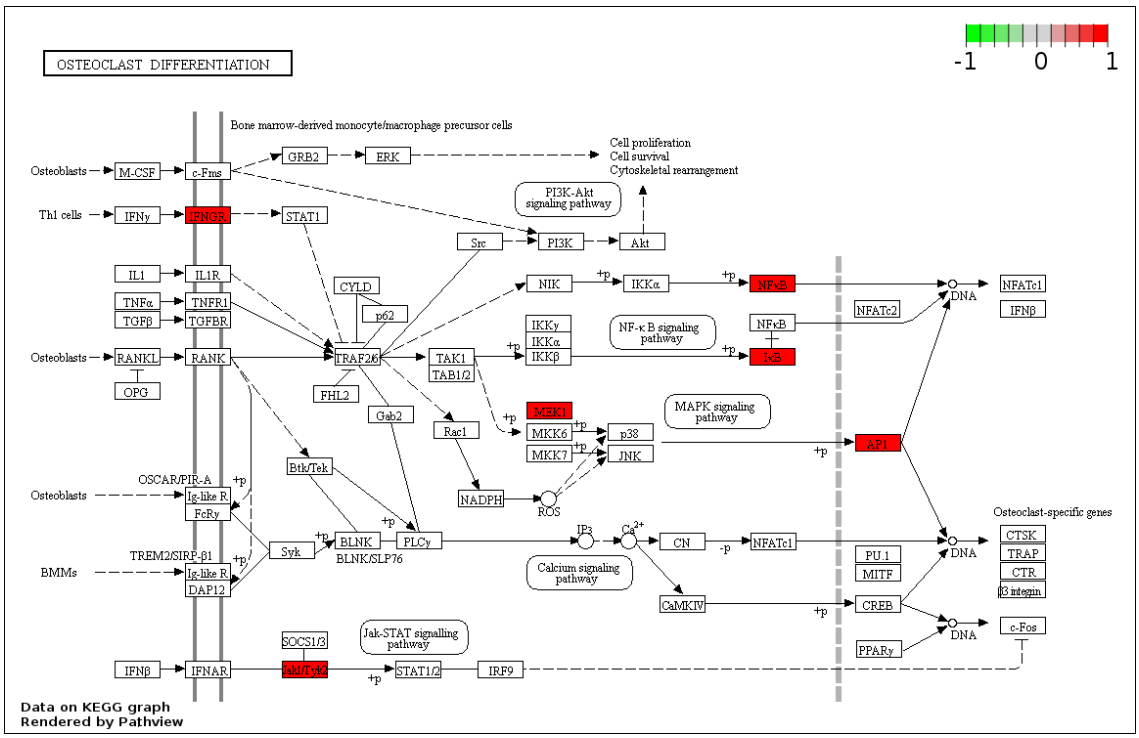



***Figura 4.121.*** *Vía de la diferenciación de osteoclastos según la base KEGG. En rojo se representan los genes sobreexpresados de forma estadísticamente significativa en el metaanálisis global de la decitabina.*

Otra de las vías desreguladas por la decitabina fue la "diferenciación de células Th1 y Th2" (FDR = 0,0296) (**Figura 4.122**). Entre los genes desregulados en esta vía aparecen sobreexpresados los genes que codifican la cadena beta del receptor del interferón gamma (*IFNGR2*, *z*-valor = 3,94, *p*-valor < 0,0001) y la subunidad gamma del receptor de interleucina 2 (*IL2RG*, *z*-valor = 11,96, *p*-valor < 0,0001). Estudios previos con decitabina en pacientes oncológicos han demostrado que este compuesto es un gran potenciador de la producción de células T productoras de interferón gamma (IFN-γ)[505], que es una citocina con un potente efecto inhibidor de la proliferación celular en MM[506]. Entre las células T productoras de IFN-γ se encuentran los linfocitos Th1 que también son productoras de interleucina 2 (IL-2)[507], cuyos altos niveles en suero de pacientes con MM han sido asociados con una mayor supervivencia[508]. Por tanto, la sobreexpresión de los receptores tanto de IFN-γ como de IL-2 en las células plasmáticas de MM, debida a la desmetilación de sus promotores por la decitabina, podría favorecer la acción de estas dos

citocinas promoviendo un efecto antimieloma como respuesta al tratamiento con este compuesto.

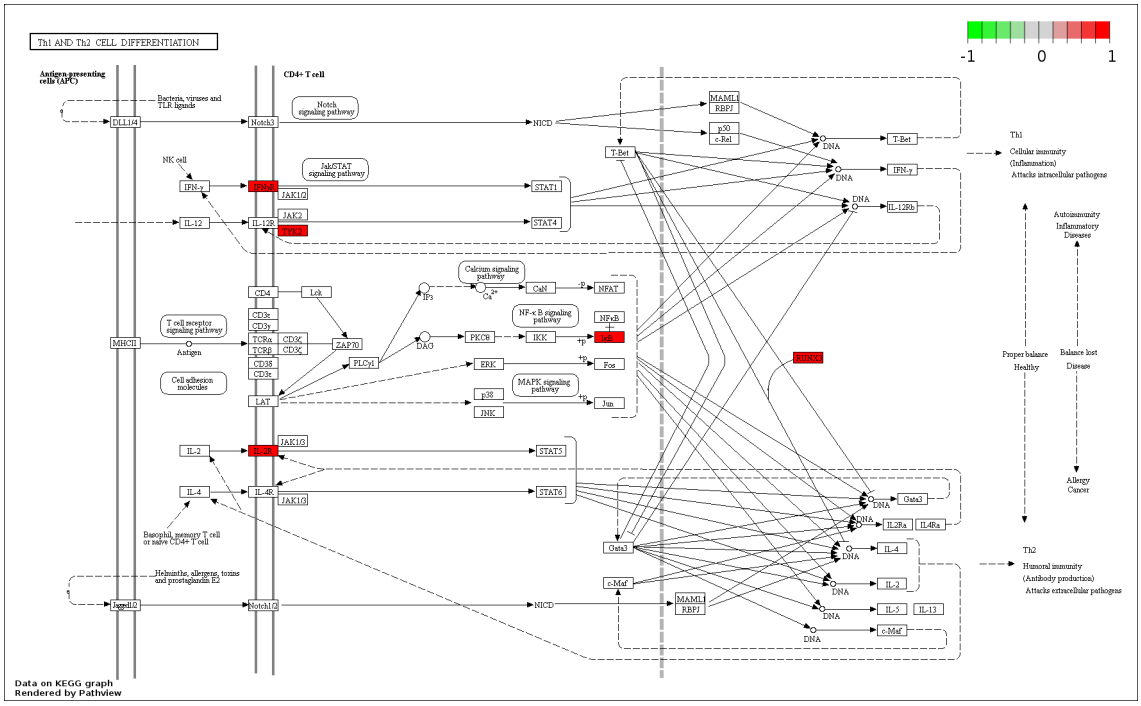



***Figura 4.122.*** *Vía de la diferenciación de células Th1 y Th2 según la base KEGG. En rojo se representan los genes sobreexpresados de forma estadísticamente significativa en el metaanálisis global de la decitabina.*

### 4.3.9. JQ1

El último compuesto sobre el que se procedió a la revisión sistemática con metaanálisis fue el inhibidor de bromodominio JQ1. Este compuesto actualmente no es empleado en el tratamiento en pacientes de MM, sin embargo, se decidió su inclusión en el presente trabajo debido al amplio número de experimentos sobre HMCLs que existen reportados en la bibliografía. Se realizó una búsqueda sistemática de estudios en repositorios *online* de HMCLs tratadas con JQ1 y se localizaron 115 series en GEO, 90 series en ArrayExpress y 254 muestras correspondientes a 26 series en SRA. Tras la eliminación de los elementos duplicados en los tres repositorios se seleccionaron para su revisión detallada 132 series. De estas 132 series revisadas, tres cumplieron los criterios de inclusión y exclusión necesarios para ser incluidas en el metaanálisis. La comprobación de la posible subdivisión de las series en diferentes estudios en función de las concentraciones de fármaco empleadas, el tiempo de tratamiento o la utilización de varias líneas celulares, determinó que las series GSE31365 y GSE44929 fuesen divididas en tres y 7 estudios, respectivamente. De esta manera, el número final de estudios considerados para el metaanálisis fue de 11. El diagrama de flujo que se muestra en la

**Figura 4.123**, detalla el esquema de selección de estudios para el metaanálisis del compuesto JQ1 en función de los diferentes criterios de inclusión y exclusión.

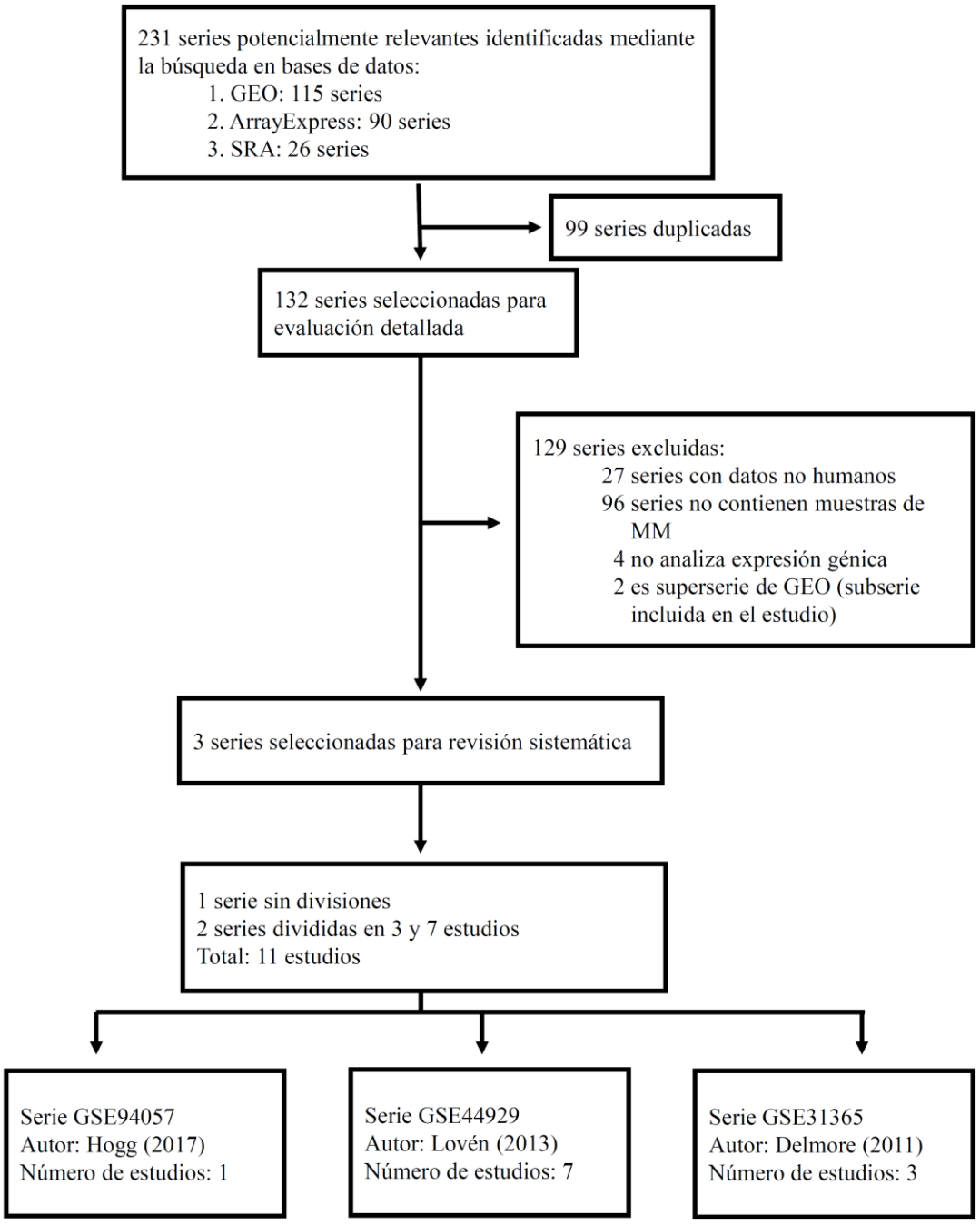



***Figura 4.123**. Diagrama de flujo del proceso de selección de estudios incluidos en el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con JQ1.*

A continuación estos 11 estudios seleccionados se clasificaron en subgrupos en función de la mediana ± MAD de los tiempos de tratamiento y de la concentración aplicada de JQ1. Se establecieron los tiempos a dos y a 10 horas (mediana de seis horas) como puntos de corte del tiempo de tratamiento. Para la concentración únicamente se

dispuso de un punto de corte en 500 nM debido a que la MAD fue cero. Los resultados del agrupamiento en subgrupos se recogen en la **Tabla 4.12**.

***Tabla 4.12.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con JQ1.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (nM) |
|---|---|---|---|---|---|---|
| GSE94057 | Hogg (2017)[509] | ALF1 | Illumina HiSeq 2500 | 4 | 2 | 1000 |
| GSE44929 | Lovén (2013a)[192] | MM1-S | Affymetrix PrimeView | 4 | 6 | 5 |
| GSE44929 | Lovén (2013b)[192] | MM1-S | Affymetrix PrimeView | 4 | 6 | 50 |
| GSE44929 | Lovén (2013c)[192] | MM1-S | Affymetrix PrimeView | 4 | 6 | 500 |
| GSE44929 | Lovén (2013d)[192] | MM1-S | Affymetrix PrimeView | 4 | 6 | 5000 |
| GSE44929 | Lovén (2013e)[192] | MM1-S | Affymetrix PrimeView | 4 | 0.5 | 500 |
| GSE44929 | Lovén (2013f)[192] | MM1-S | Affymetrix PrimeView | 4 | 1 | 500 |
| GSE44929 | Lovén (2013g)[192] | MM1-S | Affymetrix PrimeView | 4 | 3 | 500 |
| GSE31365 | Delmore (2011a)[155] | MM1-S | Affymetrix Human Gene 1.0ST | 4 | 24 | 500 |
| GSE31365 | Delmore (2011b)[155] | KMS11 | Affymetrix Human Gene 1.0ST | 4 | 24 | 500 |
| GSE31365 | Delmore (2011c)[155] | OPM1 | Affymetrix Human Gene 1.0ST | 4 | 24 | 500 |

*En verde, estudios seleccionados para el subgrupo de tiempos o concentraciones bajos; en amarillo, estudios seleccionados para el subgrupo de tiempos o concentraciones intermedias; en rojo, estudios seleccionados para el subgrupo de tiempos o concentraciones altas.*

Tras establecer los subgrupos se procedió a seleccionar los genes cuyo valor absoluto del FC fuese mayor a 1,5 en los 11 estudios o, al menos, en todos los estudios de uno de los subgrupos de tiempo o concentración, excluyendo los subgrupos que solamente constasen de un estudio. De este modo, se consideraron 963 genes como candidatos para los metaanálisis. La **Figura 4.124** muestra la distribución de los ln(FC) de los 963 genes en los 11 estudios seleccionados. Puede observarse la baja dispersión de los ln(FC) en los casos donde, o bien los tiempos de tratamiento (Lóven [2013e] y Lóven [2013f]), o bien las concentraciones de JQ1 (Lovén [2013a]) son extremadamente reducidos. La influencia de estos dos factores será determinada en los siguientes apartados a través del metaanálisis por subgrupos.

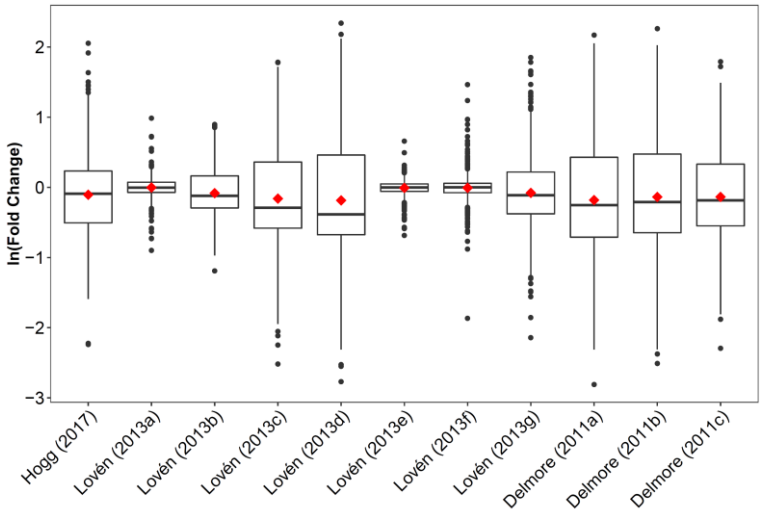
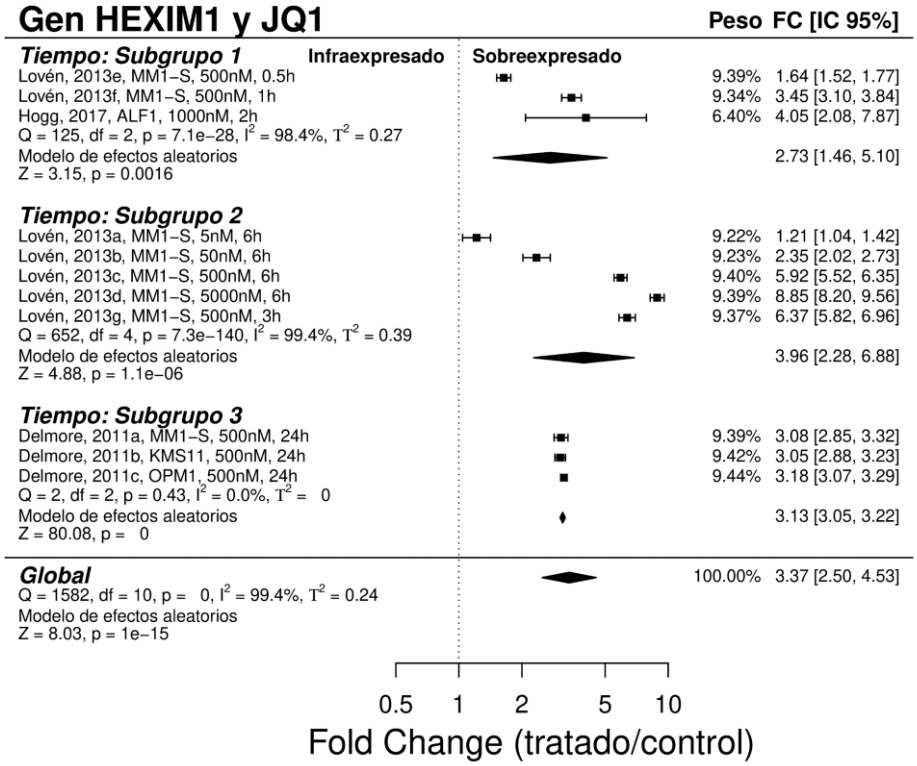



***Figura 4.124.*** *Diagrama de caja (box plot) del ln(Fold Change) de los 963 genes seleccionados para el metaanálisis de JQ1 en monoterapia en líneas celulares de MM. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

#### 4.3.9.1. Análisis por subgrupos: tiempo de tratamiento

Se establecieron tres subgrupos considerando el tiempo de tratamiento con JQ1. En un primer subgrupo (G1) se agruparon los estudios realizados a un tiempo menor o igual a dos horas, de modo que tres de los estudios cumplieron este criterio: Hogg (2017), Lovén (2013e) y Lovén (2013f). El segundo subgrupo (G2) estuvo formado por los cinco estudios cuyos tiempos de tratamiento fueron superiores a las dos horas, pero inferiores o iguales a las 10 horas: estudios (a-d) y (g) de Lovén (2013). El tercer subgrupo comprendió los tres estudios de la serie GSE31365, ya que fueron los únicos que tuvieron tiempos de tratamiento superiores a las 10 horas (estudios [a-c] de Delmore [2010]). Mediante el metaanálisis por subgrupos de tiempo fueron determinados 222 genes con *p*-valor < 0,05 en el subgrupo G1, 839 genes en el caso del subgrupo G2 y 763 genes en el subgrupo G3 (**Anexo 31**). En total, 154 de estos genes fueron comunes a los tres subgrupos, de los que 38 presentaron sobreexpresión y 65 infraexpresión al tratar con JQ1; el resto presentaron sentidos de expresión opuestos en al menos dos de los subgrupos. El diagrama de bosque del gen sobreexpresado con mayor valor de FC y del gen infraexpresado con menor valor de FC aparecen representados en la **Figura 4.125**.

**a**




**b**

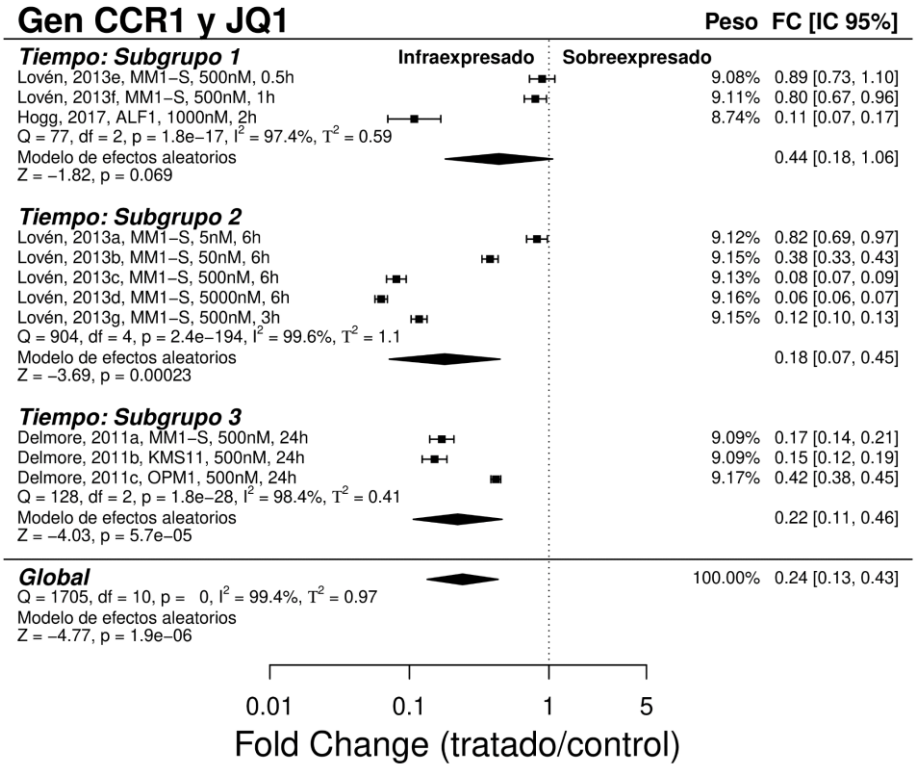



***Figura 4.125.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de tiempo de tratamiento con JQ1.* ***a)*** *Diagrama de bosque del gen HEXIM1, que fue el más sobreexpresado considerando la mediana del FC de los 11 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen CCR1, que fue el más infraexpresado considerando la mediana del FC de los 11 estudios seleccionados.*

A continuación, se comparó la expresión de los genes entre los tres subgrupos para determinar la influencia del tiempo de tratamiento. De este modo, se obtuvo que 568 genes presentaron diferencias estadísticamente significativas entre los subgrupos G1 y G2, 640 genes entre los subgrupos G1 y G3, y 450 genes entre los subgrupos G2 y G3 (**Anexo 31**). El análisis comparativo de cada uno de los subgrupos se muestra mediante diagrama de puntos en la **Figura 4.126**, donde se observa que el cambio en el ln(FC) inducido por tiempo de tratamiento aplicado el subgrupo G1 fue menor que el que produjeron los tiempos de tratamiento de los subgrupos G2 y G3, ya que los valores de ln(FC) de estos genes en G1 aparecen situados alrededor de cero.

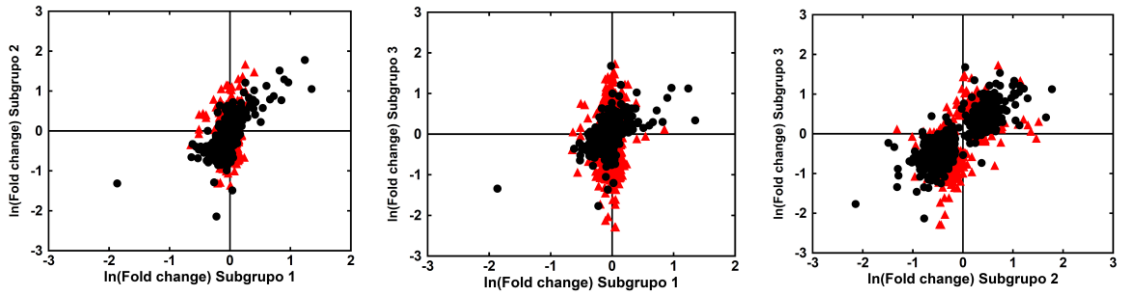



***Figura 4.126.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 963 genes estudiados donde se comparan los subgrupos de tiempo de tratamiento 1, 2 y 3. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

El análisis ORA sobre los genes que presentaron diferencias entre los subgrupos se recoge en la **Figura 4.127**. El objetivo de este análisis fue la determinación de vías o funciones biológicas que pudieran ser afectadas por el tiempo de tratamiento con el compuesto JQ1.

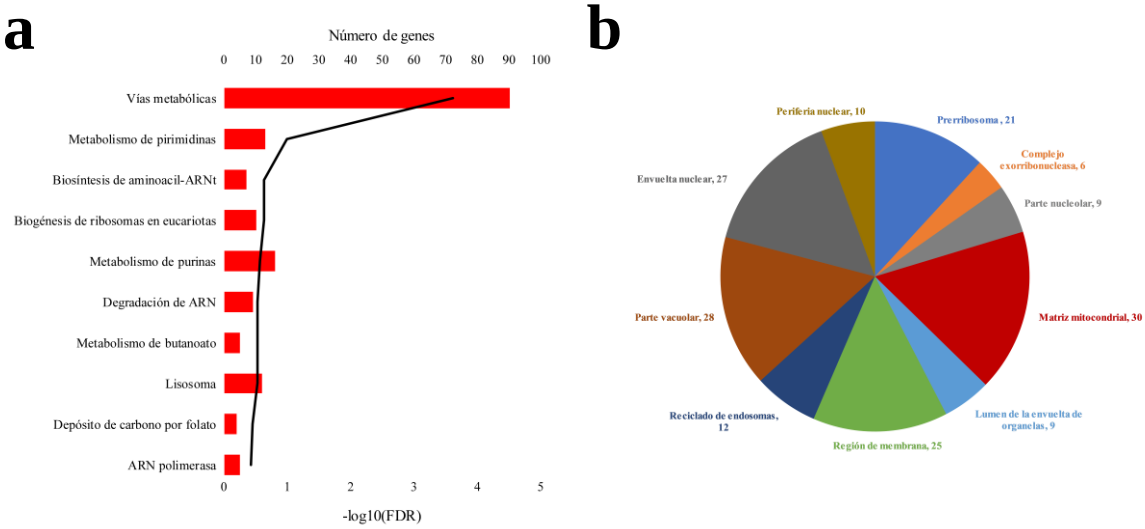



***Figura 4.127.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con JQ1. En esta figura se recogen las 10 rutas KEGG y las10 funciones GO con un menor valor de FDR.* ***a)*** *Análisis de rutas biológicas KEGG,* ***b)*** *análisis de localización GO.*

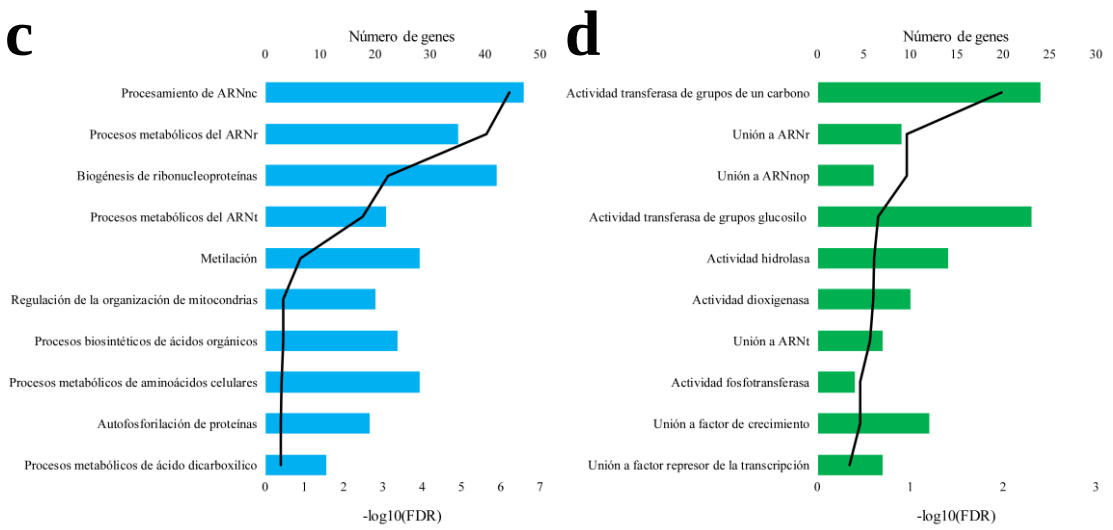



***Figura 4.127 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de tiempo de tratamiento con JQ1. En esta figura se recogen las 10 rutas KEGG y las10 funciones GO con un menor valor de FDR.* ***c)*** *Análisis de procesos biológicos GO y* ***d)*** *análisis de funciones moleculares GO.*

Una única vía KEGG ("vías metabólicas"), y varios términos GO, tanto a nivel de PB, ("procesamiento de ARNnc" o "biogénesis de ribonucleoproteínas"), como a nivel de FM ("actividad transferasa de grupos de un carbono"), resultaron estadísticamente significativos a FDR < 0,05. Para determinar el efecto del tiempo de tratamiento seleccionamos los 90 genes desregulados correspondientes a la vía KEGG "vías metabólicas. En la mayor parte del conjunto de genes estudiados se observa una tendencia a que la desregulación de la expresión génica producida por JQ1 se incrementa a medida que aumenta el tiempo de tratamiento, de manera que el mayor cambio del ln(FC) aparece en el subgrupo G3, observándose este incremento tanto en los genes sobreexpresados como en los infraexpresados (**Figura 4.128**). Por tanto, el tiempo de tratamiento con JQ1 parece ser en general un factor determinante en la modulación de la expresión de los genes diana de este compuesto.

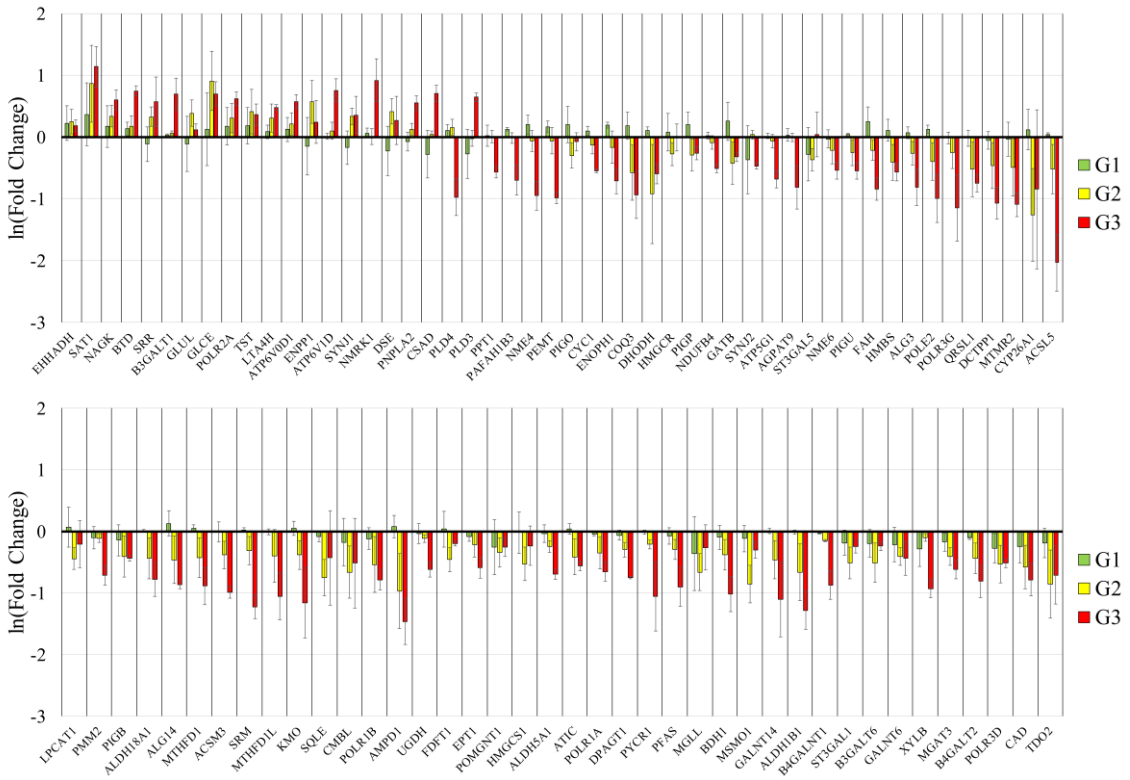
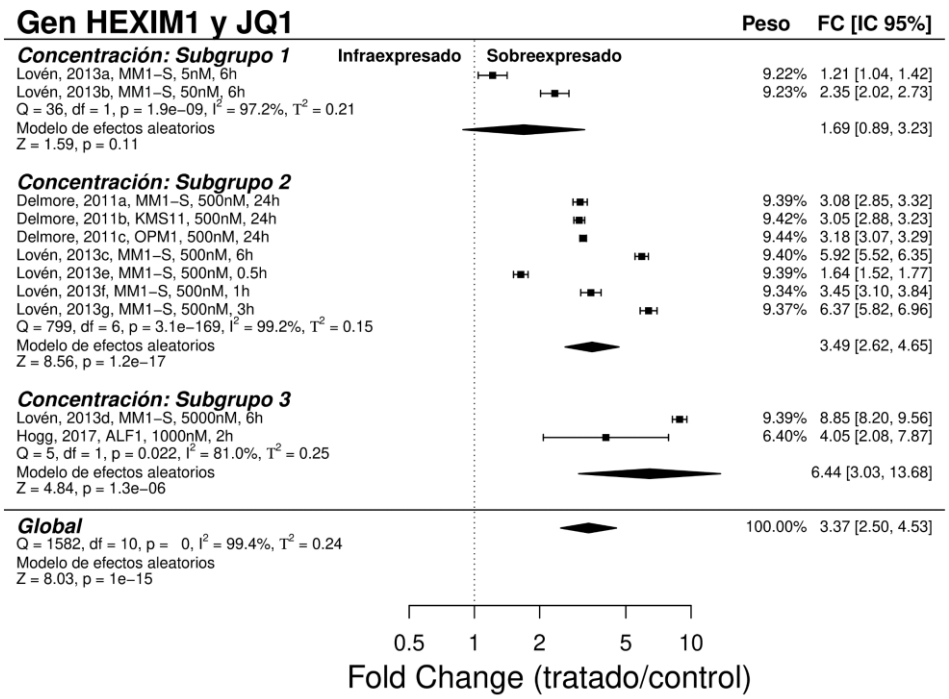



***Figura 4.128.*** *Valores promedio del ln(Fold Change) de los genes desregulados en "vías metabólicas" KEGG en los tres subgrupos de tiempo de tratamiento con JQ1 (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.9.2. Análisis por subgrupos: concentración

Al igual que en el caso del metaanálisis de tiempo de tratamiento, el metaanálisis por subgrupos de concentración de JQ1 también constó de tres subgrupos. En el primer subgrupo (G1) se agruparon los estudios con una concentración de JQ1 menor de 500 nM: Lovén (2103a) y Lovén (2013b). En lo que respecta al segundo subgrupo (G2), comprendió los 7 estudios realizados a una concentración de 500 nM: estudios (c) y (e-g) de Lovén (2013) y estudios (a-c) de Delmore (2011). Finalmente, el tercer subgrupo (G3) recogió los dos estudios de altas concentraciones, superiores a 500 nM: Hogg (2017) y Lovén (2013d). Mediante el metaanálisis por subgrupos fueron determinados 230 genes con *p*-valor < 0,05 en el subgrupo G1, 803 genes en el caso del subgrupo G2 y 503 genes en el subgrupo G3 (**Anexo 32**). El cruce de las tres listas reveló 102 genes comunes a los tres subgrupos, de los que 42 presentaron sobreexpresión y 59 infraexpresión al tratar con JQ1. En la **Figura 4.129** se representa el diagrama de bosque de los genes *HEXIM1* y *CCR1*, los genes con mayor y menor valor de FC considerando la mediana de los 11 estudios considerados para este metaanálisis, respectivamente.

**a**




**b**

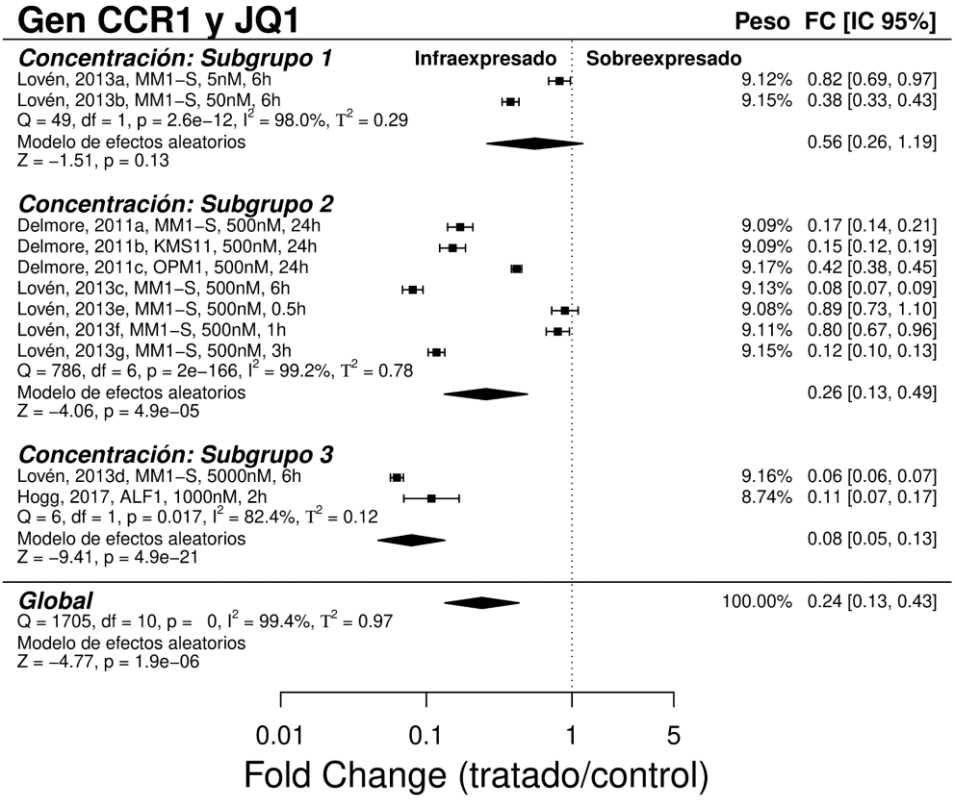



***Figura 4.129.*** *Diagrama de bosque (forest plot) del metaanálisis de efectos aleatorios por subgrupos de concentración de JQ1.* ***a)*** *Diagrama de bosque del gen HEXIM1, que fue el más sobreexpresado considerando la mediana del FC de los 11 estudios seleccionados.* ***b)*** *Diagrama de bosque del gen CCR1, que fue el más infraexpresado considerando la mediana del FC de los tres estudios seleccionados.*

La comparación de los resultados del metaanálisis entre los tres subgrupos obtuvo que 220 genes presentaron diferencias estadísticamente significativas entre los subgrupos G1 y G2, 363 genes entre los subgrupos G1 y G3, y 250 genes entre los subgrupos G2 y G3 (**Anexo 32**). En la **Figura 4.130** se muestran los genes diferenciales entre cada par de subgrupos (en rojo) respecto al total de genes (en negro). La mayor parte de los cambios observados en estos genes consistieron en un incremento o una disminución de la intensidad de la expresión; el número de genes en los que se observó un cambio de sentido de expresión entre dos de los grupos fue escaso.

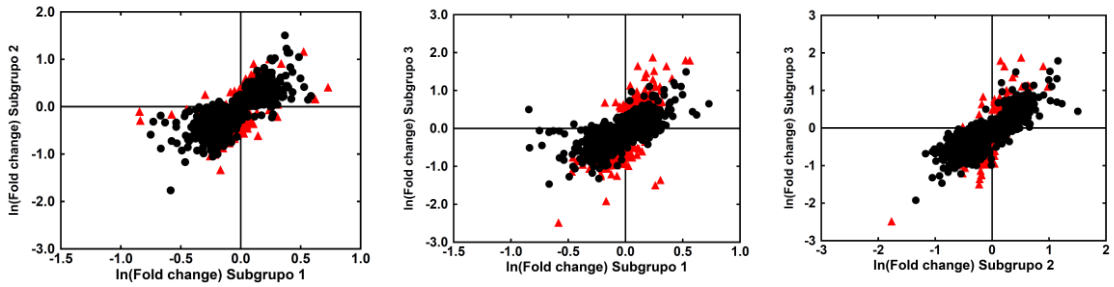



***Figura 4.130.*** *Diagrama de puntos de los valores de ln(FC) obtenidos para los 963 genes estudiados donde se comparan los subgrupos de concentración de JQ1. En rojo se muestran los genes que mostraron diferencias estadísticamente significativas entre cada par de subgrupos.*

El análisis ORA para la determinación de las vías biológicas KEGG o funciones GO afectadas por la concentración de JQ1 sobre los genes con diferencias estadísticamente significativas entre, al menos, dos de los subgrupos, no reveló ningún resultado estadísticamente significativo a FDR < 0,05. Los resultados de este análisis se muestran en la **Figura 4.131**.

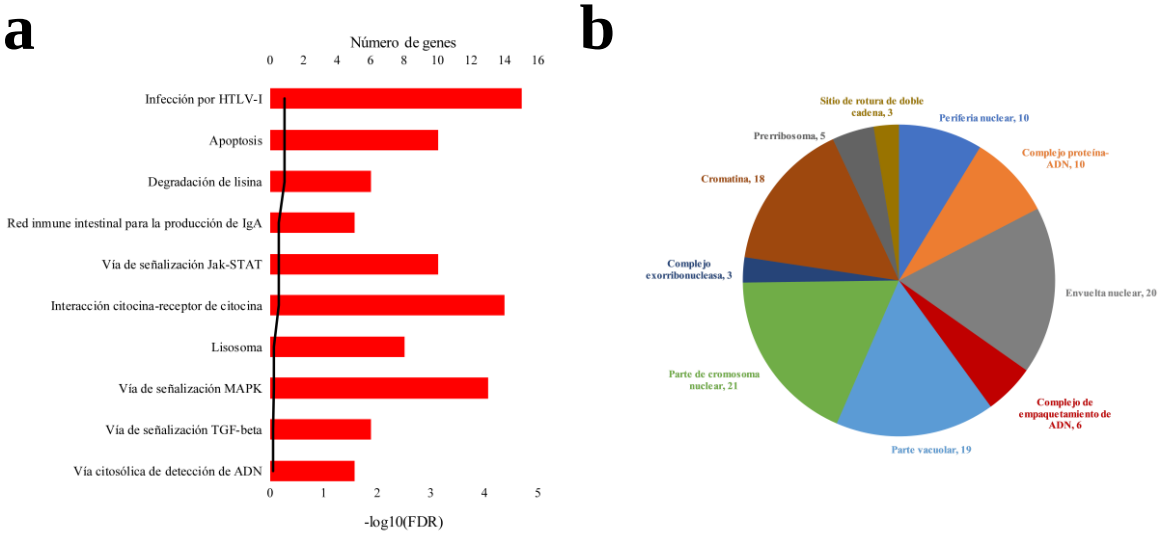



***Figura 4.131.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de JQ1. En cada panel se recogen las 10 rutas KEGG y los 10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

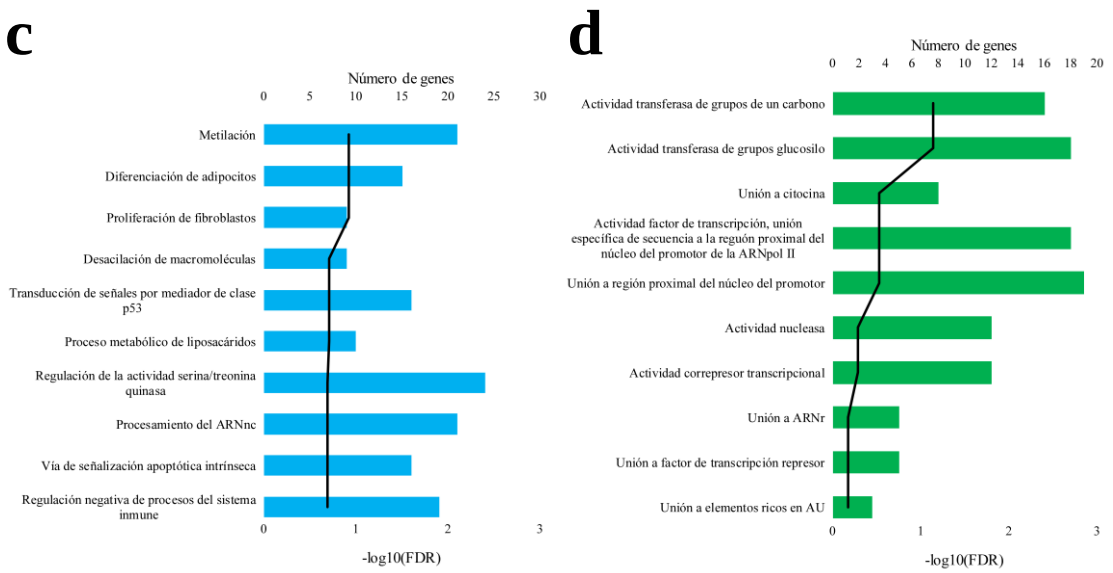



***Figura 4.131 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas entre los subgrupos de concentración de JQ1. En cada panel se recogen las 10 rutas KEGG y los 10 términos GO con un menor valor de FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

A pesar de que no hubo resultados significativos, se decidió comprobar la influencia de la concentración de JQ1 sobre los genes de una de las vías KEGG que podrían ser más representativas del mecanismo de acción de este compuesto: la vía KEGG "apoptosis". El efecto que se observó sobre estos genes fue que, en general, las concentraciones de los subgrupos G2 y G3, y principalmente del G3, produjeron un incremento del cambio de la expresión génica mayor que las menores concentraciones aplicadas en el subgrupo G1. (**Figura 4.132**). Por tanto, al igual que ocurriese con el tiempo de tratamiento, parece que la concentración aplicada de JQ1 también influiría en el grado de desregulación de la expresión de los genes diana de este compuesto.

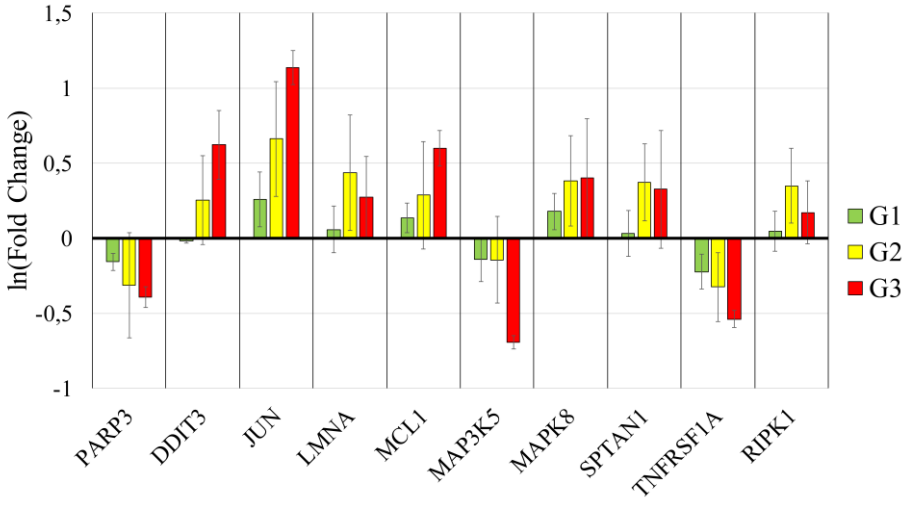



***Figura 4.132.*** *Valores promedio del ln(Fold Change) de los genes desregulados en la vía de apoptosis en los tres subgrupos de concentración de JQ1 (G1, G2 y G3). Las barras de error representan la desviación estándar del ln(Fold Change).*

#### 4.3.9.3. Metaanálisis global de JQ1

El análisis global considerando los 11 estudios reveló una diferencia en el tamaño del efecto sobre la expresión génica estadísticamente significativa a *p*-valor < 0,05 de 870 genes, de los que 316 presentaron sobreexpresión y 554 infraexpresión en las muestras tratadas con JQ1 (**Anexo 33**). El análisis ORA sobre los 870 genes se muestra en la **Figura 4.133**.

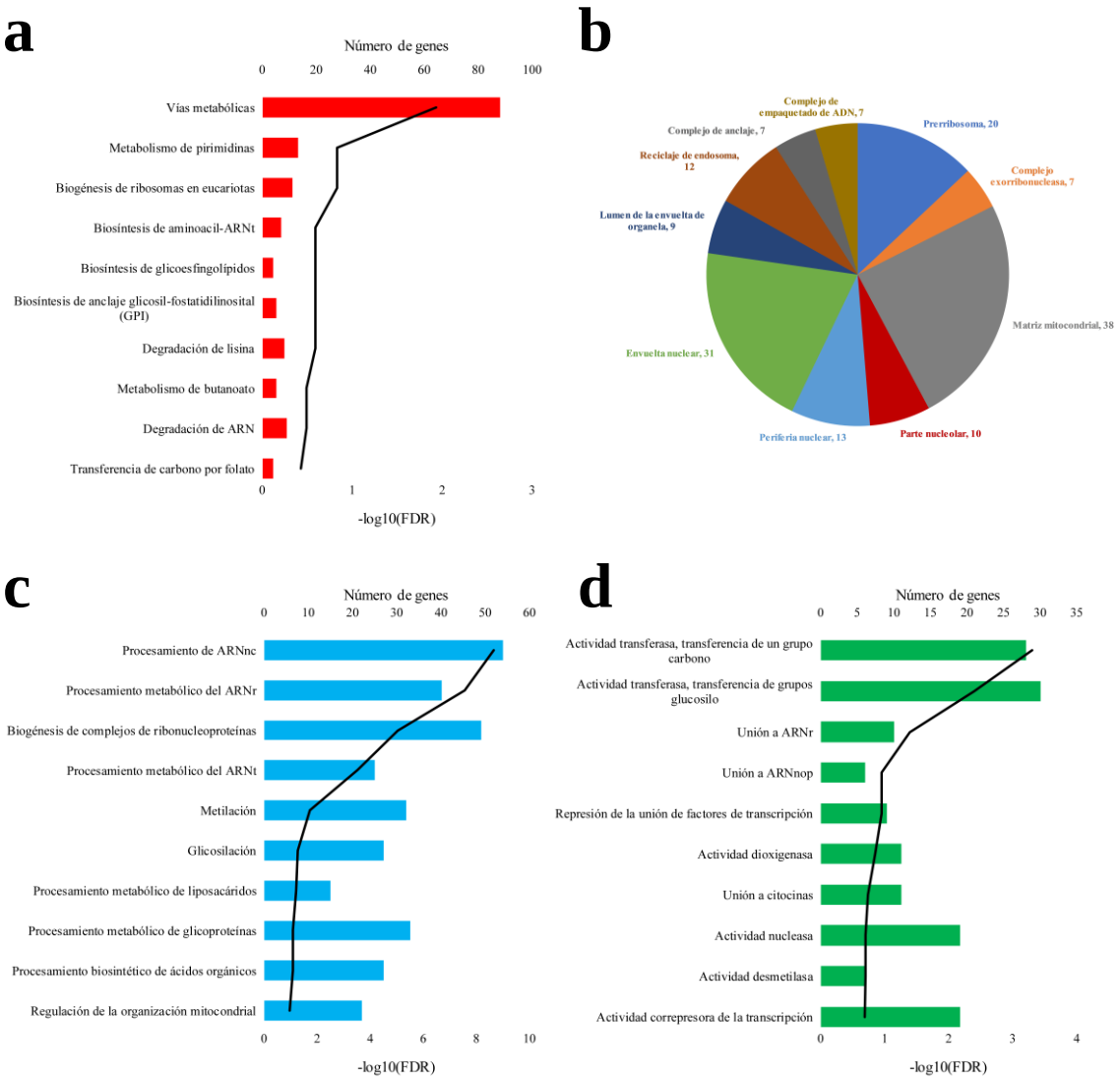



***Figura 4.133.*** *Análisis de sobrerrepresentación sobre vías KEGG y términos GO considerando los 870 genes con una diferencia del tamaño del efecto estadísticamente significativa en el metaanálisis de la expresión génica de líneas celulares tratadas con JQ1. En cada panel se recogen las 10 rutas KEGG o los10 términos GO con menor FDR.* ***a)*** *TOP 10 vías biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Con 88 genes desregulados, la ruta KEGG "vías metabólicas" fue la vía con mayor sobrerrepresentación en número de genes para el compuesto JQ1. Sin embargo, esta vía implica multitud de funciones relacionadas con el metabolismo general de la célula, lo que hace que sea demasiado amplia como para hacer una descripción detallada de los

procesos que pueden estar alterados en relación con la misma. Por este motivo, se decidió realizar el estudio de vías KEGG sobre las dos vías siguientes en significancia, que a pesar de no superar el umbral del FDR de 0,05, los genes recogidos en ellas se asociaron directamente con otras vías o PB cuya sobrerrepresentación sí fue significativa. La primera de estas dos vías fue el "metabolismo de pirimidinas" (FDR = 0,1472) (**Figura 4.134**).

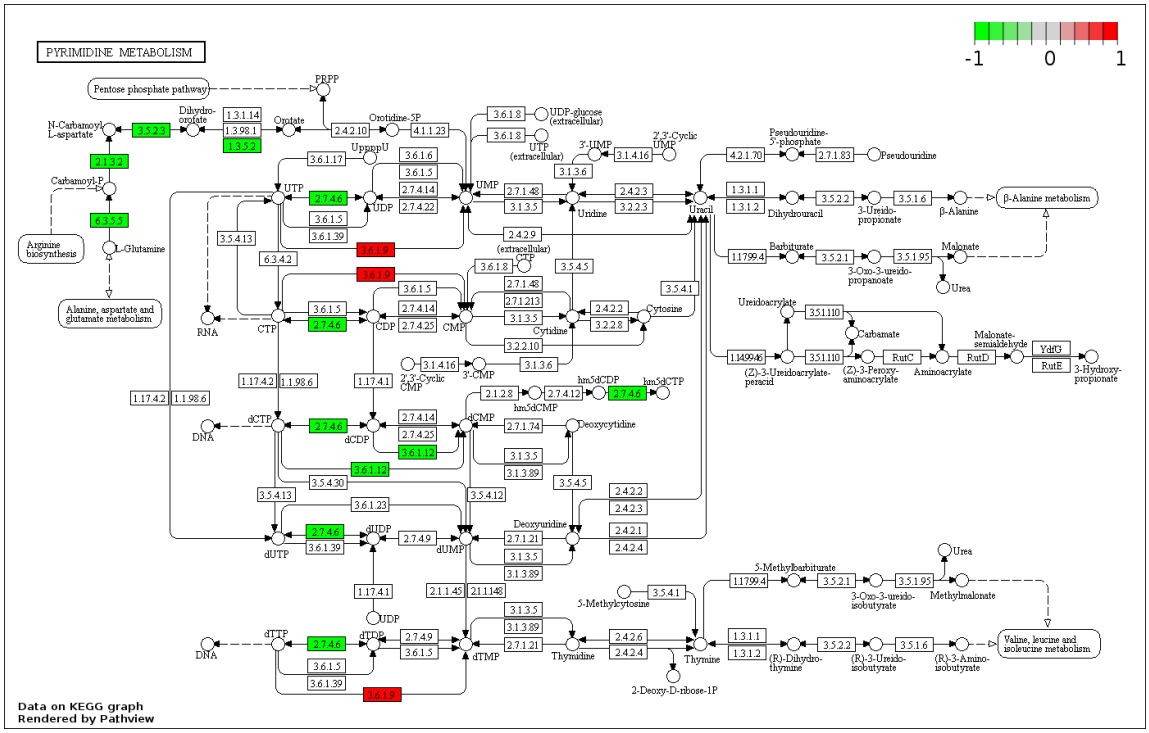



***Figura 4.134.*** *Vía del metabolismo de pirimidinas según la base KEGG. En verde se representan los genes infraexpresados y en rojo los sobreexpresados de forma estadísticamente significativa en el metaanálisis global del JQ1.*

Esta vía ha sido reportada como potencialmente afectada en el MM y, por tanto, su modulación sería de posible interés terapéutico[53]. En nuestro trabajo, fueron detectados 13 genes desregulados en esta vía, dos presentaron sobreexpresión y 11 infraexpresión al tratamiento. Resulta de especial interés la elevada proporción de genes que codifican distintas subunidades de ARN polimerasas, de manera que aparecen infraexpresados dos genes asociados a la ARN polimerasa I (*POLR1A* y *POLR1B*), sobreexpresado un gen asociado a la ARN polimerasa II (*POLR2A*) e infraexpresados otros dos genes asociados a la ARN polimerasa III (*POLR3D* y *POLR3G*). La infraexpresión de elementos asociados a las ARN polimerasas I y III, implicadas en la transcripción de ARNr podría estar asociada además con la segunda de las vías KEGG seleccionadas: la "biosíntesis de ribosomas en eucariotas" (FDR = 0,1472). Esta segunda vía estaría desregulada negativamente por el compuesto JQ1, ya que observamos que este compuesto reduce de manera significativa la expresión de 10 de sus componentes (**Figura 4.135**). Por lo tanto, podría esperarse que la biogénesis de ribosomas estuviese regulada a nivel transcripcional inhibiendo la síntesis tanto del ARNr, como de varios de los componentes de esta vía.

Este dato se confirmaría a través del análisis de los genes implicados en los PB "procesamiento metabólico del ARNr" (FDR < 0,0001) y "biogénesis de complejos de ribonucleoproteínas" (FDR < 0,0001), donde encontramos 47 y 39 genes regulados de manera negativa, respectivamente. La regulación de la biogénesis de ribosomas por JQ1 podría ser en parte ejercida a través de la infraexpresión de *MYC* (*z*-valor = -5,69, *p*-valor < 0,0001), cuyo efecto regulador ha sido demostrado a varios niveles y sobre diferentes elementos de esta vía[458]. El resultado final de todo el proceso conduciría a la inhibición del programa traduccional de la célula tumoral, lo que se ha demostrado que puede ser una estrategia terapéutica efectiva contra el MM[510].

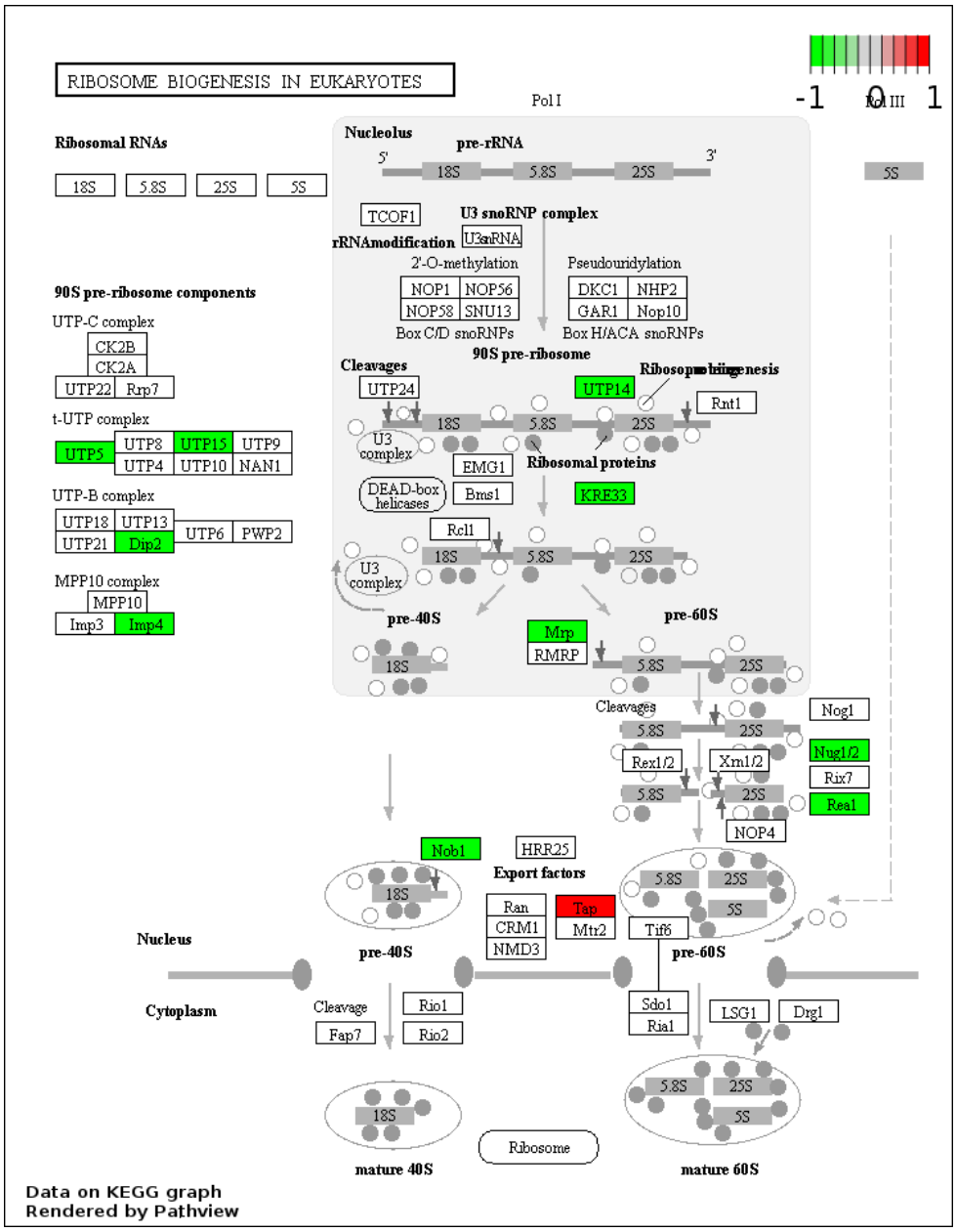



***Figura 4.135.*** *Vía de la biogénesis de ribosomas en eucariotas según la base KEGG. En verde se representan los genes infraexpresados de forma estadísticamente significativa en el metaanálisis global de JQ1, mientras en rojo se representan los genes sobreexpresados de manera estadísticamente significativa en ese mismo metaanálisis.*

### 4.3.10. Análisis del sesgo de publicación

Se analizó la posibilidad de la existencia de sesgos de publicación en cada uno de los compuestos analizados mediante metaanálisis utilizando la prueba de regresión de Egger. Aunque en todos los metaanálisis se detectó un sesgo de publicación estadísticamente significativo ($p < 0,05$) en algunos de los genes estudiados, en ningún caso supuso un porcentaje superior o igual al 50% del total de genes analizados (**Tabla 4.13)**. Por este motivo, el sesgo de publicación no fue considerado como representativo del conjunto de genes para ninguno de los 9 fármacos estudiados. Los resultados de las regresiones de Egger individuales de cada uno de los genes estudiados en este trabajo aparecen en los correspondientes **Anexos** en los que se recoge el metaanálisis global de cada uno de los 9 compuestos analizados.

***Tabla 4.13.*** *Resultados del análisis del sesgo de publicación en los 9 fármacos en los que se llevó a cabo la revisión sistemática con metaanálisis.*

| Fármaco | # genes analizados | Genes p < 0,05 en regresión de Egger | Porcentaje (%) |
|---|---|---|---|
| **Melfalán** | 1460 | 390 | 26,71 |
| **Dexametasona** | 141 | 15 | 10,64 |
| **Bortezomib** | 863 | 242 | 28,04 |
| **Lenalidomida** | 1164 | 351 | 30,15 |
| **Pomalidomida** | 212 | 56 | 26,42 |
| **Panobinostat** | 2056 | 409 | 19,89 |
| **Azacitidina** | 145 | 28 | 19,31 |
| **Decitabina** | 123 | 14 | 11,38 |
| **JQ1** | 963 | 189 | 19,63 |

*El análisis del sesgo de publicación se realizó mediante la regresión de Egger*

### 4.3.11. Otros fármacos

En este apartado se incluyen los estudios de expresión génica sobre fármacos que, o bien no pudieron ser incluidos en el metaanálisis debido a la carencia de estudios que cumpliesen los criterios de inclusión, como fue el caso del interferón γ (**Anexo 9**), o bien los únicos estudios de que se disponía fueron trabajos del servicio de Hematología de Salamanca, como fue el caso de los compuestos moduladores del *splicing* alternativo amilorida y TG003.

#### 4.3.11.1. Amilorida

El efecto de la amilorida sobre la expresión génica fue evaluado en este trabajo mediante técnicas de metaanálisis. Para ello, se dispuso de datos de expresión génica de RNA-seq en las HMCLs KMS12-BM y JJN-3 depositadas en GEO bajo el identificador GSE95077. Estos datos fueron utilizados previamente para la determinación del *pipeline*

de análisis de RNA-seq en el **Capítulo 1**. Los dos estudios en los que fue dividida esta serie aparecen recogidos en la **Tabla 4.14**.

***Tabla 4.14.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con amilorida.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (mM) |
|---|---|---|---|---|---|---|
| GSE95077 | Rojas (2017a)[169] | KMS12-BM | Illumina HiSeq 2500 | 6 | 24 | 0,1 |
| GSE95077 | Rojas (2017b)[169] | JJN-3 | Illumina HiSeq 2500 | 6 | 24 | 0,1 |

En un siguiente paso se procedió a la selección de los genes candidatos para llevar a cabo el metaanálisis. Como resultado se obtuvieron 786 genes comunes, para los que la distribución de sus ln(FC) es la que aparece en la **Figura 4.136**, en la que se observa un mayor cambio de expresión, tanto a nivel de sobreexpresión como de infraexpresión, en el estudio Rojas (2017a), correspondiente a la línea celular KMS12-BM.

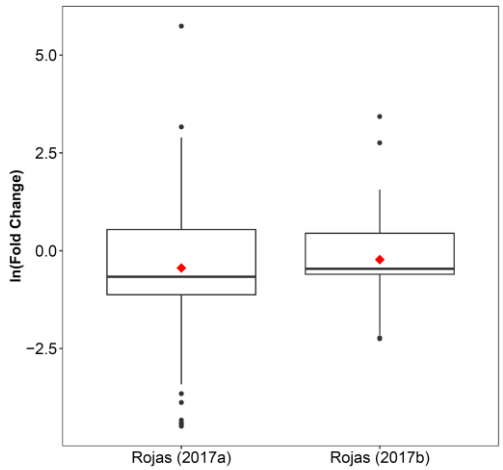



***Figura4.136.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC])de los 786 genes seleccionados para el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con amilorida. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

A continuación, se procedió al estudio con metaanálisis de los 786 genes previamente seleccionados, revelando diferencias estadísticamente significativas a *p*-valor < 0,05 en el tamaño del efecto en 675 genes (**Anexo 34**). En cuanto a la dirección del cambio de expresión, 212 genes presentaron sobreexpresión, mientras que 463 genes mostraron infraexpresión al tratamiento con la amilorida. En la **Figura 4.137** se representa el análisis ORA de estos genes significativos considerando como fuente de vías y funciones biológicas las bases KEGG y GO.

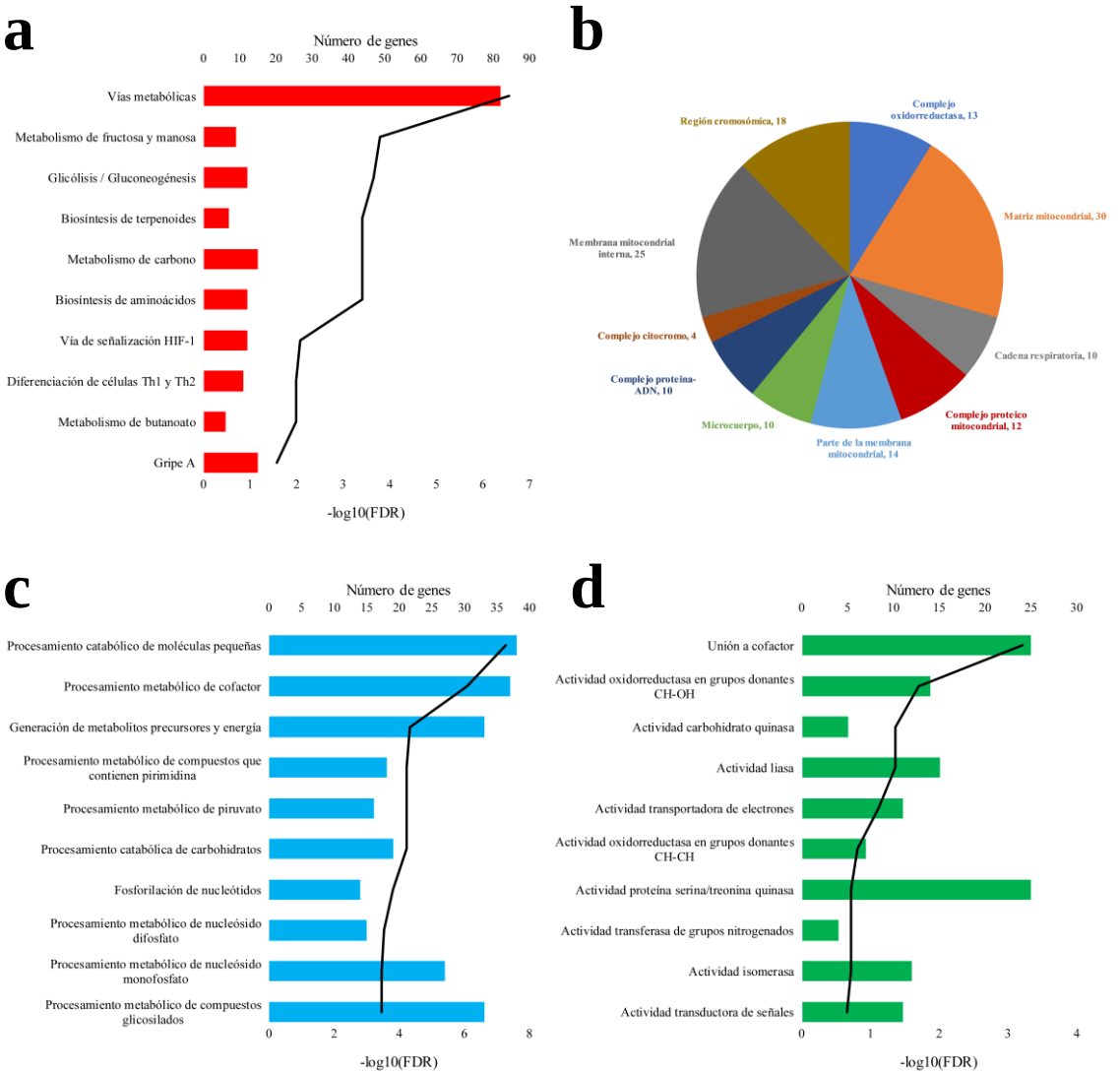



***Figura 4.137.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas en el metaanálisis global para el tratamiento con amilorida. En esta figura se recogen las 10 rutas KEGG y los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Este análisis ORA muestra una gran cantidad de vías y términos sobrerrepresentados de manera estadísticamente significativa relacionados con procesos metabólicos. La implicación de los procesos metabólicos en el desarrollo del cáncer se ha visto como un proceso crucial, ya que las células cancerígenas tienen un metabolismo especial que puede promover la proliferación y el escape de la vía de muerte celular, así como favorecer la producción de metabolitos que lleven a la oncogénesis[511]. En el MM, la investigación de la implicación del metabolismo en el desarrollo de la enfermedad viene aumentando de forma gradual en los últimos años. En particular, la glucosa y la glutamina son los componentes metabólicos más estudiados en las células de MM[512]. Estudios recientes en MM también han demostrado la participación de genes asociados al metabolismo, como *LDHA* y *HIF1A*, en el desarrollo de resistencia a fármacos, por lo que algunos autores proponen la necesidad de desarrollar nuevas drogas cuya diana sea

la regulación de las vías metabólicas[513]. En este sentido, la amilorida podría ser un compuesto de interés en el tratamiento del MM, ya que su capacidad reguladora de vías metabólicas ha sido descrita previamente en vías como el ciclo de Krebs[514] o la síntesis de proteínas[515], lo que, unido a los resultados obtenidos en este trabajo, confirmaría el potencial efecto modulador de la amilorida sobre el metabolismo. Una de las vías desreguladas asociadas al metabolismo, con un interés particular por lo expuesto anteriormente, es la "vía de señalización HIF-1" (FDR = 0,0084) (**Figura 4.138**), ya que entre los genes que aparecen desregulados en esta ruta está *LDHA* (*z*-valor = -14,59, *p*-valor < 0,0001). El gen *LDHA* codifica la proteína lactato deshidrogenasa A, que es una enzima asociada a la glicolisis anaeróbica, implicada también en procesos de desarrollo, invasión y metástasis en procesos cancerígenos[516] mediante el conocido efecto Warburg[517]. Su silenciamiento en células mielomatosas mediante ARN cortos de interferencia (shRNA) conduce a la restauración de la sensibilidad al bortezomib en HMCLs resistentes[513]. Por lo tanto, la acción de la amilorida sobre *LDHA* en las células mielomatosas podría tener un efecto beneficioso en el tratamiento del MM, como terapia de rescate.

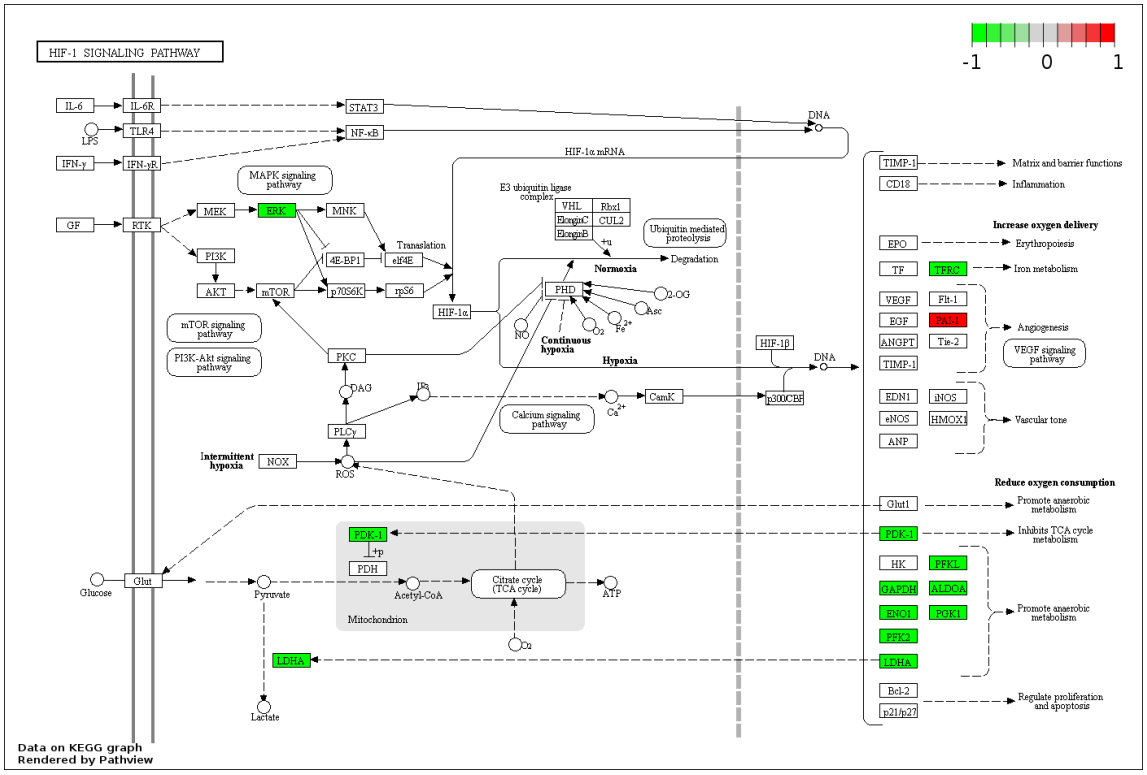



***Figura 4.138.*** *Vía de señalización de HIF-1 según la base KEGG. En verde se representan los genes infraexpresados de forma estadísticamente significativa en el metaanálisis global de la amilorida, mientras en rojo se representan los genes sobreexpresados de manera estadísticamente significativa en ese mismo metaanálisis.*

En lo que respecta a la modulación del *splicing* alternativo promovida por la amilorida, hay que destacar que no se detectó entre las 10 rutas y términos más significativos ningún elemento sobrerrepresentado asociado directamente a este proceso

(**Figura 4.137**). Sin embargo, a nivel génico sí se detectó sobreexpresión de dos genes que codifican proteínas componentes de la vía del espliceosoma, el gen *DHX16* (*z*-valor = 7,94, *p*-valor < 0,0001) y el gen *PRPF3* (*z*-valor = 5,80, *p*-valor < 0,0001), este último previamente reportado en el trabajo de Rojas y colaboradores en células de pacientes de MM tratadas con amilorida[169], lo que sugiere un mecanismo de modulación del *splicing* alternativo a través de la regulación de los niveles de mRNA de algunos de los componentes del espliceosoma.

#### 4.3.11.2. TG003

El compuesto TG003 es un benzotiazol que actúa como potente inhibidor de la familia de quinasas Clk. Al igual que ocurre con la amilorida, este compuesto también ha sido asociado con procesos de modulación del *splicing* alternativo. En este trabajo se evaluó, mediante técnicas de metaanálisis, el efecto del TG003 sobre la expresión génica en HMCLs. Se dispuso para ello de datos de RNA-seq de las HMCLs KMS12-BM y JJN-3 sometidas a tratamiento en monoterapia con este compuesto, además de los correspondientes controles tratados con DMSO, todos ellos utilizados previamente para el desarrollo del pipeline de RNA-seq en el **Capítulo 1** y depositados en GEO con el número de serie GSE95077. Las características generales de los dos estudios correspondientes a las dos líneas celulares se recogen en la **Tabla 4.15**.

***Tabla 4.15.*** *Estudios seleccionados para el metaanálisis de efectos aleatorios de la expresión génica en líneas celulares de mieloma múltiple tratadas con TG003.*

| Serie | Estudio | Línea Celular | Plataforma | N | Tiempo (h) | Concentración (mM) |
|---|---|---|---|---|---|---|
| GSE95077 | Rojas (2018a)* | KMS12-BM | Illumina HiSeq 2500 | 6 | 24 | 0,4 |
| GSE95077 | Rojas (2018b)* | JJN-3 | Illumina HiSeq 2500 | 6 | 24 | 0,4 |

** Estudios no publicados.*

Se consideraron 481 genes candidatos para el metaanálisis seleccionando aquellos genes cuyo valor absoluto del FC fue mayor a 1,5 en los dos estudios. La distribución de los ln(FC) de estos genes se muestra como diagrama de caja en la **Figura 4.139**, observándose un patrón muy similar en los dos estudios, con una ligera tendencia a la infraexpresión en ambos casos.

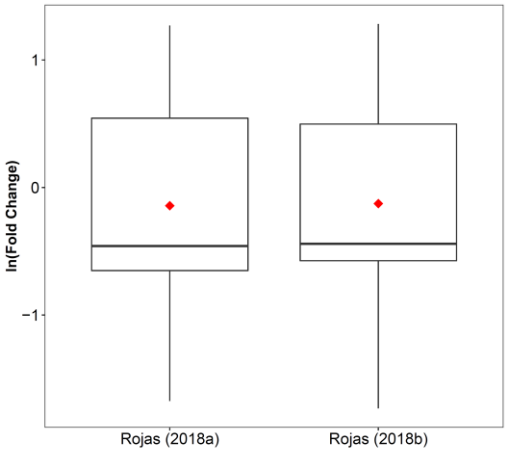



***Figura 4.139.*** *Diagrama de caja (box plot) del ln(Fold Change) (ln[FC])de los 481 genes seleccionados para el metaanálisis de la expresión génica en líneas celulares de mieloma múltiple tratadas con TG003. El diamante rojo representa el promedio del ln(FC) en cada estudio.*

Finalmente, se llevó a cabo el estudio con metaanálisis de estos 481 genes previamente seleccionados, de manera que se observaron diferencias estadísticamente significativas a *p*-valor < 0,05 en el tamaño del efecto en 470 genes (**Anexo 35**). En cuanto a la dirección del cambio de expresión, 187 genes presentaron sobreexpresión, mientras que 283 genes mostraron infraexpresión al tratamiento con TG003. En la **Figura 4.140** se representa el análisis ORA de estos genes significativos considerando como las vías biológicas KEGG y los términos GO.

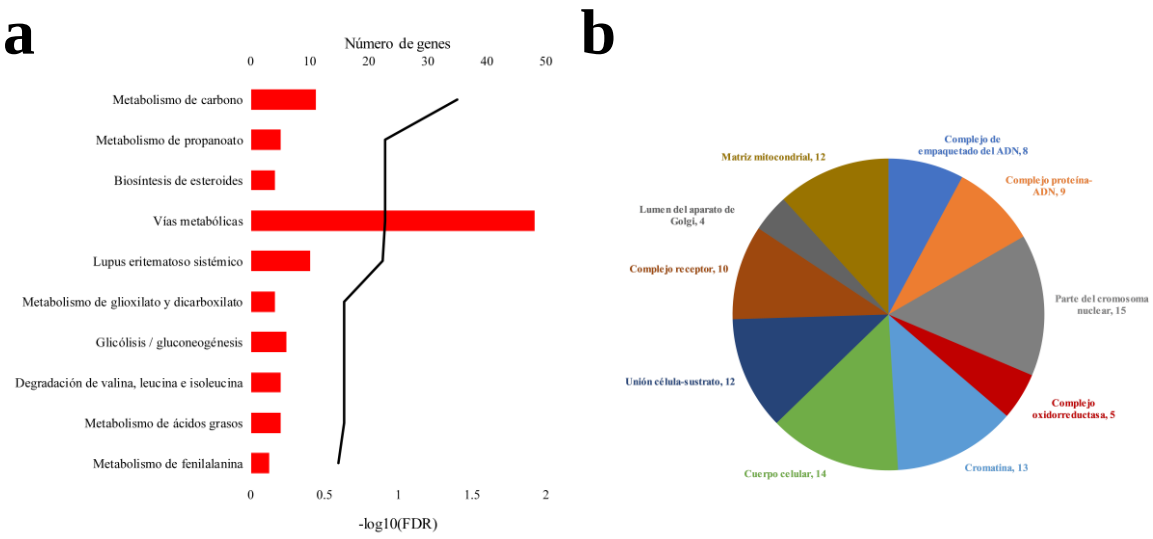



***Figura 4.140.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas en el metaanálisis global del TG003. En esta figura se recogen las 10 rutas KEGG y los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.*

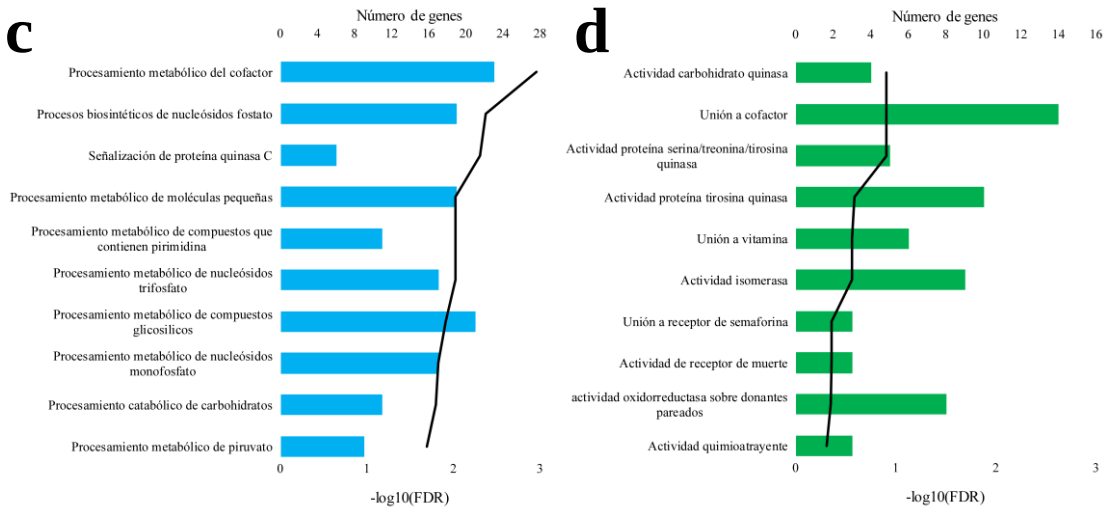



***Figura 4.140 (continuación).*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas en el metaanálisis global del TG003. En esta figura se recogen las 10 rutas KEGG y los10 términos GO con un menor valor de FDR.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Al igual que sucedió con la amilorida, el análisis ORA para el TG003 mostró una gran cantidad de términos GO sobrerrepresentados de forma estadísticamente significativa asociadas a procesos metabólicos, aunque en este caso, solamente una vía KEGG, la vía del "metabolismo de carbono", superó el umbral de significancia (FDR = 0,0401). Las repercusiones de la desregulación de los procesos metabólicos en MM ya fueron discutidas en el apartado de la amilorida, apareciendo nuevamente desregulado en el caso del TG003 el gen *LDHA* (*z*-valor = -4,37, *p*-valor < 0,0001), lo que podría sugerir un mecanismo de acción de este compuesto sobre el metabolismo muy similar al producido por la amilorida.

En cuanto a la desregulación de los procesos de *splicing* alternativo, de manera similar la amilorida, no aparecieron vías o términos sobrerrepresentados asociados a este proceso. Sin embargo, tampoco fue detectado ningún gen desregulado asociado a la maquinaria del espliceosoma, lo que conduce a la hipótesis de que la regulación de esta vía por el TG003 podría estar produciéndose a otro nivel diferente del mRNA génico.

#### 4.3.11.3. Interferón γ

En este trabajo, se realizó una búsqueda sistemática en repositorios de datos online de series de datos de expresión génica en HMCLs tratadas en monoterapia con interferón. Como se muestra en el **Anexo 9**, solamente fue detectado un estudio correspondiente a la serie GSE94134, con lo que no fue viable la realización de un metaanálisis. Este estudio constaba de cuatro muestras de expresión génica medida con RNA-seq (Illumina HiSeq 2500) de la línea celular ALF1, de las que dos muestras estaban tratadas con IFN-γ a una concentración de 1 ng/mL por dos horas, y dos muestras control tratadas con el vehículo

DMSO. Sobre estas muestras se realizó un análisis de expresión diferencial utilizando el algoritmo *edgeR* en su versión GLM. Mediante este análisis se detectaron 15 genes estadísticamente significativos a FDR < 0,05 (**Anexo 36**), sobre los que se realizó un análisis ORA considerando las bases de datos KEGG y GO (**Figura 4.141**).

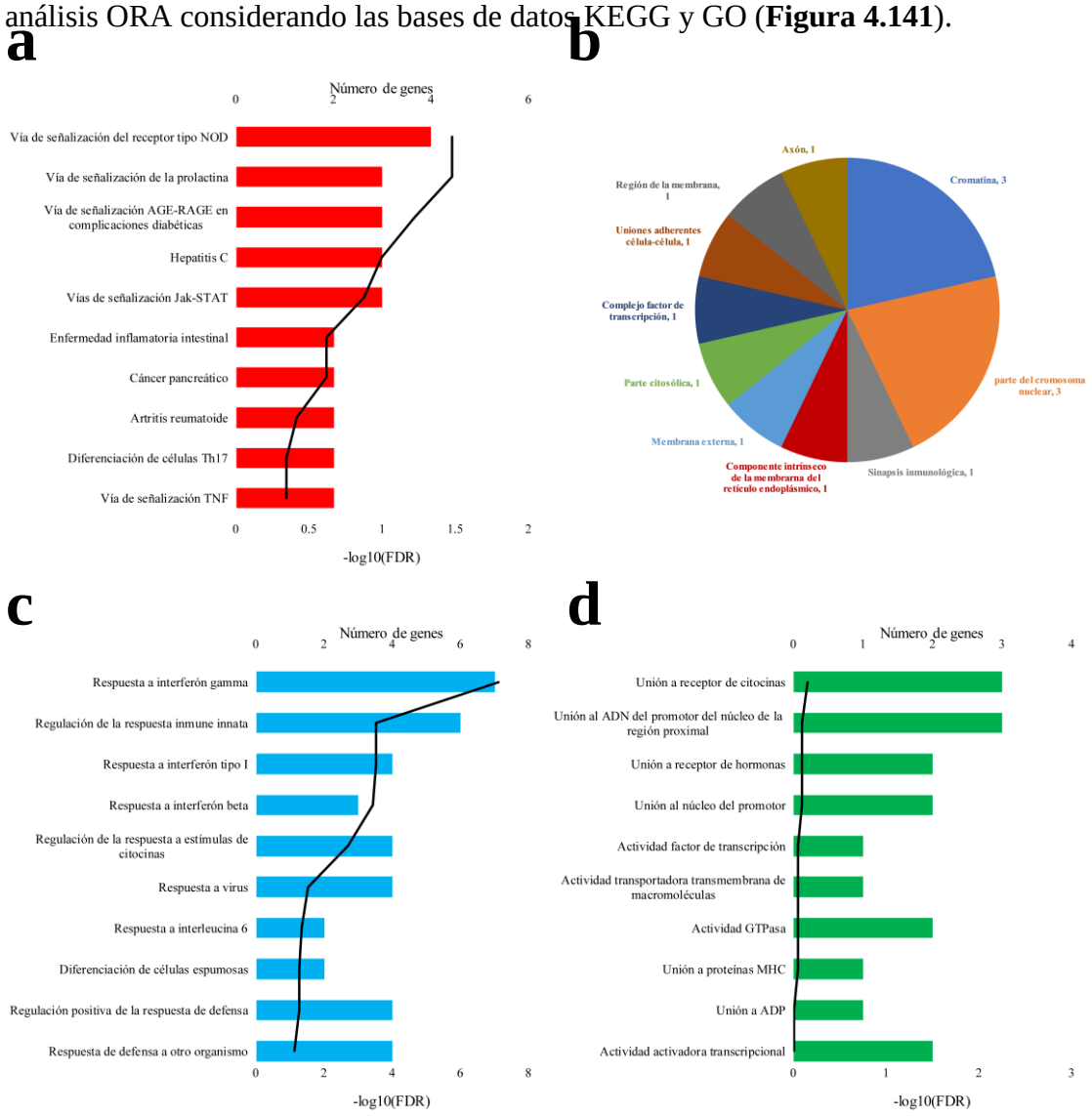



***Figura 4.141.*** *Análisis de sobrerrepresentación de los genes con diferencias de expresión estadísticamente significativas para el tratamiento con interferón γ. En esta figura se recogen las 10 rutas KEGG y los10 términos GO con un menor valor de FDR.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 componentes celulares GO.* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Como era esperable, el PB con un menor valor de FDR fue la "respuesta a interferón gamma" (FDR < 0,0001), de manera que 8 de los 15 genes desregulados pertenecían a este proceso. Entre estos 8 genes, se encuentra el factor regulatorio del interferón *IRF1* (FC = 25,56, FDR < 0,0001). La proteína codificada por este gen se une al ADN en una región regulatoria de la transcripción llamada "elementos de respuesta estimulados por interferón" o ISRE, del inglés *interferon-stimulated response element*[518], lo que conduce a la modulación de la expresión de los genes diana. Otro de los genes desregulados en

este PB es el activador transcripcional *STAT1* (FC = 3,26, FDR < 0,0001), cuya modulación a través de IFN-γ se ha demostrado crucial para la transcripción de genes cruciales en la respuesta a esta molécula[519]. Otra función de este gen es su actividad como supresor tumoral y su expresión se ha asociado a buen pronóstico en diversos tipos de tumores sólidos[520], si bien su efecto promotor tumoral también se ha descrito en otros muchos trabajos[520]. En el presente análisis también se detectó la sobreexpresión de otro gen de la familia *STAT*, en concreto del gen *STAT3* (FC = 1,55, FDR = 0,0140), cuya correlación con el crecimiento y la supervivencia de las células tumorales ha sido reportada en varios trabajos en cáncer[521]. Existe evidencia experimental de que la regulación de *STAT1* y *STAT3* es una regulación cruzada, de modo que la modulación de la ratio de expresión entre estos dos genes puede producir un efecto proapoptótico o proliferativo en la célula, en función de si esta ratio es favorable a *STAT1* o a *STAT3*, respectivamente[522]. Por este motivo, el resultado obtenido en este trabajo requerirá un estudio de mayor profundidad, ya que la sobreexpresión de ambos genes por IFN-γ en esta HMCL no concuerda con este modelo regulatorio propuesto en la literatura. No obstante, una posible explicación a este fenómeno podría ser la aparición de un mecanismo de resistencia que haría frente a la sobreexpresión de *STAT1* producida por el IFN-γ, ya que precisamente la activación de *STAT3* juega un papel de gran importancia en mecanismos de resistencia en un amplio espectro de tratamientos[523].

### 4.3.12. Comparación de las firmas génicas de los fármacos antimieloma

Teniendo en cuenta que algunos de los fármacos analizados comparten mecanismos de acción se llevó a cabo una comparación de las 12 listas de genes para determinar la presencia o ausencia de una firma transcripcional común entre ellos. La intersección entre las 12 listas de genes se realizó mediante la herramienta online de cálculo de diagramas de Venn de la Universidad de Gent (http://bioinformatics.psb.ugent.be/webtools/Venn/). El resultado de este cruce fue representado como una matriz de similitud, tal y como se recoge en la **Figura 4.142**.

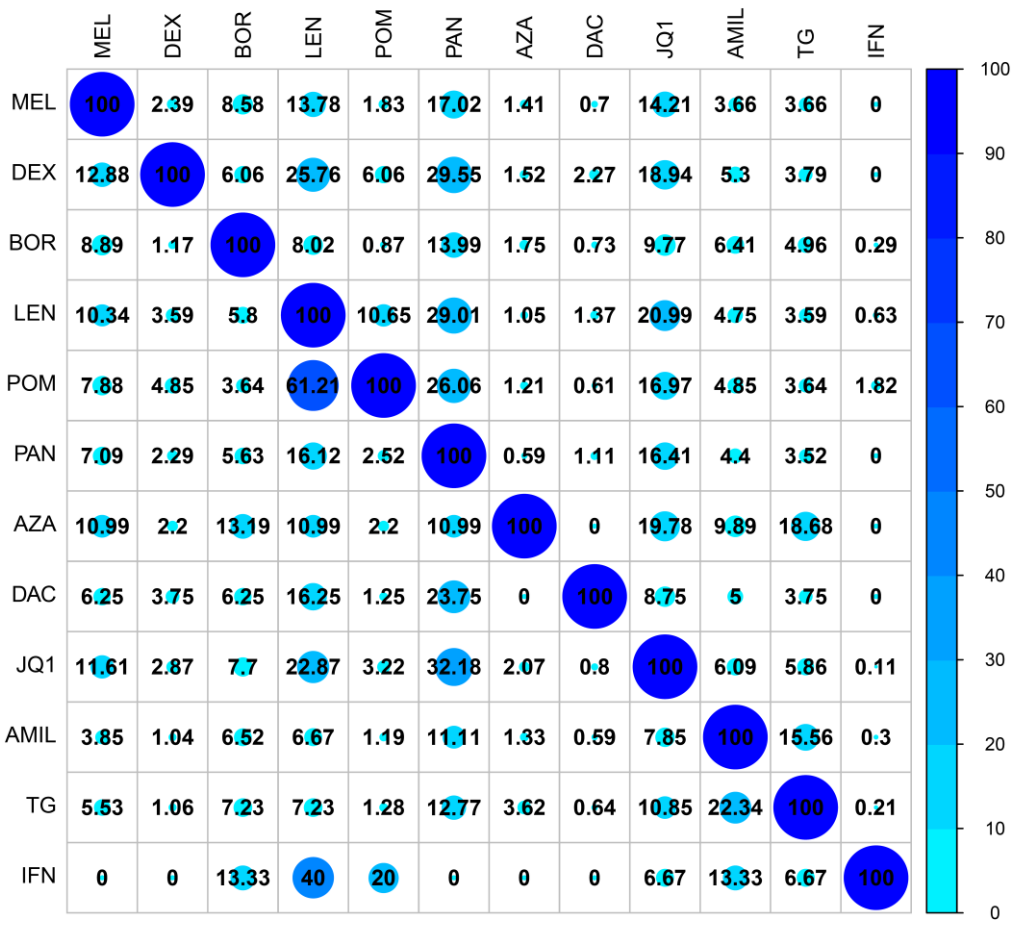
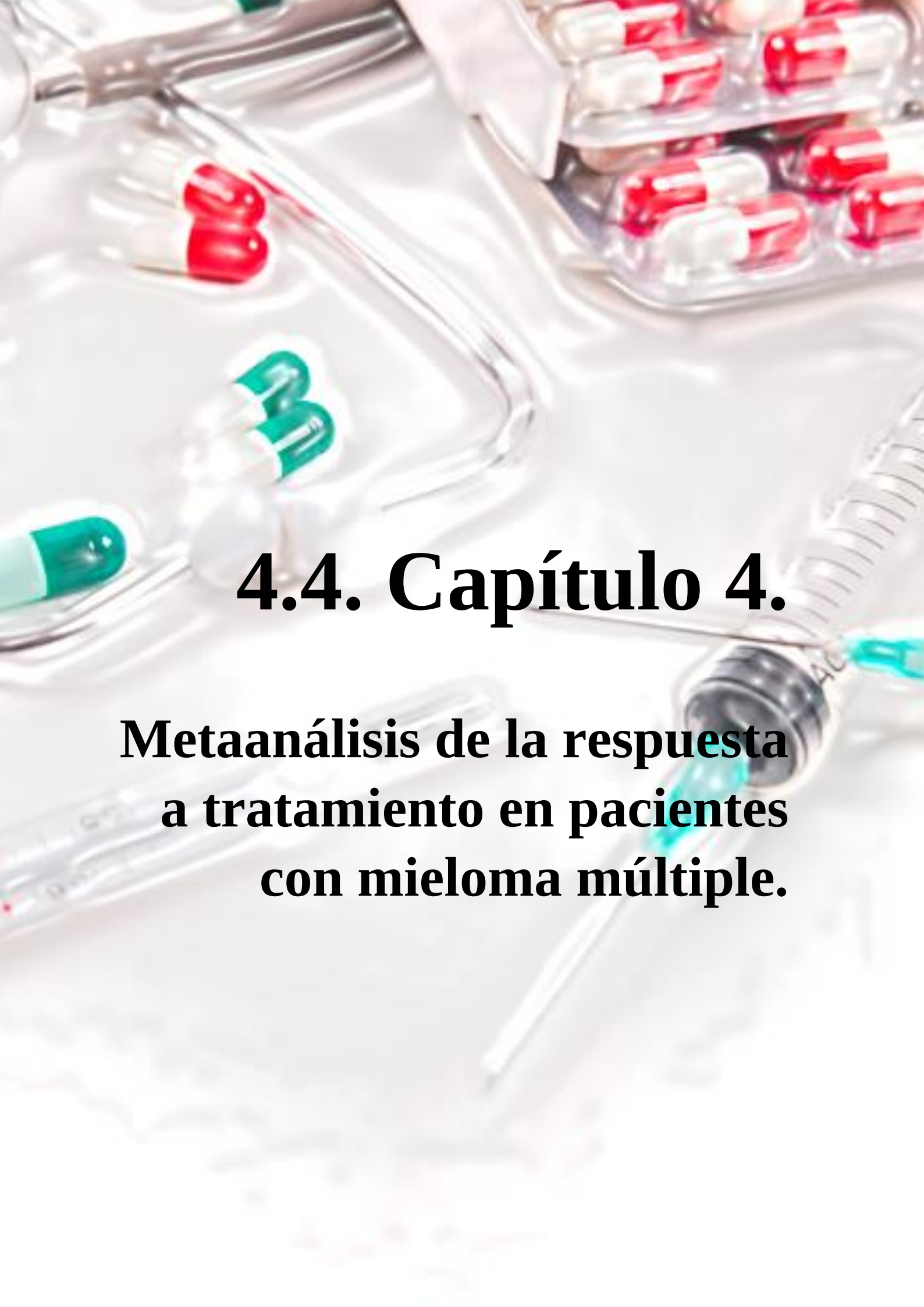



***Figura 4.142.*** *Matriz de similitud para las listas de genes estadísticamente significativos en los análisis de 12 fármacos aplicados en monoterapia en líneas celulares de mieloma múltiple. Cada círculo representa el porcentaje de similitud entre dos fármacos, el cúal es también indicado numéricamente. Los valores porcentuales son relativos al total de genes desregulados en la lista correspondiente al fármaco indicado en el eje de ordenadas (Y). MEL: melfalán, DEX: dexametasona, BOR: bortezomib, LEN: lenalidomida, POM: pomalidomida, PAN: panobinostat, AZA: azacitidina, DAC: decitabina, JQ1: JQ1, AMIL: amilorida, TG: TG003, IFN: interferón γ.*

La mayor similitud entre fármacos se detectó en la comparación entre los dos IMiDs, lenalidomida y pomalidomida, alcanzando un 61,2 % de similitud respecto a la lista con menor número de genes (pomalidomida). El resto de comparaciones mostraron porcentajes de similitud inferiores al 50%, siendo la comparación entre lenalidomida e IFN-γ con un 40%, considerando como referencia los 15 genes desregulados con IFN-γ, la de mayor semejanza. El hecho que podría explicar que los mayores porcentajes de similitud se registren entre los dos IMiDS y además entre la lenalidomida y el IFN-γ, es que los tres compuestros están implicados en funciones inmunomoduladoras a través de la regulación de los niveles de interleucinas[150, 524], y además los niveles del propio IFN-γ están regulados por ambos IMiDs[98].

En lo que respecta a los menores porcentajes de similitud, resulta llamativo el bajo solapamiento entre las firmas de expresión génica de la azacitidina y su deoxiderivado decitabina, no encontrándose genes coincidentes entre ambos fármacos. Estudios previos en LMA sobre agentes desmetilantes ya reportaron estas diferencias en el potencial regulatorio transcripcional de estos compuestos[525], y podría ser una de las causas de los diferentes mecanismos de acción, efectos y respuestas que ambos compuestos presentan en el tratamiento de diferentes tipos de cáncer[139, 142, 526].



# 4.4. Capítulo 4.

## Metaanálisis de la respuesta a tratamiento en pacientes con mieloma múltiple.

El objetivo de este capítulo fue la determinación de una firma de expresión génica de la respuesta a varios regímenes de tratamiento en pacientes con MM. Para ello, se realizó una búsqueda bibliográfica sistemática de series con datos de expresión génica medidos con tecnologías masivas de alto rendimiento utilizando como palabra clave *"myeloma"* en los repositorios GEO, ArrayExpress y SRA. Esto se tradujo en la detección de 288 series en GEO, 253 series en ArrayExpress y 21 series en SRA. Además, se añadió una serie adicional correspondiente al proyecto CoMMpass de la *Multiple Myeloma Research Foundation* (MMRF, https://research.themmrf.org y www.themmrf.org). Tras la eliminación de los elementos duplicados en los tres repositorios, fueron seleccionadas 367 series, de las que 10 cumplieron los criterios de inclusión y exclusión propuestos en la **Sección de Material y métodos**. Los datos aquí expuestos corresponden a búsquedas bibliográficas cuya fecha de cierre se estableció a fecha 9 de febrero de 2017. Como se muestra en el diagrama de flujo de la **Figura 4.143**, se establecieron seis posibles grupos de análisis en función del régimen de tratamiento aplicado a los pacientes de las 10 series seleccionadas:

1. Bortezomib en monoterapia.
2. Terapias basadas en bortezomib en combinación o sin combinar con otros fármacos, exceptuando IMiDs.
3. Terapias basadas en la combinación de bortezomib e IMiDs.
4. Terapias basadas en la combinación de vincristina, adriamicina y dexametasona (VAD)
5. Trasplante autólogo de células madre hematopoyéticas (ASCT, del inglés *Autologous Stem Cell Transplant*)

Dos de las 10 series no pudieron ser asignadas a ninguno de estos grupos de tratamiento, por lo que tuvieron que ser descartadas. La causa de esta exclusión fue la alta heterogeneidad en los regímenes de tratamiento que recibieron los pacientes recogidos en estas dos series, no detectándose ninguna agrupación de muestras de tamaño suficiente como para llevar a cabo un estudio de expresión génica, ya que el tamaño mínimo de muestras por grupo de tratamiento fue establecido en tres muestras.

En lo que respecta a los grupos de tratamiento correspondientes a la terapia VAD y el ASCT, no se consideraron en los análisis posteriores debido a que en ambos casos una única serie cumplió los criterios de inclusión. En el resto de los regímenes de tratamiento mencionados se procedió al estudio mediante metaanálisis. El procedimiento completo de análisis se detallará en cada uno de los apartados de este capítulo.

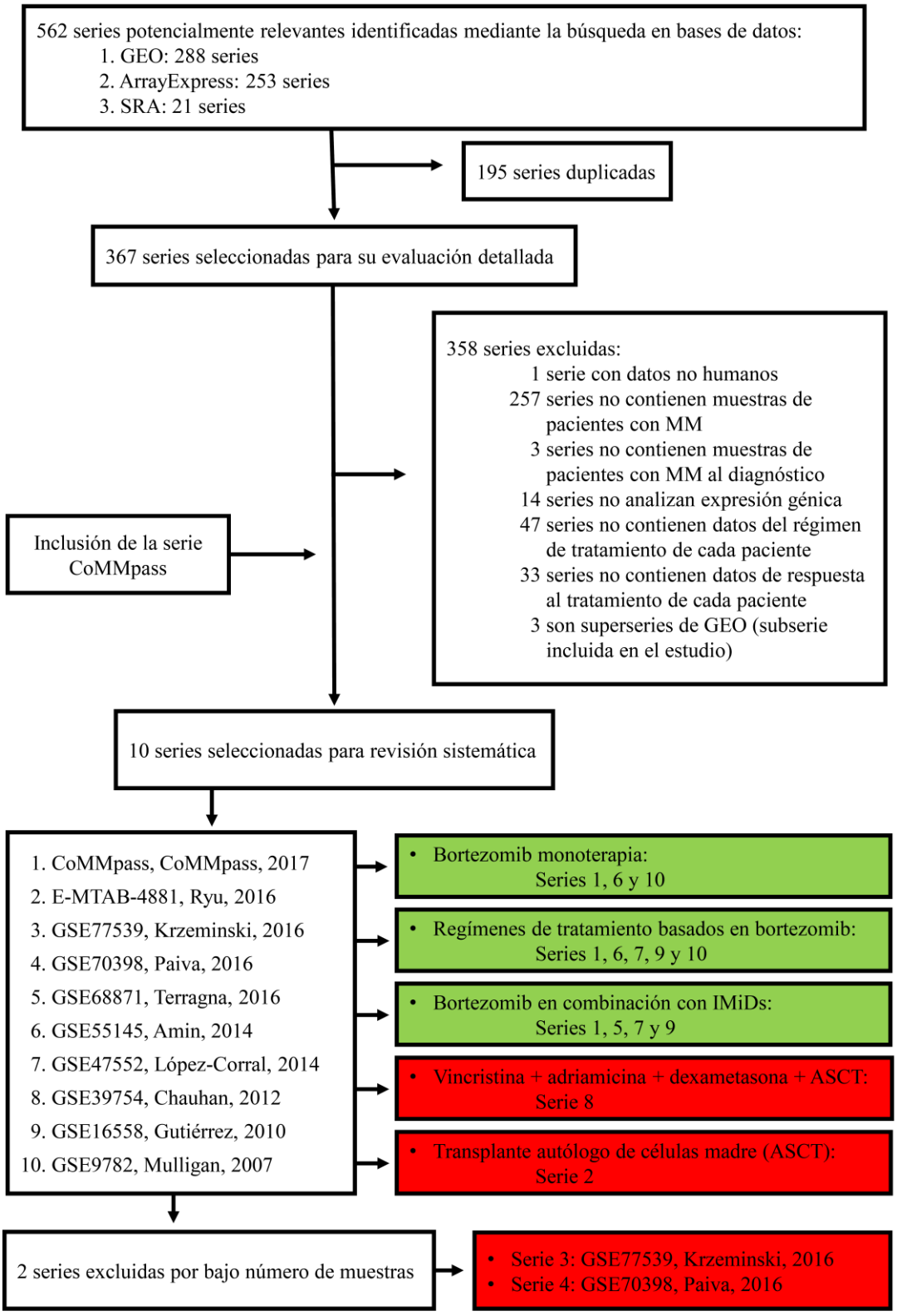



***Figura 4.143.*** *Diagrama de flujo de la selección de series candidatas al estudio mediante metaanálisis de la respuesta a distintos regímenes de tratamiento en MM. En verde, regímenes de tratamiento seleccionados para el metaanálisis, en rojo, regímenes de tratamiento excluidos.*

### 4.4.1. Bortezomib en monoterapia

Como paso previo al metaanálisis, se seleccionaron los pacientes tratados con bortezomib en monoterapia de las tres series que se indican en la **Figura 4.143** y se determinaron los genes comúnmente interrogados por las correspondientes plataformas de análisis de expresión génica. El número total de genes comunes fue de 15.855 genes. Sin embargo, en la serie GSE9782, debido a que no se dispuso de los archivos CEL no procesados, se admitieron genes duplicados, por lo que constaría de 31.823 "grupos de sondas" o *probesets* correspondientes a 15.855 genes. A continuación, se determinó la respuesta alcanzada por cada paciente y se procedió a su estratificación mediante dos criterios:

1) Pacientes respondedores al tratamiento (OR, del inglés "*overall response*") frente a los que no respondieron (NR).
2) Pacientes que alcanzaron respuesta completa (RC) al tratamiento frente a los que no la alcanzaron (grupo "Resto").

Sobre estos dos grupos se llevó a cabo el estudio de la respuesta mediante metaanálisis.

#### 4.4.1.1. Pacientes respondedores *versus* no respondedores

En la categoría de pacientes OR se consideraron todos los pacientes que alcanzaron al menos respuesta parcial (RP) en cada uno de los estudios seleccionados. Mientras que en el grupo de pacientes NR se incluyeron los que mantuvieron la enfermedad estable (EE) y los que progresaron (EP).

Como se muestra en la **Tabla 4.16**, los estudios de Amin (2014) y Mulligan (2007) presentaron en los dos grupos de respuesta un buen equilibrio en cuanto al número de muestras (ratio OR/NR 1,39 y 0,76, respectivamente). Sin embargo, en el estudio CoMMpass (2017) el número de muestras en ambos grupos fue mucho menor, con una ratio OR/NR mayor al de los otros dos estudios (ratio OR/NR = 2), con lo que es posible que este estudio tenga una menor potencia estadística pudiendo conducir a la obtención de una lista de genes menos estable[527].

***Tabla 4.16.*** *Estudios seleccionados para el metaanálisis de la expresión génica en pacientes respondedores (OR) frente a no respondedores (NR) en régimen de tratamiento con bortezomib en monoterapia.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | OR | NR |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 6 | 3 |
| GSE55145 | Amin, 2014[197] | Affymetrix Human Exon 1.0 ST Array | 39 | 28 |
| GSE9782 | Mulligan, 2007[528] | Affymetrix Human Genome U133A y U133B Arrays | 73 | 96 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org).*

Una vez determinados los grupos de respuesta se analizó la expresión génica diferencial en cada uno de los tres estudios mediante los algoritmos *limma* o *edgeR* según correspondiese a muestras de microarray o RNA-seq, respectivamente. En el caso del estudio CoMMpass (2017) se detectaron 1.540 genes con diferencias de expresión a *p*-valor < 0,05, de los que 770 genes se encontraban sobreexpresados y 770 infraexpresados en OR frente a NR. Respecto al estudio de Amin (2.014), 1.255 genes mostraron diferencias de expresión a *p*-valor < 0,05, de los que 638 genes estaban sobreexpresados, mientras que 617 genes mostraban infraexpresión en la comparación OR *vs.* NR. Por último, en el estudio de Mulligan (2007), se detectaron 4.269 genes con diferencias de expresión a *p*-valor < 0,05, que se redujo a 3.549 genes excluyendo los duplicados. De estos, 2.215 genes presentaron sobreexpresión, 1.283 infraexpresión y 51 presentaron al menos un *probeset* en cada sentido de expresión.

El cruce mediante diagramas de Venn de los genes obtenidos en los tres estudios se realizó sin considerar el sentido de la expresión, ya que será posteriormente el estudio mediante metaanálisis el que sancione la variabilidad de los genes entre los tres estudios. Mediante este cruce, se identificaron 36 genes comúnmente desregulados en los tres estudios, y un total de 712 genes, incluyendo los anteriores 36, aparecieron desregulados de manera común al menos en dos de los estudios (**Figura 4.144)**.

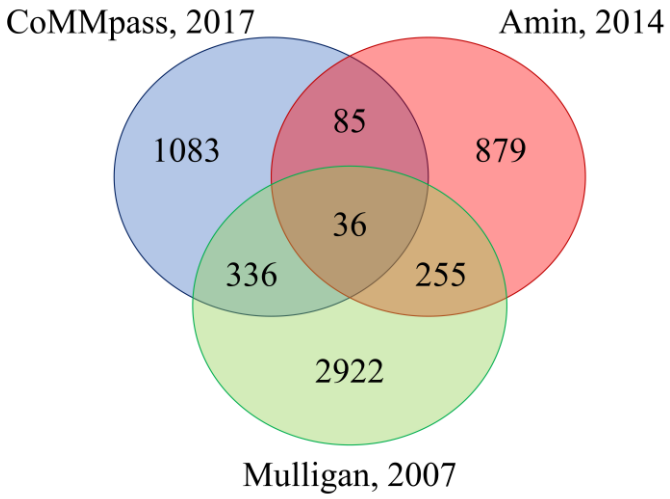



***Figura 4.144****. Diagrama de Venn de los genes seleccionados en los tres estudios para la comparación de expresión génica diferencial entre los pacientes respondedores y los no respondedores en el tratamiento con bortezomib en monoterapia.*

Hay que puntualizar dos aspectos en este análisis de expresión diferencial. En primer lugar, se decidió utilizar el algoritmo *edgeR* para los datos de RNA-seq porque en este estudio se trabajó con un nivel de $\alpha = 0,05$, y este fue el método que obtuvo mejor rendimiento a este nivel de significancia, tal y como fue descrito en el **Capítulo 1**. En segundo lugar, en este análisis se ha utilizado el *p*-valor para determinar qué genes muestran diferencias de expresión estadísticamente significativas, en lugar del FDR u

otro tipo de ajuste de la probabilidad para contrastes múltiples. Se decidió actuar de esta manera debido a que el objetivo de estos análisis fue la selección de variables para ser analizadas mediante metaanálisis. En un análisis exhaustivo de la expresión génica diferencial, cuyo objetivo fuese la determinación en una serie de muestras de una firma de expresión génica, se habrían requerido dichos ajustes de control de la tasa de error para comparaciones múltiples[529].

Tras la selección de los 712 genes comunes, se procedió al metaanálisis, encontrando que 233 genes mostraron un tamaño del efecto estadísticamente significativo (*p*-valor < 0,05) en los pacientes OR respecto a los NR (**Anexo 37**). Si consideramos el sentido de la expresión, 135 de los genes mostraron sobreexpresión y 98 infraexpresión en los pacientes OR. Sobre estos 233 genes seleccionados se realizó el análisis ORA en rutas biológicas KEGG y en funciones GO, como se muestra en la **Figura 4.145**.

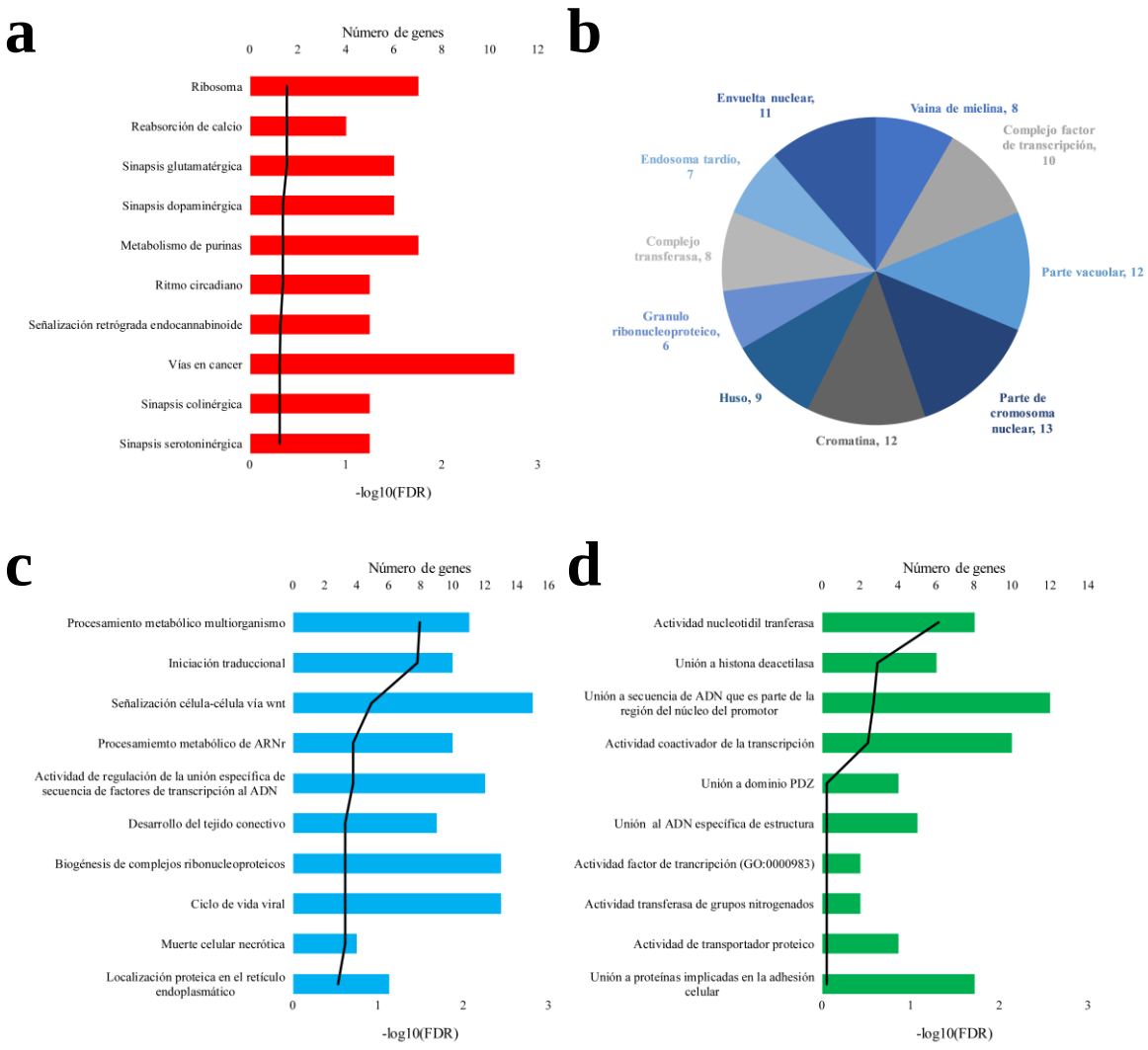



***Figura 4.145.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en el metaanálisis de la respuesta al tratamiento con bortezomib en monoterapia.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localización celular GO (CC),* ***c)*** *TOP 10 procesos biológicos GO (PB) y* ***d)*** *TOP 10 funciones moleculares GO (FM).*

En lo que respecta a las vías KEGG, ningún término alcanzó una sobrerrepresentación estadísticamente significativa. Sin embargo, en lo referente a las ontologías génicas (GO) se detectaron dos PB y una FM estadísticamente significativos. Entre los PB la función de "iniciación traduccional" fue estadísticamente significativa (FDR = 0,035) debido al elevado número de genes diferencialmente expresados desregulados que codifican proteínas ribosomales, como *RPL24*, *RPL29*, *RPL32*, *RPS15A* y *RPS25*, lo cual concuerda con el hecho de que la vía KEGG con mejor posición en cuanto a significancia, aunque no superó el umbral del FDR = 0,05, fue la vía "ribosoma" (FDR = 0,413). En este PB también aparecen desregulados tres factores de iniciación de la traducción como son *EIF3D*, *EIF3F* y *EIF3G*, así como genes cruciales en la modificación postraduccional mediante ubiquitinación como es el gen *UBA52* y en la regulación transcripcional como *RARA*. Todos estos genes aparecen infraexpresados en los pacientes OR, lo que podría indicar una regulación negativa de la iniciación de la traducción en estos pacientes. Esta regulación negativa podría ser debida a la infraexpresión de los genes *LAMTOR1* (*z*-valor = -2,31, *p*-valor = 0,0207) y *LAMTOR4* (*z*-valor = -2,70, *p*-valor = 0,0069) que codifican dos de las subunidades del complejo *ragulator-Rag* que actúa como regulador positivo del complejo mTORC1[530], activador de la traducción de proteínas. De este modo, los niveles bajos de *LAMTOR1* y *LAMTOR4*, podrían conducir a la inactivación del complejo mTORC1[531], produciendo la inhibición de la traducción de proteínas. Todo esto, unido a la infraexpresión del factor de iniciación *EIF3F* (*z*-valor = -1,97, *p*-valor = 0,0475), podría corroborar esta hipótesis, ya que, como se ha descrito previamente, la disminución de la actividad de mTORC1 se correlaciona con la degradación de este factor de iniciación[532].

Por último, este análisis fue completado mediante el estudio del sesgo de publicación utilizando la prueba de asimetría de Egger. De esta manera se determinó que 259 de los 712 genes iniciales podrían mostrar sesgo de publicación (*p*-valor < 0,05). No obstante, al ser una proporción inferior al 50% (36,4%) no fue considerado problemático al no ser una tendencia generalizada del conjunto global de genes.

#### 4.4.1.2. Pacientes que alcanzan respuesta completa *versus* resto

Para los tres estudios seleccionados en el análisis de la RC, se incluyeron en el grupo RC todas las respuestas que tenían como mínimo una IFE negativa. El tamaño muestral y la plataforma de análisis empleadas en cada uno de estos estudios aparecen recogidos en la **Tabla 4.17**, donde puede observarse la gran descompensación en el número de muestras de los dos grupos de respuesta analizados. Las ratios RC/Resto para los estudios CoMMpass (2017), Amin (2014) y Mulligan (2007) fueron de 0,5 (~2 veces), 0,31 (~3,23 veces) y 0,08 (~12,5 veces), respectivamente. Esta falta de balance entre los grupos podría crear problemas si se aplicasen test de permutaciones para el estudio de expresión diferencial ya que sesgaría el *p*-valor calculado[533], por tanto, tuvo que tenerse en cuenta esta consideración a la hora de determinar la expresión génica diferencial.

***Tabla 4.17.*** *Estudios seleccionados para el metaanálisis de la expresión génica en pacientes con respuesta completa (RC) frente al resto de respuestas en régimen de tratamiento con bortezomib en monoterapia.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | RC | Resto |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 3 | 6 |
| GSE55145 | Amin, 2014[197] | Affymetrix Human Exon 1.0 ST Array | 16 | 51 |
| GSE9782 | Mulligan, 2007[528] | Affymetrix Human Genome U133A y U133B Arrays | 13 | 156 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org).*

El análisis de la expresión génica diferencial en el estudio CoMMpass (2017), reveló que 1.543 genes muestran diferencias de expresión a *p*-valor < 0,05 en los pacientes que alcanzaron RC respecto al resto de pacientes. De estos genes, 950 presentaron sobreexpresión y 593 infraexpresión en los pacientes con RC. En cuanto a los estudios realizados con *microarrays*, en el estudio de Amin (2014) se detectaron 604 genes a *p*-valor < 0,05, de los que 310 genes estuvieron sobreexpresados y 294 infraexpresados en los pacientes con RC. En el estudio de Mulligan (2007) se identificaron 1.297 genes, de los que, excluyendo los duplicados, 443 estaban sobreexpresados y 751 infraexpresados en RC, y 15 genes mostraban sentidos de expresión opuestos en dos o más sondas. El cruce de los genes seleccionados de los tres estudios se muestra en el diagrama de Venn de la **Figura 4.146**. Cuatro de los genes seleccionados fueron comunes a los tres estudios, y un total de 221 genes fueron comunes a, al menos, dos de los estudios.

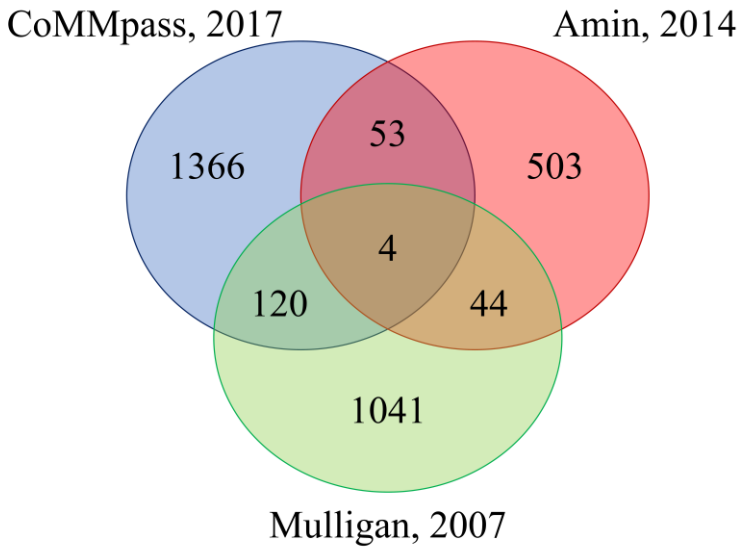



***Figura 4.146.*** *Diagrama de Venn de los genes seleccionados en los tres estudios para la comparación de expresión génica diferencial entre los pacientes que alcanzan respuesta completa frente al resto de pacientes bajo el tratamiento con bortezomib en monoterapia.*

A través del metaanálisis de estos 221 genes comunes se detectaron un total de 26 genes estadísticamente significativos a *p*-valor < 0,05, de los que 12 presentaron sobreexpresión y 14 infraexpresión en los pacientes que alcanzaron RC (**Anexo 38**). En la **Tabla 4.18** se muestra un mapa de calor con los valores de ln(FC) de estos 26 genes.

***Tabla 4.18.*** *Mapa de calor (heatmap) de los 26 genes estadísticamente significativos en el metaanálisis de efectos aleatorios para el estudio de la expresión génica en pacientes con respuesta completa frente al resto de respuestas para el régimen de tratamiento con bortezomib en monoterapia.*

| Símbolo | Valor *z* | *p*-valor | CoMMpass (2017) | Amin (2014) | Mulligan (2007) |
|---|---|---|---|---|---|
| BCLAF1 | 5,669 | 1,43E-08 | | | |
| C10orf118 | 4,987 | 6,12E-07 | | | |
| NKAIN2 | 3,349 | 8,11E-04 | | | |
| PXDC1 | 3,188 | 0,001 | | | |
| KLHL14 | 2,941 | 0,003 | | | |
| HEBP2 | 2,602 | 0,009 | | | |
| USP1 | 2,413 | 0,016 | | | |
| SOCS6 | 2,247 | 0,025 | | | |
| FAM114A1 | 2,243 | 0,025 | | | |
| SPRED1 | 2,228 | 0,026 | | | |
| TMEM144 | 2,143 | 0,032 | | | |
| USP54 | 2,129 | 0,033 | | | |
| NIN | -4,838 | 1,31E-06 | | | |
| DCLK1 | -4,480 | 7,47E-06 | | | |
| LAMTOR4 | -4,429 | 9,47E-06 | | | |
| UBASH3B | -3,746 | 1,80E-04 | | | |
| TIMM13 | -3,711 | 2,07E-04 | | | |
| BCAS3 | -3,419 | 6,29E-04 | | | |
| NUP210 | -3,402 | 6,70E-04 | | | |
| TRAF7 | -3,239 | 0,001 | | | |
| CLYBL | -2,510 | 0,012 | | | |
| QSOX2 | -2,439 | 0,015 | | | |
| GATM | -2,401 | 0,016 | | | |
| FRMD8 | -2,105 | 0,035 | | | |
| HMGA1 | -2,021 | 0,043 | | | |
| LGR4 | -1,965 | 0,049 | | | |

*Los parámetros del valor z y el p-valor hacen referencia a los valores del estadístico z y de la probabilidad estadística obtenidos en el metaanálisis, respectivamente. El color de la celda es rojo si el ln(FC) del gen en ese estudio es positivo (gen sobreexpresado), verde si el ln(FC) es negativo (gen infraexpresado). A mayor intensidad del color rojo o verde, mayor o menor valor del ln(FC), respectivamente.*

El análisis de sobrerrepresentación de genes en rutas biológicas KEGG sobre los 26 genes no obtuvo ninguna vía representada por más de un gen. Sin embargo, sí se observó una alta representación de genes en algunas funciones GO, aunque no estadísticamente significativa, tal como se muestra en la **Figura 4.147**.

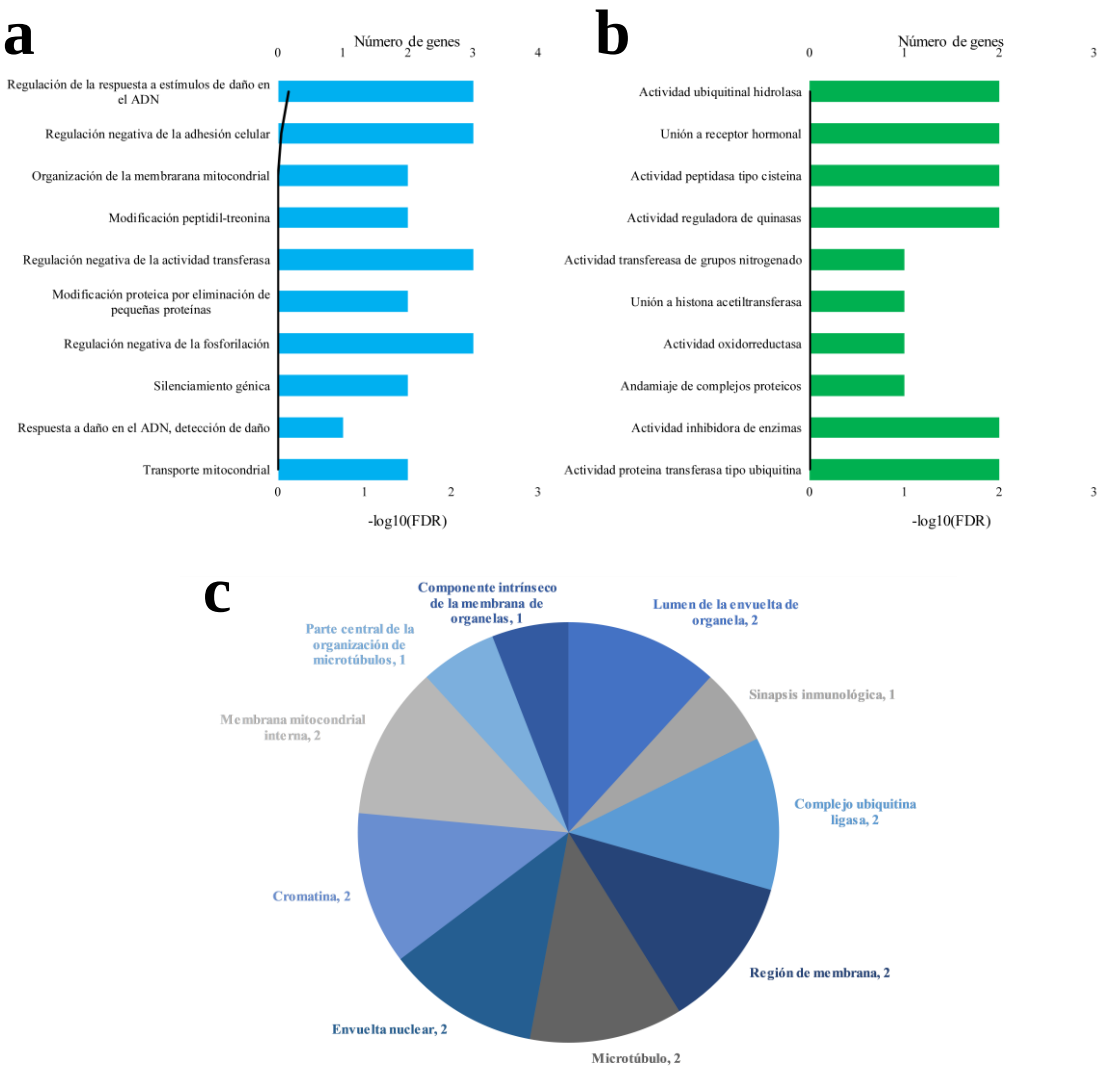



***Figura 4.147.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en el metaanálisis de la respuesta completa en el régimen de tratamiento con bortezomib en monoterapia.* ***a)*** *TOP 10 procesos biológicos GO,* ***b)*** *TOP 10 funciones moleculares GO y* ***c)*** *TOP 10 localizaciones celulares GO.*

Debido a la ausencia de significancia estadística no se llevó a cabo el estudio de las funciones asociadas con los resultados del metaanálisis, pero sí de los genes desregulados presentes en los dos PB representados en un mayor número de genes. Ante la igualdad en el número de genes de cuatro PB, se decidió aplicar un segundo criterio de selección escogiendo los dos PB con menores valores del FDR. Estos dos PB fueron la "regulación de la respuesta a estímulos de daño del ADN" (FDR = 0,757) y la "regulación negativa de la adhesión celular" (FDR = 0,916). En lo que al primer PB respecta, agrupó a tres genes sobreexpresados en los pacientes con RC: *SPRED1* (*z*-valor = 2,23, *p*-valor = 0,0259), *USP1* (*z*-valor = 2,41, *p*-valor = 0,0158) y *BCLAF1* (*z*-valor = 5,67, *p*-valor < 0,0001). En cuanto al PB de "regulación negativa de la adhesión celular" recogió otros tres genes, dos de ellos infraexpresados en los pacientes con RC, *BCAS3* (*z*-valor = -3,42, *p*-valor = 0,0006), *UBASH3B* (*z*-valor = -3,75, *p*-valor = 0,0002); y un gen sobreexpresado: *SOCS6* (*z*-valor = 2,25, *p*-valor = 0,0246). Todos estos genes están

implicados de una u otra forma en procesos carcinogénicos. De este modo, los genes *BCAS3* y *UBASH3B* promueven la tumorogénesis y la metástasis en cáncer de mama, respectivamente[534, 535], *SOCS6* inhibe la proliferación celular en cáncer de estómago[536], *BCLAF1* favorece la muerte celular en MM[537], *SPRED1* actúa como supresor tumoral en melanoma de mucosas[538] y en el caso del gen *USP1*, su inhibición reduce la viabilidad de las células de MM[539]. En el caso de los cinco primeros genes, el sentido de su expresión en nuestro trabajo podría estar promoviendo un fenotipo antitumoral del que se beneficiaría el bortezomib para llevar a cabo su mecanismo de acción de una manera más eficaz y conseguir la RC de los pacientes. Sin embargo, en el caso del gen *USP1* su sobreexpresión no concordaría con esta hipótesis, ya que niveles altos de este gen al diagnóstico en el grupo que alcanzó RC estaría promoviendo la viabilidad de las células mielomatosas gracias a su función de desubiquitinación de proteínas. Esta paradoja debe ser investigada en profundidad para determinar qué papel puede tener este gen en la respuesta a bortezomib.

Finalmente, para completar el análisis de la RC se procedió al estudio del sesgo de publicación sobre la lista de 221 genes. La regresión de Egger reveló un posible sesgo de publicación en 71 genes ($p$-valor < 0,05), suponiendo este valor un 32,1% de los genes estudiados. Este resultado no fue considerado como representativo del conjunto total de genes analizados.

## 4.4.2. Terapias basadas en el uso de bortezomib

El análisis de la respuesta a tratamiento en este apartado se llevó a cabo sobre las series de pacientes cuyo régimen de tratamiento estuvo basado en el bortezomib, independientemente de si este régimen se aplicó en monoterapia o en combinación con otros compuestos, siempre que no fuesen IMiDs. De esta manera, fueron incluidos todos los estudios previamente analizados en el **Apartado 4.4.1.** Además, se incluyeron dos estudios adicionales en el análisis de la RC, que no pudieron ser añadidos a la aproximación OR *vs.* NR debido al bajo número de muestras que presentó el grupo NR. Además, como paso previo al metaanálisis se determinó el número de genes comúnmente interrogados por las plataformas de análisis de la expresión génica de todos los estudios seleccionados, arrojando un total de 15.084 genes comunes.

### 4.4.2.1. Pacientes respondedores *versus* no respondedores

Para realizar este análisis se dispuso de tres series cuyos grupos recogidos en el **Apartado 4.4.1**. El análisis llevado a cabo en este apartado fue muy similar al ya citado del **Apartado 4.4.1**, con la salvedad de que la serie CoMMpass (2017) recogió un número más amplio de pacientes al ser considerados todos los regímenes de tratamiento que incluyesen bortezomib, a excepción de las combinaciones con IMiDs. El otro punto de exclusividad de este análisis respecto al del **Apartado 4.4.1** es la lista inicial de genes, ya

que al haber sido incluidas dos nuevos estudios en el estudio de la RC, la lista de genes de partida comunes a todos los estudios es ligeramente diferente, comprendiendo 15.084 genes, frente a los 15.855 del **Apartado 4.4.1**. En cuanto a las ratios entre el número de muestras de los dos grupos de respuesta, los estudios Amin (2014) y Mulligan (2007) conservan los expuestos en el **Apartado 4.4.1.1**. En lo relativo al estudio CoMMpass (2017), el grupo de pacientes NR presentó un número de muestras muy inferior al recogido en el grupo de pacientes OR, de manera que la ratio OR/NR fue de 10,29. Las características de estas series y de los estudios correspondientes se recogen en la **Tabla 4.19**.

***Tabla 4.19.*** *Estudios seleccionados para el metaanálisis de la respuesta en pacientes tratados con regímenes de tratamiento basados en bortezomib en monoterapia o en cualquier combinación con otros fármacos salvo IMiDs.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | OR | NR |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 144 | 14 |
| GSE55145 | Amin, 2014[197] | Affymetrix Human Exon 1.0 ST Array | 39 | 28 |
| GSE9782 | Mulligan, 2007[528] | Affymetrix Human Genome U133A y U133B Arrays | 73 | 96 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org). OR = pacientes respondedores; NR = pacientes no respondedores.*

El análisis de expresión génica diferencial se realizó sobre los 15.084 genes comunes a los tres estudios previamente seleccionados mediante la comparación de la expresión génica del grupo de pacientes OR frente a la expresión del grupo de NR. De este modo, en el estudio CoMMpass (2017) fueron seleccionados 1.513 genes, de los que 644 genes presentaron sobreexpresión y 869 infraexpresión en los pacientes respondedores. En el caso del estudio de Amin (2014), se detectaron 1.211 genes de los que 611 estuvieron sobreexpresados y 600 infraexpresados. Finalmente, en el estudio de Mulligan (2007) se seleccionaron 3.453 genes no duplicados, de los que 2.157 presentaron sobreexpresión, 1.245 infraexpresión y 51 fueron interrogados por dos *probesets* con diferentes sentidos de expresión génica. En todos los casos se estableció el *p*-valor = 0,05 como punto de corte para de selección de genes, de modo que todos los genes por debajo de este valor fueron elegidos para los siguientes análisis. El cruce de los genes seleccionados en los tres estudios se muestra en el diagrama de Venn de la **Figura 4.148**.

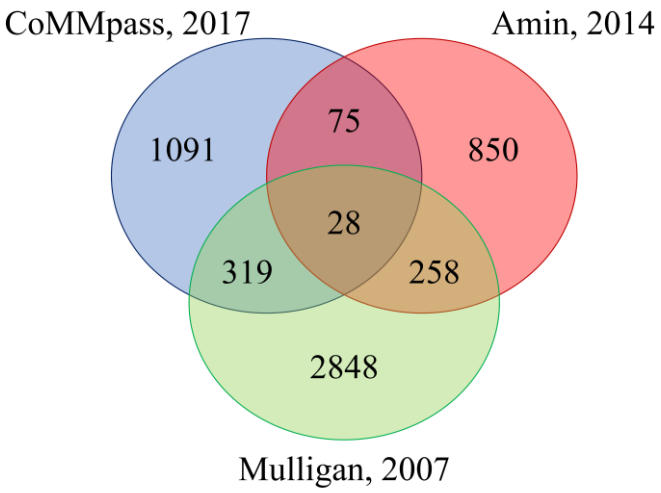



***Figura 4.148.*** *Diagrama de Venn de los genes seleccionados en los tres estudios para la comparación de expresión génica diferencial entre los pacientes respondedores frente a no respondedores. El régimen de tratamiento aplicado en todos los estudios estuvo basado en bortezomib, o bien en monoterapia, o bien en combinación con cualquier otro compuesto farmacológico, salvo IMiDs.*

El estudio mediante metaanálisis de la expresión génica en los tres estudios se realizó sobre los 680 genes que fueron comunes, al menos, a dos de los estudios, revelando un tamaño del efecto estadísticamente significativo ($p$-valor < 0,05) en 143 genes, de los que 91 presentaron sobreexpresión y 52 infraexpresión en los pacientes OR (**Anexo 39**). Tras la selección de los genes estadísticamente significativos, se procedió a su estudio detallado mediante el análisis ORA sobre rutas y funciones biológicas. El resultado considerando las 10 rutas biológicas KEGG más relevantes, así como las 10 funciones GO a los niveles de PB, FM y CC, aparece recogido en la **Figura 4.149**.

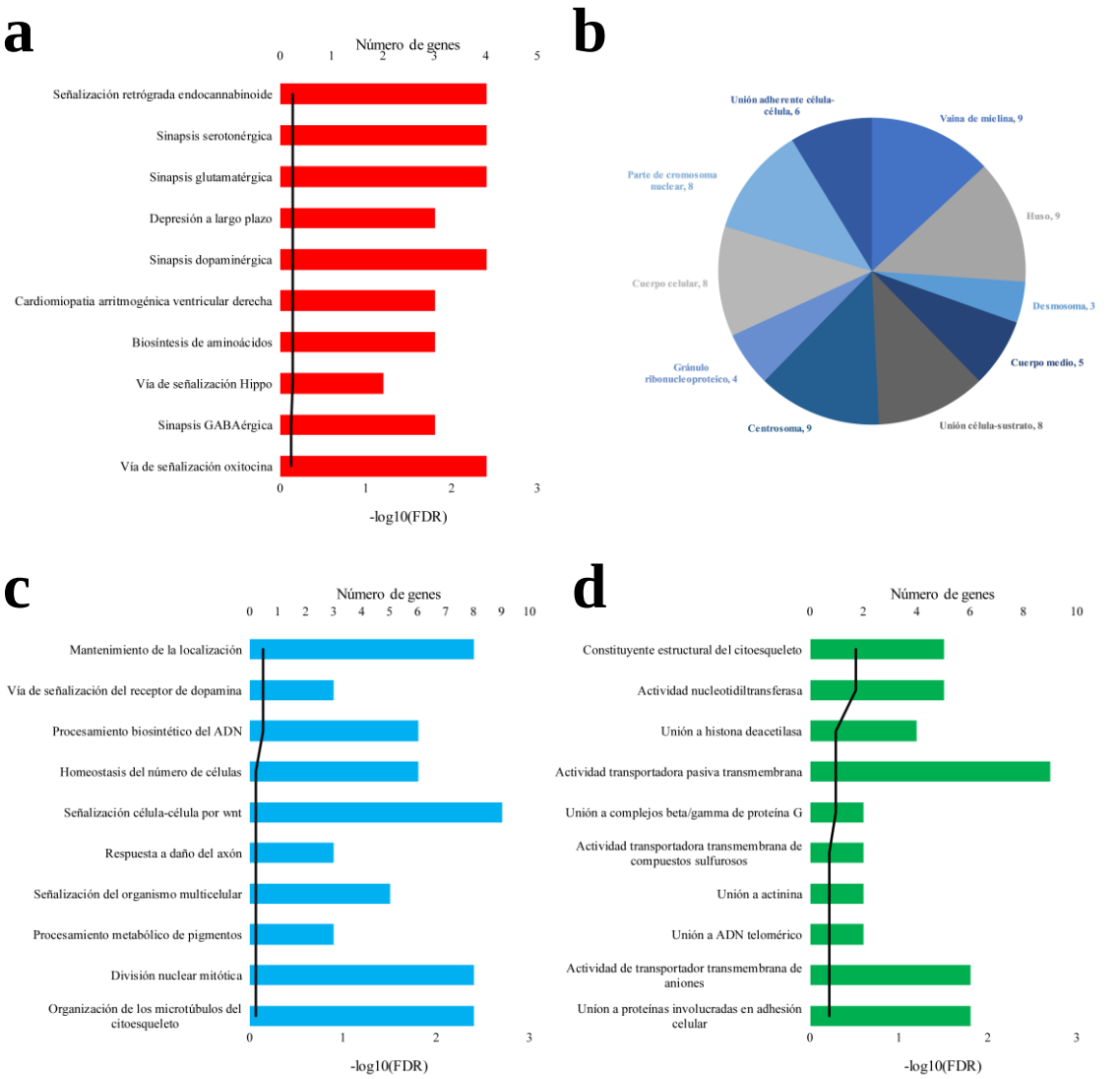



***Figura 4.149.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en los tres estudios seleccionados para el metaanálisis de la respuesta al tratamiento con regímenes de tratamiento basados en bortezomib, excepto aquellos en combinación con IMiDs.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localizaciones celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Ningún término del análisis ORA presentó una sobrerrepresentación estadísticamente significativa a FDR < 0,05. Por este motivo, se decidió analizar las características de los genes implicados en los dos términos representados en un mayor número de genes, que fueron, el PB "señalización célula-célula por wnt" (FDR = 0,8642), y la FM "actividad transportadora pasiva transmembrana" (FDR = 0,5197) con 9 genes cada uno. Entre los genes que aparecen desregulados en el primero de estos términos GO se encontraba el supresor tumoral *CYLD* (*z*-valor = 2,37, *p*-valor = 0,0176)[540]. En MM, la pérdida de expresión de este gen, tanto a nivel de ARNm, como proteico se ha asociado a la progresión de la enfermedad y a un pronóstico adverso, debido principalmente a la activación de vías como la vía de señalización NF-κB, a través del incremento de la localización nuclear de las proteínas p50 y p65[541]. Precisamente, el gen que codifica la

proteína p50, *NFKB1*, aparece infraexpresado de manera estadísticamente significativa en pacientes OR (*z*-valor = -2,23, *p*-valor = 0,0256). Por tanto, la sobreexpresión de *CYLD* unida a la infraexpresión de *NFKB1* indicaría una inactivación de la vía de señalización de NF-κB en los pacientes OR, que en contraposición al mecanismo de resistencia que promovería la activación de esta vía[542], podría estar favoreciendo la respuesta al tratamiento.

En cuanto a la FM "actividad transportadora pasiva transmembrana", los 9 genes que presentan esta actividad aparecen sobreexpresados en los pacientes respondedores. La mayor parte de estos genes son codificantes de proteínas que forman canales transportadores de distintos iones como $Cl^-$ (*CLCN3* y *CLIC5*), $Na^+$ (*SCN8A*), $Ca^{2+}$ (*TRPC1*) o $K^+$ (*KCNK6*) y también aparecen intercambiadores de iones $Na^+/K^+$ como *SLC24A3*. Algunas de estas proteínas, como TRPC1, pueden tener funciones relevantes en la respuesta a drogas de los MM mediante la modificación de la permeabilidad de la membrana[543].

Finalmente, como último paso del estudio con metaanálisis, se revisó la posibilidad de la existencia de sesgo de publicación mediante la regresión de Egger sobre los estudios de metaanálisis de los 680 genes detectándose la presencia de este sesgo a *p*-valor < 0,05 en 152 genes. Esto supone un 22,3% de los genes analizados, con lo que no se consideró representativo del conjunto total de genes.

#### 4.4.2.2. Pacientes que alcanzan respuesta completa *versus* resto

El estudio de la RC en pacientes tratados con regímenes basados en bortezomib se llevó a cabo sobre cinco estudios. La estratificación de la respuesta llevada a cabo sobre estos cinco estudios ya ha sido descrita en el **Apartado 4.4.1**. En lo relativo a las ratios entre los grupos de respuesta de este análisis, los estudios de Amin (2014) y Mulligan (2007) conservaron las mismas ratios que en el **Apartado 4.4.1.2**. Sin embargo, en lo que respecta a los tres estudios restantes, dos de ellos, López-Corral (2014) y Gutiérrez (2010), presentaron un gran equilibrio entre los dos grupos, con ratios RC/Resto de 1,4 y 1,2. Por su parte, el estudio CoMMpass (2017) presentó una fuerte descompensación de los dos grupos de respuesta, con una ratio RC/Resto = 0,12 (8,3 veces). Las características de los cinco estudios seleccionados para este análisis aparecen recogidas en la **Tabla 4.20**.

***Tabla 4.20.*** *Estudios seleccionados para el metaanálisis de la respuesta completa (RC) en pacientes tratados con regímenes de tratamiento basados en bortezomib en monoterapia o en cualquier combinación con otros fármacos salvo IMiDs.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | RC | Resto |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 17 | 141 |
| GSE55145 | Amin, 2014[197] | Affymetrix Human Exon 1.0 ST Array | 16 | 51 |
| GSE47552 | López-Corral, 2014[189] | Affymetrix Human Gene 1.0 ST Array | 7 | 5 |
| GSE16558 | Gutiérrez, 2010[544] | Affymetrix Human Gene 1.0 ST Array | 7 | 6 |
| GSE9782 | Mulligan, 2007[528] | Affymetrix Human Genome U133A y U133B Arrays | 13 | 156 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org).*

El análisis de expresión diferencial mediante el algoritmo *edgeR* en el estudio CoMMpass (2017) reveló la expresión diferencial de 1.269 genes a *p*-valor < 0,05, de los que 443 mostraron sobreexpresión y 826 infraexpresión en los pacientes con RC. En el caso de los análisis con el algoritmo *limma* en los cuatro estudios restantes, se obtuvieron los resultados que se describen a continuación: en el estudio de Amin (2014) se detectaron 574 genes diferencialmente expresados de los que 292 presentaron sobreexpresión y 282 infraexpresión en RC. En el estudio de López-Corral (2014), 873 genes mostraron expresión diferencial a *p*-valor < 0,05, de los que 463 estuvieron sobreexpresados y 410 infraexpresados en los pacientes que alcanzaron RC. En cuanto al estudio de Gutiérrez (2010), se obtuvieron 1.153 genes a *p*-valor < 0,05, de los que 620 fueron genes sobreexpresados y 533 infraexpresados en los pacientes con RC. Finalmente, en el estudio de Mulligan (2007) se detectaron 1.150 genes sin elementos duplicados, de los que 430 genes estuvieron sobreexpresados en RC, 705 infraexpresados y 15 presentaron al menos dos *probesets* en sentidos opuestos de expresión.

El cruce de las cinco listas de genes mediante diagrama de Venn reveló la inexistencia de genes comúnmente desregulados en los cincos estudios a *p*-valor < 0,05. Los resultados de este cruce se recogen en la **Figura 4.150**.

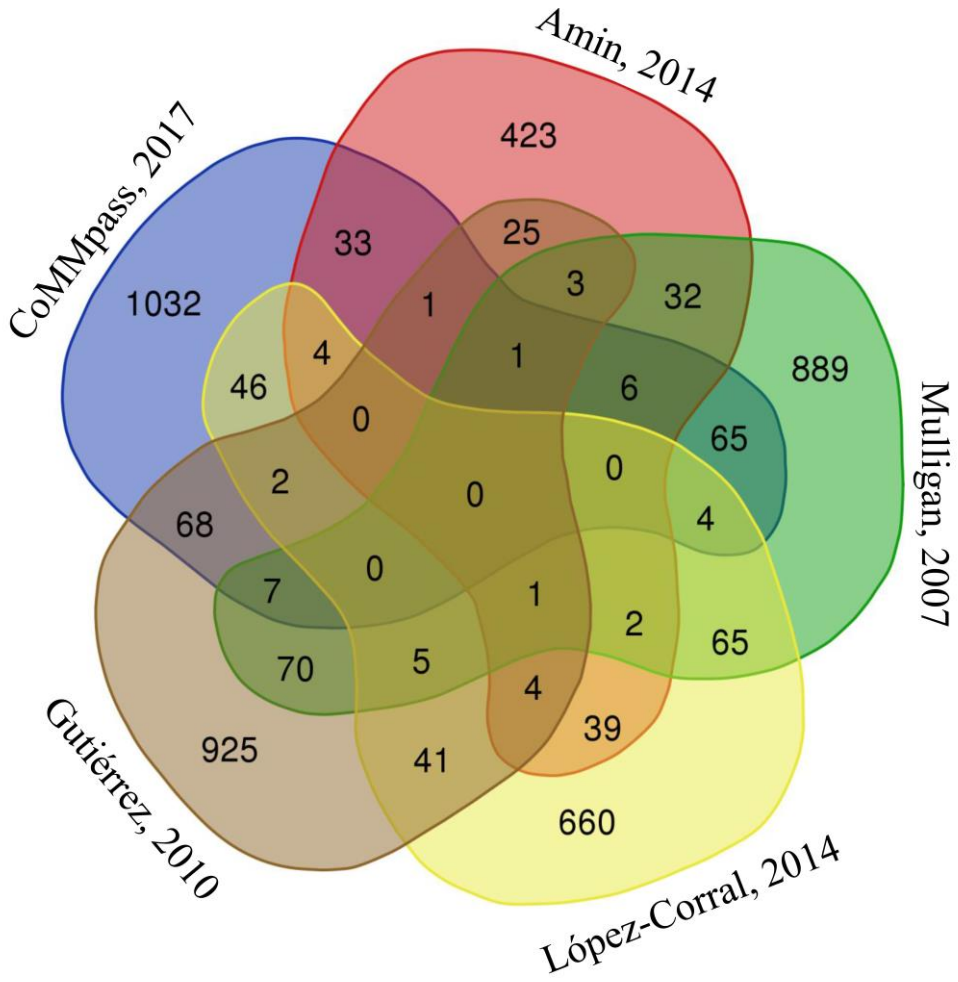



***Figura 4.150.*** *Diagrama de Venn de los genes seleccionados en los cinco estudios en los que se determinó la expresión génica diferencial entre los pacientes que alcanzaron respuesta completa frente al resto de pacientes. El régimen de tratamiento aplicado en todos los estudios estuvo basado en bortezomib, o bien en monoterapia, o bien en combinación con cualquier otro compuesto farmacológico, salvo IMiDs*

A continuación, se realizó un metaanálisis de efectos aleatorios considerando los 542 genes que fueron comunes, al menos, a dos de los estudios. Este análisis detectó 76 genes en los que la diferencia en el tamaño del efecto de los cinco estudios fue estadísticamente significativa a *p*-valor < 0,05, de los que 20 presentaron sobreexpresión y 56 infraexpresión en los pacientes que alcanzaron RC (**Anexo 40**). Se procedió a ejecutar el análisis ORA de genes utilizando como referencia las bases de vías y términos biológicos KEGG y GO sobre estos 76 genes (**Figura 4.151**).

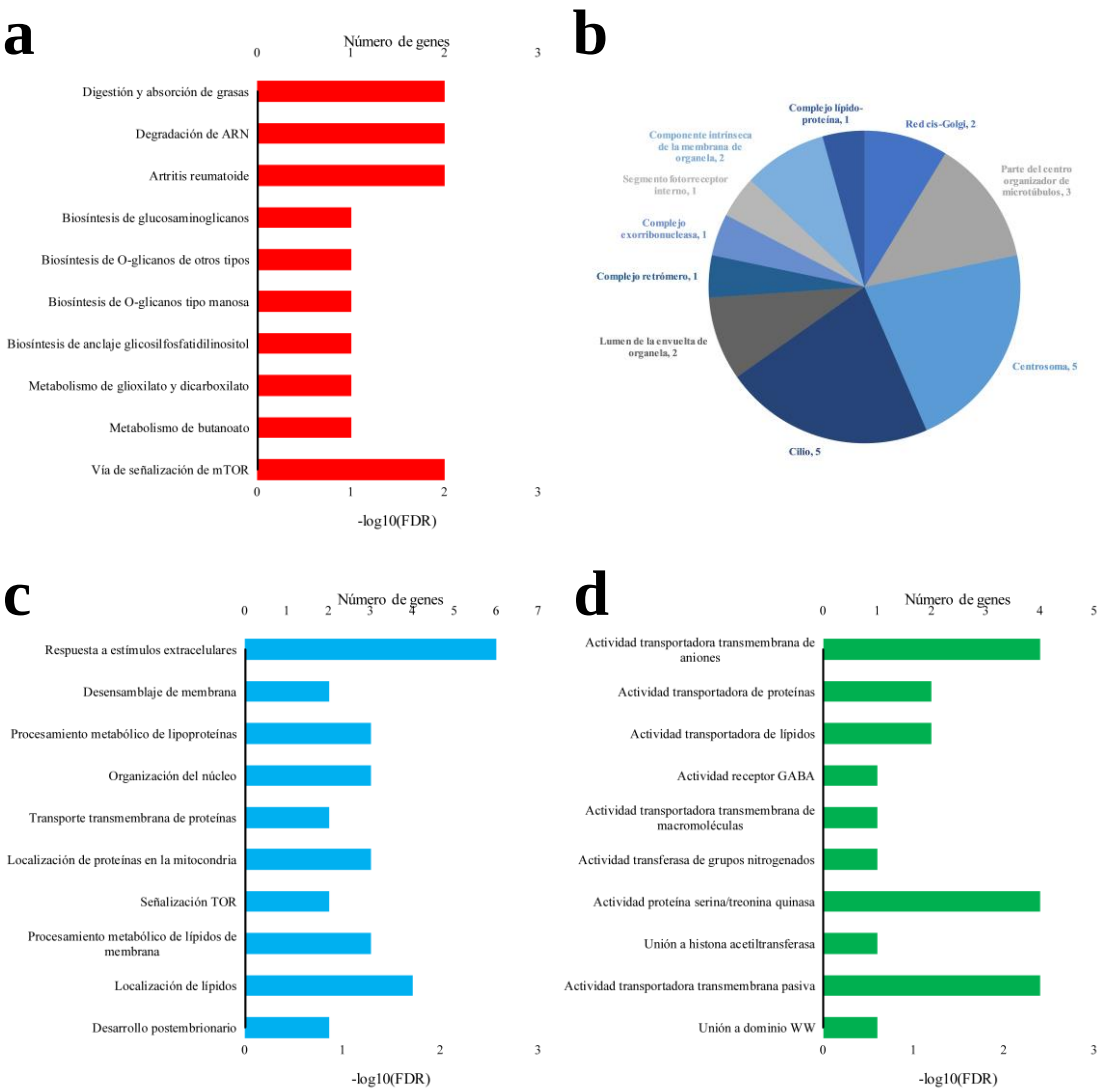



***Figura 4.151.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en el metaanálisis de la respuesta completa al tratamiento con regímenes de tratamiento basados en bortezomib, excepto aquellos en combinación con IMiDs.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localizaciones celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Ninguna de las vías KEGG o términos GO obtuvo resultados estadísticamente significativos a FDR < 0,05. Por esta razón se procedió al estudio de los genes presentes en el PB "respuesta a estímulos extracelulares" (FDR = 1) al ser, con seis genes, el término con un mayor número de genes asociados. Entre estos seis genes, todos ellos infraexpresados en los pacientes con RC, dos de ellos, *ABCA1* (*z*-valor = -4,03, *p*-valor = 0,0001) y *KLF10* (*z*-valor = -3,69, *p*-valor = 0,0002), han sido previamente estudiados en trabajos de MM. En el caso de *ABCA1*, su infraexpresión por fármacos antineoplásicos a nivel de ARNm ha sido asociada con la promoción de la citotoxicidad celular a través de la acumulación intracelular de colesterol[545]. En lo que respecta a *KLF10*, se trata de un gen supresor tumoral, y su infraexpresión ha sido asociada con un incremento de la proliferación celular y de la inhibición de la apoptosis a través de la activación de la vía PTEN/AKT, también en MM[546]. Por tanto, la infraexpresión de ambos genes produce

efectos contrapuestos en la célula mielomatosa, por lo que sería necesaria una investigación más avanzada sobre el posible mecanismo que estos genes puedan tener sobre la célula tumoral del MM.

Por último, se comprobó la posible existencia de sesgo de publicación sobre los 542 genes seleccionados utilizando la regresión de Egger. Este estudio reveló una posible presencia de sesgo en 70 genes, lo que supuso un 12,9% del total de genes analizados, porcentaje inferior al 50% considerado en esta tesis como problemático sobre el conjunto global de genes.

### 4.4.3. Terapia combinada de bortezomib y agentes inmunomoduladores

El análisis realizado en este apartado se llevó a cabo sobre las series de pacientes cuyos regímenes de tratamiento consistieron en el uso combinado del fármaco bortezomib junto con IMiDs. En el caso de los IMiDs, no se hizo distinción en si el compuesto utilizado fue talidomida, lenalidomida o pomalidomida. De esta manera, se seleccionaron cuatro estudios para este análisis, con 16.462 genes interrogados de manera común por las plataformas de análisis de la expresión génica correspondientes a estos estudios.

#### 4.4.3.1. Pacientes respondedores *versus* no respondedores

Se consideraron dos estudios para el análisis de la expresión génica en pacientes OR frente a los pacientes NR: CoMMpass (2017) y Terragna (2016). En el caso del estudio CoMMpass (2017), la estratificación de la respuesta ya fue descrita anteriormente en el **Apartado 4.4.1**. Respecto al estudio de Terragna (2016) la estratificación se realizó de la misma manera, considerando pacientes OR los que alcanzaron como mínimo RP, mientras que el grupo de pacientes NR estuvo formado exclusivamente por pacientes con EE. Ambos estudios presentaron un fuerte desequilibrio entre los grupos de respuesta, de modo que las ratios OR/NR fueron de 25,4 y 15,9, para CoMMpass (2017) y Terragna (2016), respectivamente, debido a que la mayor parte de los pacientes que reciben este régimen de tratamiento logran algún tipo de respuesta. Las características de estos dos estudios se detallan en la **Tabla 4.21**.

***Tabla 4.21.*** *Estudios seleccionados para el metaanálisis de la respuesta en pacientes tratados con regímenes de tratamiento basados en la combinación bortezomib e IMiDs.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | OR | NR |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 330 | 13 |
| GSE68871 | Terragna, 2016[194] | Affymetrix Human Genome U133 Plus 2.0 Array | 111 | 7 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org). OR = pacientes respondedores; NR = pacientes no respondedores.*

De este modo, en el análisis de expresión diferencial OR *vs.* NR sobre el estudio CoMMpass (2017) se detectaron 1.905 genes diferencialmente expresados a *p*-valor < 0,05, de los que 1.000 estaban sobreexpresados y 905 infraexpresados en los pacientes que respondieron al tratamiento. En cuanto al estudio de Terragna (2016) se identificaron 752 genes a *p*-valor < 0,05, de los que 318 estaban sobreexpresados y 434 infraexpresados en los pacientes OR. El cruce de las dos listas de genes aparece representado a través de un diagrama de Venn en la **Figura 4.152**.

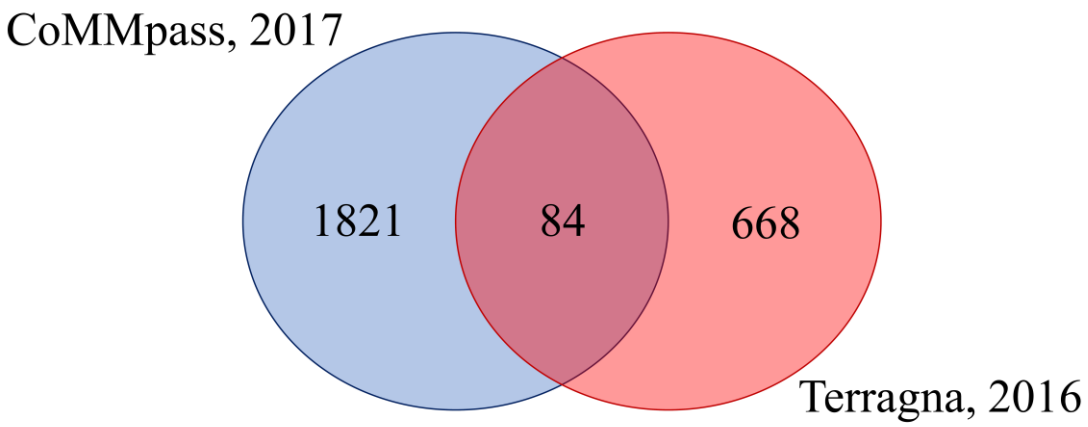



***Figura 4.152.*** *Diagrama de Venn de los genes seleccionados en los dos estudios seleccionados para el análisis de la expresión génica diferencial entre los pacientes respondedores frente a los no respondedores. El régimen de tratamiento aplicado en todos los estudios fue bortezomib en combinación con IMiDs*

Sobre los 84 genes comunes a los dos estudios obtenidos en el cruce de las dos listas de genes se llevó a cabo el metaanálisis de efectos aleatorios, en el que se encontraron 44 genes que presentaban diferencias en el tamaño del efecto estadísticamente significativas (*p*-valor < 0,05) (**Anexo 41**). De estos 44 genes, 13 estaban sobreexpresados y 31 infraexpresados en los pacientes respondedores. Los análisis de sobrerrepresentación sobre esta lista de genes utilizando las bases KEGG y GO aparecen recogidos en la **Figura 4.153**.

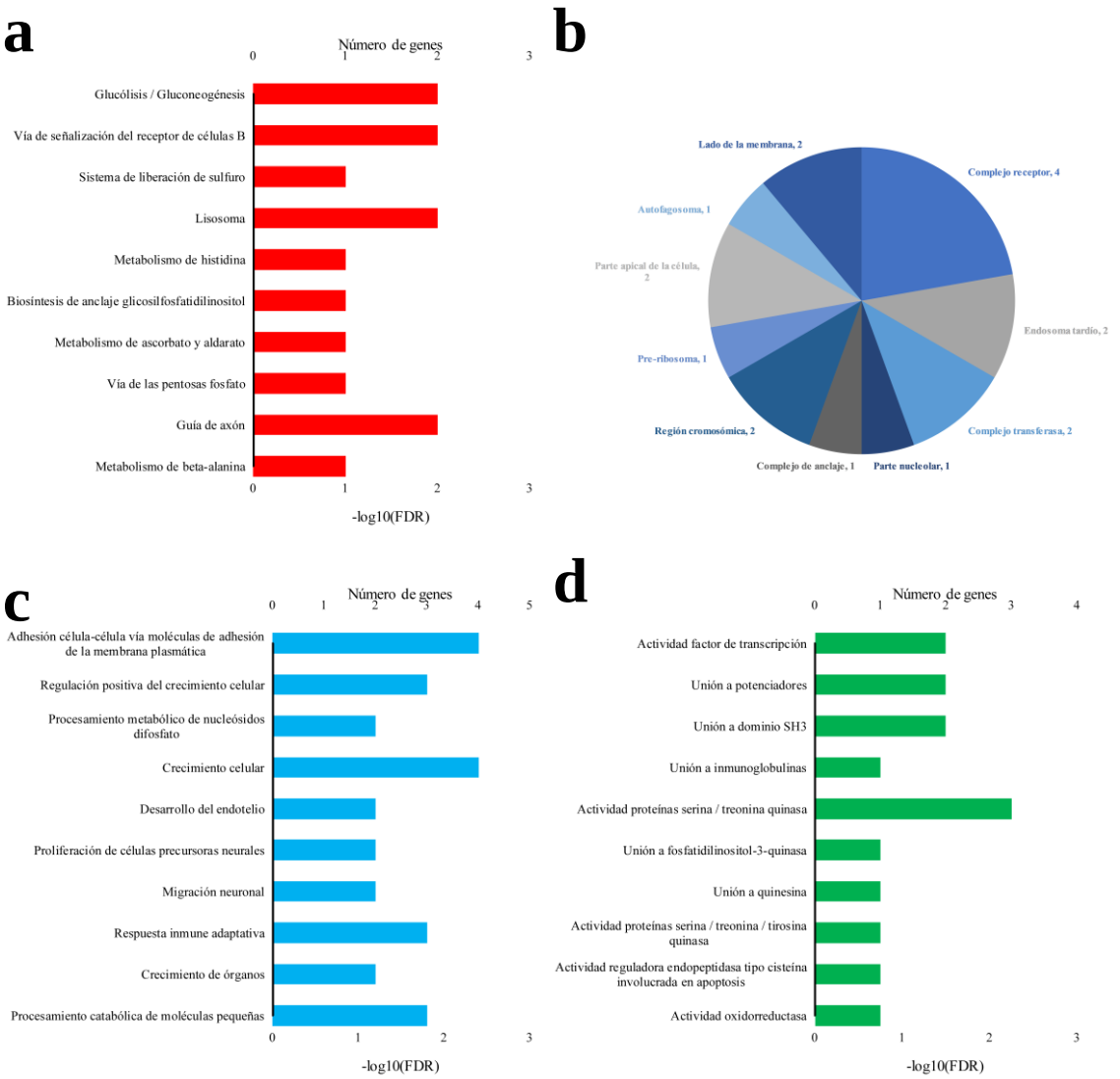



***Figura 4.153.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en los dos estudios seleccionados para el metaanálisis de la respuesta al tratamiento con regímenes de tratamiento basados en bortezomib en combinación con IMiDs.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localizaciones celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y d) TOP 10 funciones moleculares GO.*

Ninguno de los términos GO o vías KEGG resultó estadísticamente significativo, por este motivo se decidió analizar únicamente los genes presentes en los términos que comprendieron un mayor número de genes, que en este caso fueron los PB "adhesión célula-célula vía moléculas de adhesión de la membrana plasmática" (FDR = 1) y el "crecimiento celular" (FDR = 1). Entre los dos PB se recogieron un total de 12 genes desregulados, entre los que se encuentra el gen *ACVR1* (*z*-valor = 3,84, *p*-valor = 0,0001), que tiene un papel muy relevante en la vía de la proteína morfogénica ósea (BMP), implicada en el desarrollo y la reparación del sistema esquelético, y en el caso del MM, en la diferenciación osteoblástica junto con la vía Wnt[547]. Otro de los genes interesantes entre este grupo de genes, fue el gen *RPS6KA1* (*z*-valor = -3,11, *p*-valor = 0,0019), que forma parte de la familia de proteínas quinasas p90RSK, la cual comprende cuatro

isoformas cuya actividad se ha asociado con el control de la proliferación, la migración y la diferenciación celular, entre otras funciones[548]. En el caso concreto del gen *RPS6KA1*, ha sido implicado en procesos como la modulación de la metástasis en cáncer de pulmón[549] o la promoción de la invasión en melanoma[550], sin embargo, su papel en MM no ha sido estudiado, a pesar de que se ha comprobado que compuestos inhibidores de otros miembros de esta familia de genes, como el gen *RPS6KA3*, presentan actividad antimieloma[551].

Como apunte final al metaanálisis llevado a cabo en este apartado, hay que indicar que no fue posible proceder con el análisis de sesgo de publicación mediante la regresión de Egger, ya que esta técnica requiere más de dos estudios para ser ejecutada.

#### 4.4.3.2. Pacientes que alcanzan respuesta completa *versus* resto

Para el análisis de la RC en pacientes tratados con bortezomib en combinación con IMiDs se seleccionaron los estudios CoMMpass (2017), Terragna (2016), López-Corral (2014) y Gutiérrez (2010). El grupo de RC estuvo formado en todos los estudios por los pacientes que tenían como mínimo una IFE negativa; el resto de los pacientes fueron incluidos en el grupo "Resto" en el análisis. En lo relativo al equilibrio de los dos grupos de respuesta, los estudios de López-Corral (2014), con una ratio RC/Resto de 0,8 (1,2 veces), y Gutiérrez (2010), con ratio = 1,5, presentaron un mayor equilibrio que los estudios CoMMpass (2017) (ratio = 0,25 [~4 veces]) y Terragna (2016) (ratio = 0,15 [~6,9 veces]). La diferencia de equilibrio entre los estudios fue probablemente debida al tamaño muestral de los mismos, ya que fueron precisamente los estudios con un mayor número de muestras los que estuvieron menos balanceados. Las características generales de los cuatro estudios aparecen recogidas en la **Tabla 4.22**.

***Tabla 4.22.*** *Estudios seleccionados para el metaanálisis de la respuesta completa (RC) en pacientes tratados con regímenes de tratamiento basados en la combinación bortezomib e IMiDs.*

| Serie | Estudio | Plataforma | N | |
|---|---|---|---|---|
| | | | RC | Resto |
| CoMMpass | CoMMpass, 2017* | Illumina HiSeq2000 o HiSeq2500 | 69 | 274 |
| GSE68871 | Terragna, 2016[194] | Affymetrix Human Genome U133 Plus 2.0 Array | 15 | 103 |
| GSE47552 | López-Corral, 2014[189] | Affymetrix Human Gene 1.0 ST Array | 9 | 11 |
| GSE16558 | Gutiérrez, 2010[544] | Affymetrix Human Gene 1.0 ST Array | 6 | 4 |

** Datos generados como parte de la iniciativa de medicina personalizada de la Multiple Myeloma Research Foundation (MMRF, https://research.themmrf.org y www.themmrf.org).*

El estudio de expresión diferencial sobre estos cuatro estudios mostró alta variabilidad en cuanto al número de genes detectados a *p*-valor < 0,05. Así, el estudio en el que se detectó un mayor número de genes fue en el estudio CoMMpass (2017) donde fueron detectados 1.921 genes de los que 907 estaban sobreexpresados y 1.014

infraexpresados. En los estudios de Terragna (2016) y López-Corral (2014) el nivel de detección fue intermedio con 1.041 genes (459 sobreexpresados y 582 infraexpresados) y 963 genes (686 sobreexpresados y 277 infraexpresados) detectados, respectivamente. La menor detección se produjo en el estudio de Gutiérrez (2010) con 401 genes detectados a *p*-valor < 0,05 (175 sobreexpresados y 226 infraexpresados). El cruce de las cuatro listas de genes además reveló la inexistencia de genes comunes a los cuatro estudios, aunque sí que hubo coincidencias entre parejas y tríos de estudios, tal como puede observarse en la **Figura 4.154**.

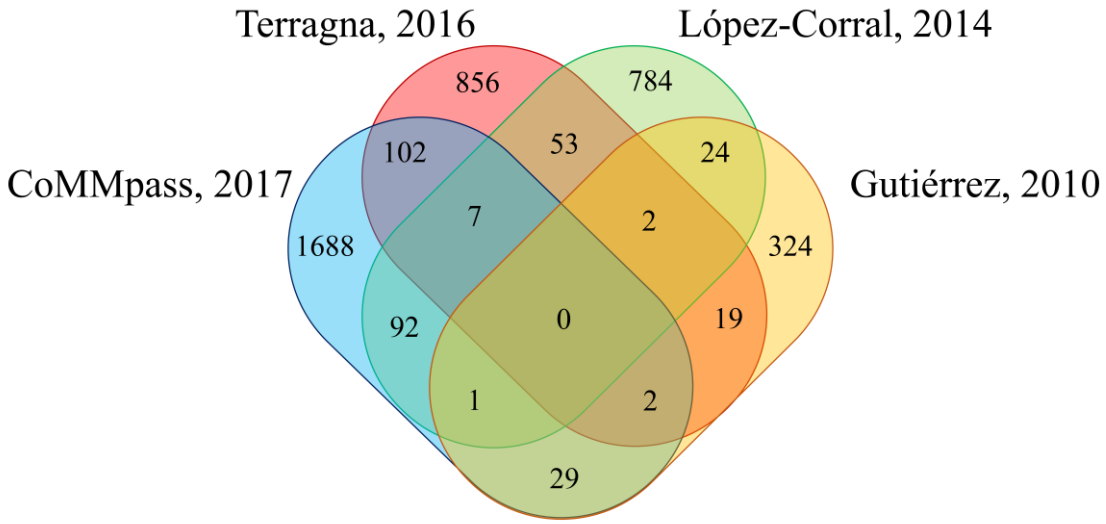



***Figura 4.154.*** *Diagrama de Venn de los genes seleccionados en los cuatro estudios para la comparación de expresión génica diferencial entre los pacientes que alcanzaron RC frente al resto de pacientes. El régimen de tratamiento aplicado en todos los estudios estuvo basado en bortezomib combinado con IMiDs*

A continuación, se llevó a cabo el metaanálisis de efectos aleatorios considerando los 331 genes comunes a, al menos, dos de los estudios. Se determinaron 54 genes cuyo tamaño del efecto fue estadísticamente significativo a un *p*-valor < 0,05 (**Anexo 42**). De los 54 genes, 31 presentaron sobreexpresión en los pacientes que alcanzaron RC, mientras que 23 genes estuvieron infraexpresados en estos pacientes. Sobre estos genes se llevó a cabo el análisis de sobrerrepresentación en vías KEGG y términos GO, cuyos resultados se recogen en la **Figura 4.155**.

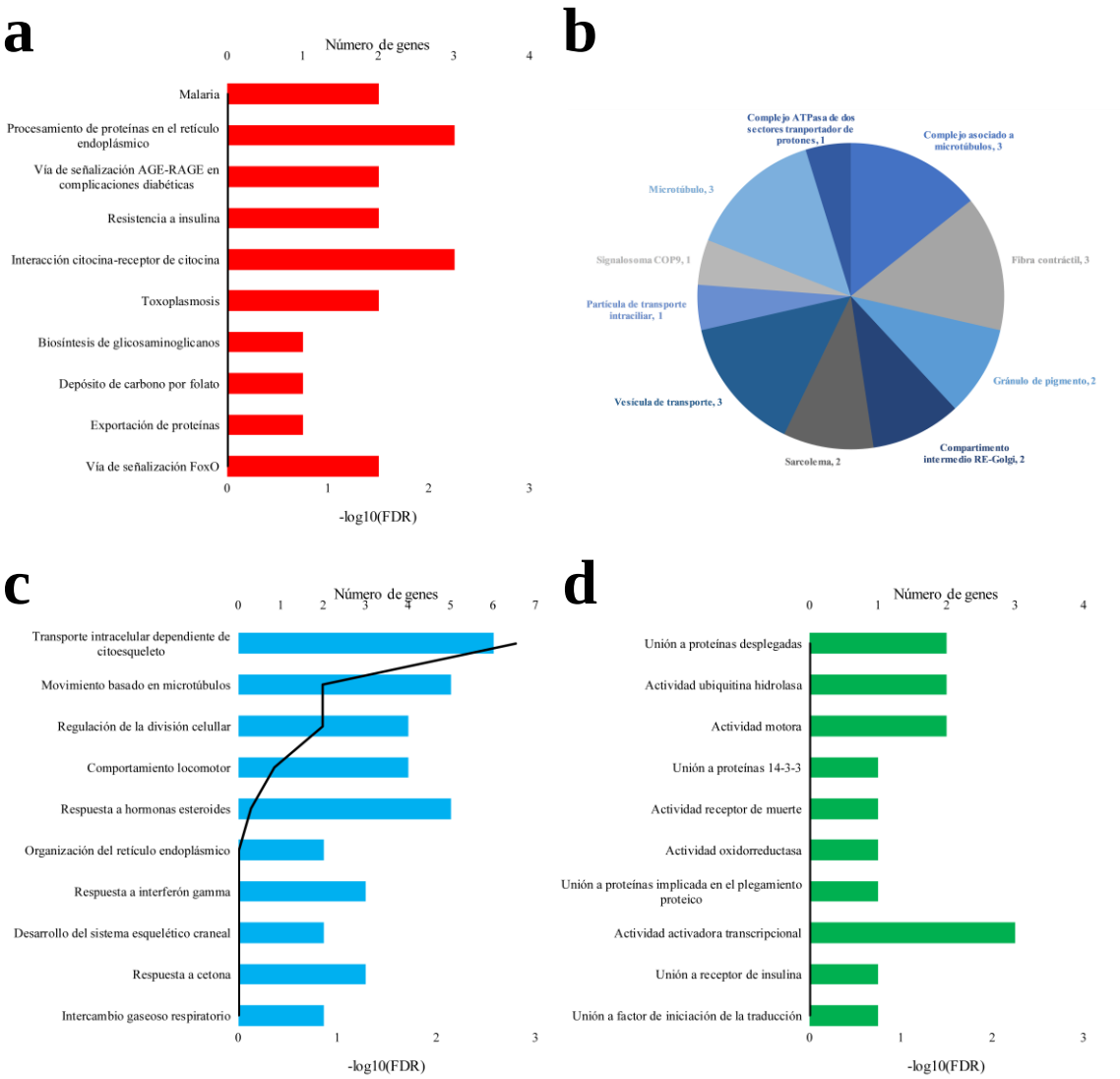



***Figura 4.155.*** *Análisis de sobrerrepresentación sobre vías KEGG y ontologías génicas (GO) considerando los genes estadísticamente significativos en el metaanálisis de la respuesta completa a regímenes de tratamiento con bortezomib en combinación con IMiDs.* ***a)*** *TOP 10 rutas biológicas KEGG,* ***b)*** *TOP 10 localizaciones celulares GO,* ***c)*** *TOP 10 procesos biológicos GO y* ***d)*** *TOP 10 funciones moleculares GO.*

Únicamente el PB "transporte intracelular dependiente de citoesqueleto" resultó estadísticamente significativo con un FDR = 0,0016 y representado por seis genes, todos ellos sobreexpresados en los pacientes con RC. De esta manera, la sobreexpresión de genes como *KIF16B* (*z*-valor = 5,54, *p*-valor < 0,0001), implicado en el transporte de endosomas[552], así como de otros genes involucrados en diversos tipos de transporte intracelular, como *BBS12* (*z*-valor = 2,26, *p*-valor = 0,0241), *APBA1* (*z*-valor = 2,97, *p*-valor = 0,0029), *MYRIP* (*z*-valor = 3,49, *p*-valor = 0,0005) o *KIFAP3* (*z*-valor = 2,72, *p*-valor = 0,0064), junto con el conocimiento previo de que genes como los que codifican proteínas de la familia RAB, implicada en la regulación del tráfico intracelular de vesículas y el transporte de proteínas[553], conducen a la sensibilidad a compuestos como la lenalidomida[554], podrían sugerir la importancia de los procesos de transporte intracelular en la respuesta a IMiDs.

# 4.5. Capítulo 5.

## Predicción de la respuesta al tratamiento en pacientes con mieloma múltiple.

El análisis de predicción de la respuesta se realizó sobre los mismos grupos de tratamiento que en el **Apartado 4.4**. En este caso, cada grupo de tratamiento fue estratificado en función de la respuesta alcanzada por los pacientes en las siguientes aproximaciones analíticas:

1) Pacientes respondedores versus no respondedores
2) Pacientes que alcanzan respuesta completa versus resto de pacientes
3) Respuesta codificada en tres grupos
4) Respuestas múltiples o multirrespuesta

Como punto de partida para los análisis de predicción en las aproximaciones 1) y 2) se seleccionaron los genes estadísticamente significativos ($p$-valor < 0,05) de los metaanálisis correspondientes del **Apartado 4.4**. En el caso de las aproximaciones 3) y 4) se seleccionaron los genes que fueron estadísticamente significativos a $p$-valor < 0,05 en el análisis con los algoritmos *limma* o *edgeR* considerando múltiples grupos (> 2 grupos). Para la aproximación 3), tras hacer esta selección estadística de genes, se procedió al cruce de las listas de genes de cada uno de los estudios incluidos, de manera que fueron seleccionados para el análisis de predicción aquellos genes que fuesen comunes en al menos dos de los estudios. Respecto a la aproximación 4), fue un caso particular, ya que debido a que cada estudio tiene su propia codificación de la respuesta no fue posible obtener una lista de genes comunes para todos los estudios. Por tanto, la predicción se realizó sobre cada estudio considerando únicamente su propia lista de genes diferencialmente expresados procedentes del análisis de múltiples grupos. Estas predicciones aparecerán en cada apartado con el nombre de "inicial".

En todas las aproximaciones de predicción, se procedió a una segunda predicción utilizando un filtrado de genes en función del grado de importancia (VIM) sobre la respuesta en las matrices de entrenamiento utilizando el método de bosques aleatorios o *random forest* en el paquete *Boruta* en R. La importancia o VIM es un parámetro que determina qué variables son conductoras de una determinada respuesta. Las predicciones bajo estas condiciones serán denominadas en cada apartado como "boruta".

Por último, se llevó a cabo una tercera predicción con los genes de una nueva lista creada a partir de las listas de genes filtrados por *Boruta* de todos los estudios de un mismo grupo de respuesta para un mismo régimen de tratamiento. Estas predicciones serán denominadas "sumatorio".

En todos los casos para proceder con el análisis de predicción se utilizó una matriz o serie de entrenamiento del modelo, constituida por tres cuartas partes de los pacientes en cada estudio y una matriz de validación del modelo que comprendió el tercio restante de pacientes. La selección de pacientes para una u otra matriz se hizo de manera aleatoria.

Todas las predicciones se efectuaron utilizando cinco métodos predictivos, siempre que fuese posible:

a) Máquinas de soporte vectorial con pesos estadísticos (wSVM)
b) Mínimos cuadrados parciales (PLS)
c) Máquinas de soporte vectorial sin pesos estadísticos (SVM)
d) K vecinos más cercanos (KNN)
e) Bosques aleatorios (RF)

En todos los análisis de predicción se considerará como mejor predicción aquella que alcanzando la tasa de acierto global más elevada, haya sido realizada con un menor número de genes. En caso de igualdad entre dos o más métodos de predicción, se determinará como óptima aquella predicción en la que la sensibilidad o razón de verdaderos positivos (RVP) y la especificidad o razón de verdaderos negativos (RVN), así como el valor predictor positivo (VPP) y el valor predictor negativo (VPN) estén más compensados según el criterio del investigador. Este último criterio estará basado en un aspecto subjetivo para tratar de evitar dar por válida una predicción aparentemente buena en grupos poco equilibrados.

### 4.5.1. Bortezomib en monoterapia

El análisis de predicción para este régimen de tratamiento se llevó a cabo exclusivamente con estudios de expresión génica con datos disponibles de respuesta al tratamiento con bortezomib en monoterapia. En cada una de las cuatro aproximaciones se detallarán los estudios analizados.

***Pacientes respondedores versus no respondedores***

La predicción de la respuesta al tratamiento con bortezomib en monoterapia se llevó a cabo sobre los estudios CoMMpass (2017), Amin (2014) y Mulligan (2007). El número de muestras utilizadas en las matrices de entrenamiento y validación en cada uno de estos tres estudios se recoge en la **Tabla 4.23**. Los tres estudios no mostraron un buen equilibrio en el número de muestras en los dos grupos de respuestas, con ratios OR/NR de 2,00 en CoMMpass (2017), 1.39 en Amin (2014) y 0,76 en el estudio de Mulligan (2007).

***Tabla 4.23****. Número de muestras en los grupos de pacientes respondedores y no respondedores en los estudios seleccionados para la predicción de la respuesta en pacientes tratados con bortezomib en monoterapia.*

| Tipo de respuesta | Matriz | CoMMpass (2017) | Amin (2014) | Mulligan (2007) |
|---|---|---|---|---|
| Responden | Entrenamiento | n = 4 | n = 26 | n = 49 |
| | Validación | n = 2 | n = 13 | n = 24 |
| No responden | Entrenamiento | n = 2 | n = 19 | n = 64 |
| | Validación | n = 1 | n = 9 | n = 32 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación*

Como lista "inicial" de genes para la predicción se utilizaron los 233 genes estadísticamente significativos del metaanálisis del **Apartado 4.4.1.1**. Sobre esta lista se aplicó en cada uno de los tres estudios el filtrado con el algoritmo *Boruta*, resultando en la selección de 16 y 20 genes en los estudios de Amin (2014) y Mulligan (2007), respectivamente. En el caso del estudio CoMMpass (2017) no fue posible aplicar este algoritmo debido al bajo tamaño muestral. Finalmente, se construyó una tercera lista de genes en la que se combinaron los genes de las dos listas filtradas de *Boruta* eliminando los genes duplicados. Esta tercera lista de 34 genes fue utilizada de manera común en los tres estudios. Las tres listas de genes resultantes, junto con una batería de cinco métodos predictivos, fueron las herramientas empleadas para llevar a cabo los análisis de predicción en los tres estudios mencionados.

En el caso del estudio CoMMpass (2017), la mejor predicción se alcanzó con los 34 genes de la lista "sumatorio", obteniendo una tasa de acierto global (TAC) del 67% con todos los métodos predictivos utilizados. Sin embargo, la especificidad se quedó en un 0% y el VPN no pudo ser calculado (**Anexo 43**), lo que indica que no fue capaz de clasificar correctamente los pacientes del grupo de no respuesta.

El estudio de Amin (2014), por su parte obtuvo una TAC del 82% con los 16 genes filtrados por *Boruta*, utilizando el método PLS con dos factores. La sensibilidad de este modelo fue del 92% y su especificidad del 67%, lo que indica que tiene una efectividad del 92% a la hora de identificar pacientes respondedores y un 67% de efectividad identificando pacientes que no responden. En cuanto a los parámetros VPP y VPN, se obtuvieron unos valores del 80% y del 86%, respectivamente. Por tanto, la probabilidad de que un paciente alcance realmente alguna respuesta si el modelo indica que el paciente responde es del 80%, mientras que la probabilidad de que un paciente no responda si el modelo indica no respuesta es del 86% (**Anexo 43**).

En lo relativo al estudio de Mulligan (2007), se obtuvo que el modelo óptimo fue el determinado a través del método KNN utilizando los 21 vecinos más próximos y los 20 genes seleccionados a través de Boruta. Esta predicción arrojó una TAC del 77% sobre la serie de validación. La sensibilidad y la especificidad de este modelo fueron del 54%y del 94%, indicando que dicho modelo tiene un 54% y un 94% de efectividad identificando pacientes respondedores y no respondedores, respectivamente. Respecto al VPP, su valor se sitúa en el 87%, y en el caso del VPN en el 73%.

Los resultados de la predicción óptima que alcanzó una mejor TAC para los tres estudios se muestran para su comparación en la **Figura 4.156**.

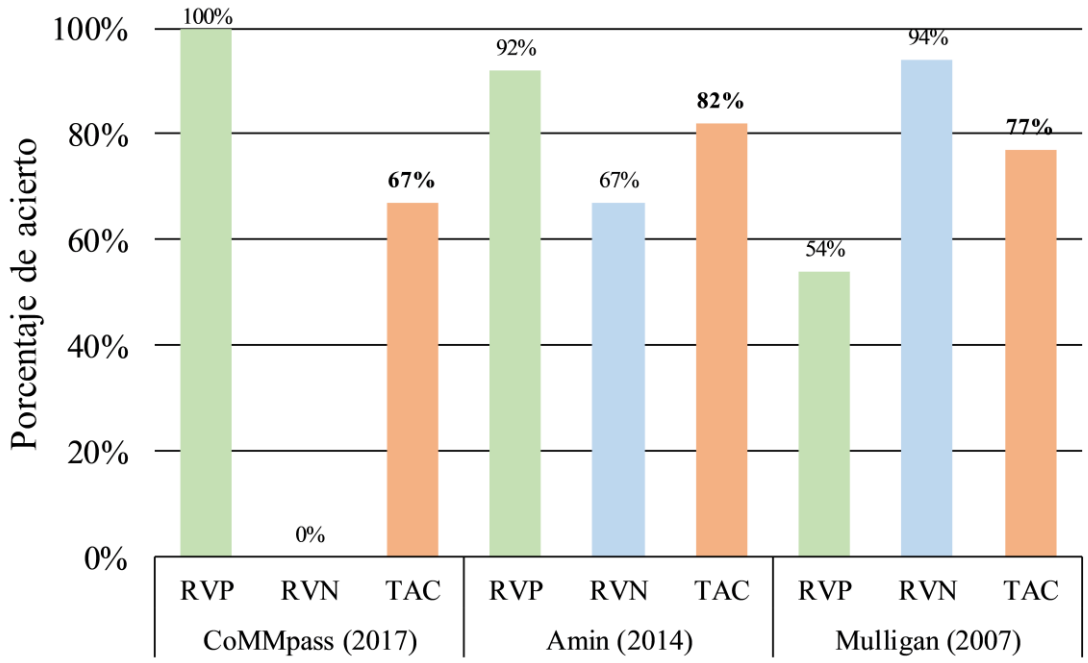



***Figura 4.156.*** *Resultados de la predicción óptima de la respuesta a bortezomib en monoterapia. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

Aunque los resultados de las predicciones de las respuestas sobre la serie CoMMpass (2017) no son óptimos al no capturar el grupo de pacientes no respondedores, los resultados en las series de Amin (2014) y Mulligan (2007) muestran la buena capacidad de estos modelos sobre la predicción de la respuesta, con altas tasas de acierto y de forma compensada, tanto en el grupo de pacientes respondedores, como en el grupo de no respondedores. Una de las posibles causas de la baja tasa de acierto en el grupo de pacientes no respondedores en el estudio CoMMpass (2017) podría ser el bajo tamaño muestral, ya que la matriz de validación solamente constó de una única muestra en este grupo de respuesta, conduciendo por tanto a un resultado poco fiable debido a la elevada arbitrariedad que conlleva el estudio de una sola muestra.

***Pacientes que alcanzan respuesta completa versus resto***

Se seleccionaron tres estudios para la predicción de la respuesta completa (RC) frente al tratamiento con bortezomib en monoterapia: CoMMpass (2017), Amin (2014) y Mulligan (2007). El número de muestras disponible en cada uno de estos tres estudios en sus respectivas matrices de entrenamiento y validación se recoge en la **Tabla 4.24**. En cuanto a la ratio de pacientes que alcanzaron RC frente a los que no la alcanzaron, se observaron en algunos casos grandes desequilibrios. El estudio más descompensado en cuanto a la ratio RC/Resto fue el estudio de Mulligan (2007) con una ratio de 0,08 (~12 veces). En el caso del estudio de Amin (2014) también se observó un cierto desequilibrio, con una ratio de 0,30 (~3,38 veces). Por último, el estudio más equilibrado en muestras

entre los dos grupos estudiados fue CoMMpass (2017), con una ratio de 0,5 (~2 veces), aunque en este último el tamaño muestral fue mucho menor.

***Tabla 4.24.*** *Número de muestras en los grupos de pacientes que alcanzaron respuesta completa (RC) y del resto de pacientes en los estudios seleccionados para la predicción de la RC en pacientes tratados con bortezomib en monoterapia.*

| Tipo de respuesta | Matriz | CoMMpass (2017) | Amin (2014) | Mulligan (2007) |
|---|---|---|---|---|
| RC | Entrenamiento | n = 2 | n = 11 | n = 9 |
| | Validación | n = 1 | n = 5 | n = 4 |
| Resto | Entrenamiento | n = 4 | n = 34 | n = 104 |
| | Validación | n = 2 | n = 17 | n = 52 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

La lista inicial de genes para las predicciones de la RC consistió en los 26 genes procedentes del metaanálisis del **Apartado 4.4.1.2**. El filtrado de esta lista con el algoritmo *Boruta* solamente se llevó a cabo en los estudios de Amin (2014) y de Mulligan (2007), de manera que fueron seleccionados 10 y seis genes, respectivamente. El algoritmo *Boruta* no pudo ser aplicado en la serie CoMMpass (2017) debido a su bajo tamaño muestral. Por último, se procedió a la construcción de una tercera lista ("sumatorio") que consistió en la suma de las dos listas de genes filtrados por *Boruta*, eliminando los elementos duplicados. Estas tres listas, junto con cinco métodos predictivos, fueron utilizados para la predicción de la RC en los tres estudios seleccionados.

El resultado del análisis de predicción para el estudio CoMMpass (2017) fue una TAC del 100% utilizando el método PLS con un factor y los 14 genes de la lista "sumatorio". Sin embargo, pese a ser una tasa de acierto óptima no puede considerarse como un resultado fiable al ser el número de pacientes con RC, tanto de la serie de entrenamiento, como de la serie de validación, extremadamente bajo (el grupo RC solamente contiene una única muestra en la serie de validación).

En el caso del estudio de Amin (2014), el método de predicción que obtuvo mejores resultados fue el PLS con 10 genes filtrados por *Boruta* y dos factores, alcanzando una TAC del 86%. Este modelo además alcanzó una RVP y una RVN del 80% y 88%, con lo que fue capaz de identificar pacientes que hacen RC y pacientes que no la alcanzan con una efectividad del 80 y el 88%, respectivamente. El VPP de este modelo se situó en 0,67 y el VPN en 0,94, por tanto, la probabilidad de que el modelo prediga que un paciente alcance RC o no la alcance, y que realmente esto ocurra, fue del 67% y del 94%, respectivamente.

Para el estudio de Mulligan (2007), los resultados de las predicciones mostraron una buena TAC utilizando wSVM con un kernel lineal, un coste de 10 y gamma de 0,25 con

los 14 genes de la lista "sumatorio". La TAC alcanzó un valor del 95% y una RVP y RVN relativamente equilibradas (75% y 96%, respectivamente). Además, los valores de VPP y VPN fueron aceptables, siendo especialmente elevado este último con un 98%, indicando este valor la probabilidad de que si el modelo predice que el paciente no alcanza RC este realmente no la alcance.

Los resultados completos de estas predicciones se recogen el en **Anexo 43**, mientras que los resultados de la predicción óptima que alcanzó una mejor TAC para los tres estudios se muestran para su comparación en la **Figura 4.157**.

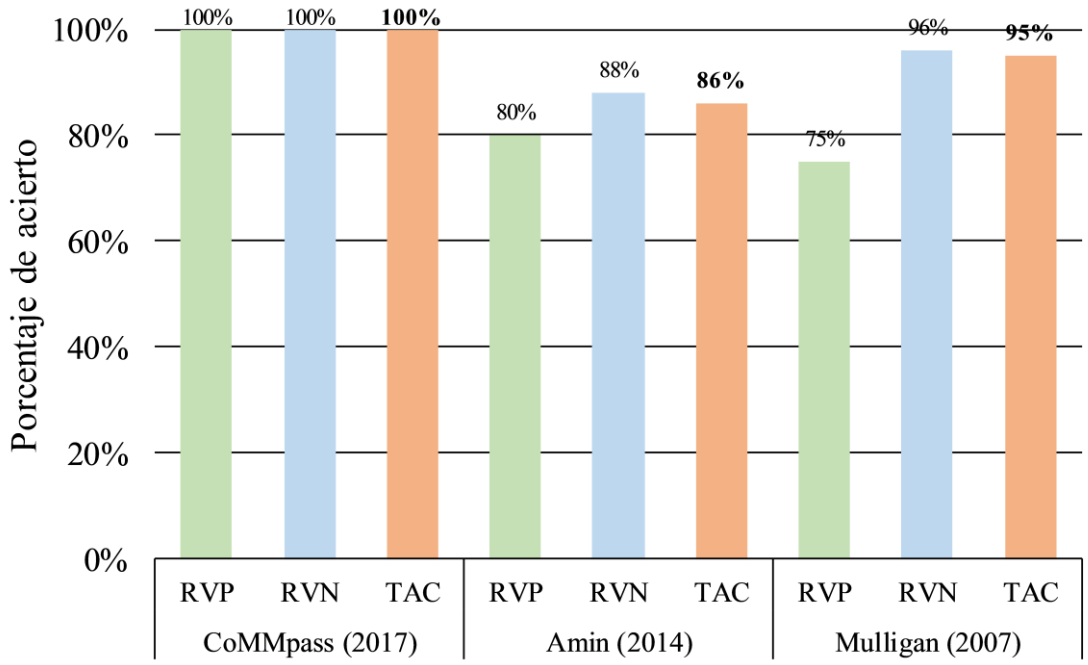



***Figura 4.157.*** *Resultados de la predicción óptima de la respuesta completa (RC) a bortezomib en monoterapia. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

Las TAC de la predicción de la RC para todos los estudios han sido elevadas, siendo en todos los casos superiores al 85% de acierto. Además, en los tres estudios, tanto la RVP, como la RVN han mostrado valores superiores al 75%, indicando esto un buen comportamiento de todos los modelos de predicción sobre la predicción tanto de la RC como de las respuestas de los pacientes que no alcanzan esta RC, respectivamente.

***Respuesta codificada en tres subgrupos***

La predicción de la respuesta en tres subgrupos se realizó sobre los estudios de Amin (2014) y Mulligan (2007), ya que fueron los únicos con muestras suficientes para formar los tres subgrupos de respuesta. La estratificación de la respuesta en estos tres subgrupos se llevó a cabo de la siguiente manera: el primer subgrupo (G1) recogió los pacientes que alcanzaron RC, el segundo subgrupo (G2), fueron agrupados los pacientes que,

respondiendo, no alcanzaron RC. Por último, en el tercer subgrupo (G3) fueron incluidos los pacientes que no alcanzaron ningún tipo de respuesta. El número de muestras seleccionadas para cada uno de estos subgrupos en los dos estudios para las matrices de entrenamiento y validación se recoge en la **Tabla 4.25**.

***Tabla 4.25.*** *Número de muestras en los tres subgrupos de pacientes para los estudios seleccionados en la predicción de la respuesta en pacientes tratados con bortezomib en monoterapia.*

| Subgrupo de respuesta | Matriz | Amin (2014) | Mulligan (2007) |
|---|---|---|---|
| G1 | Entrenamiento | n = 11 | n = 9 |
| | Validación | n = 5 | n = 4 |
| G2 | Entrenamiento | n = 19 | n = 48 |
| | Validación | n = 10 | n = 24 |
| G3 | Entrenamiento | n = 15 | n = 56 |
| | Validación | n = 7 | n = 28 |

*En cada subgrupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Como paso previo a los análisis predictivos se procedió a la selección de una lista común de genes a los dos estudios seleccionados para esta aproximación. Para ello, se realizó un análisis de expresión diferencial (ED) con el algoritmo *limma* para tres grupos. En el estudio de Amin (2014) se detectaron 991 genes con diferencias de expresión estadísticamente significativas a *p*-valor < 0,05. Por su parte, en el estudio de Mulligan (2007), el número de genes estadísticamente significativos fue de 3.778, de los que 3.106 genes no estuvieron duplicados. Hay que aclarar que se ha utilizado el *p*-valor en lugar del FDR debido a que este es un paso de selección de variables, serán los métodos predictivos los que determinen la influencia de estas variables seleccionadas sobre la predicción de la respuesta al tratamiento. Finalmente, se procedió al cruce de las dos listas de genes, siendo 215 los genes desregulados de manera común a los dos estudios que constituyeron la lista "inicial" de genes. Posteriormente, se generaron otras dos listas de genes a partir de esta lista "inicial". Una de ellas exclusiva de cada estudio aplicando un filtrado de los genes mediante el algoritmo *Boruta*, y la otra lista sumando los genes de las dos listas generadas por *Boruta* y eliminando los elementos duplicados (lista "sumatorio"). Con estas tres listas se procedió al análisis predictivo

Con respecto al estudio de Amin (2014), el mejor modelo predictivo fue el PLS con dos factores y los 215 genes seleccionados inicialmente, obteniendo una TAC del 77%, siendo las tasas de acierto por subgrupos de respuesta del 80%, 90% y 57% para los subgrupos G1, G2 y G3, respectivamente. Aunque las tasas de acierto de los subgrupos G1 y G2 son altas, la del subgrupo G3 es relativamente baja, ya que solamente alcanzó un 57% de acierto. La causa de esta mala predicción fue debida a que, de los 7 pacientes sujetos a validación, solamente cuatro fueron predichos de manera satisfactoria, mientras

que tres pacientes, una EE y dos EP, fueron clasificados dentro del subgrupo G2. La tabla de contingencia correspondiente al modelo de predicción óptimo se recoge en la **Tabla 4.26**.

***Tabla 4.26.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de Amin (2014) que obtuvo mejores tasas de acierto (PLS, dos factores).*

| PLS, NF = 2, 215 genes | | Subgrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **4** | 0 | 0 |
| | G2 | 1 (1 RC-IF) | **9** | 3 (1 EE, 2 EP) |
| | G3 | 0 | 1 (1 MBRP) | **4** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzaron dichos pacientes, entre paréntesis.*

El estudio de Mulligan, por su parte, obtendría resultados similares en cuanto a la TAC, ya que alcanzó un 75% de acierto con el método PLS con dos factores, mediante el uso de 283 genes, 215 genes si consideramos los genes no duplicados. Sin embargo, la predicción del G1 en este estudio trajo un 0% de aciertos, con lo que los pacientes que alcanzaron RC no pudieron ser discriminados del resto Las tasas de acierto sobre los subgrupos G2 y G3 fueron del 75% y del 86%, respectivamente. La predicción sobre el subgrupo G1 en este caso hace que el modelo propuesto no sea óptimo, ya que dos de los pacientes de este subgrupo son clasificados como G2 y otros dos como G3 (**Tabla 4.27**). Una de las causas de esta falta de poder predictivo sobre este subgrupo puede ser el bajo número de muestras y la ausencia de balance este número respecto a los subgrupos G2 y G3 lo que hace que G1 esté infrarrepresentado en el modelo predictivo.

***Tabla 4.27.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de Mulligan (2007) que obtuvo mejores tasas de acierto (PLS, dos factores).*

| PLS, NF = 2, 215 genes* | | Subrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **0** | 0 | 0 |
| | G2 | 2 (2 RC) | **18** | 4 (2 EE, 2 EP) |
| | G3 | 2 (2 RC) | 6 (3 RP, 3 RM) | **24** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzaron dichos pacientes, entre paréntesis.*

Los resultados completos de las predicciones de los dos estudios pueden ser consultados en el **Anexo 43**, mientras que los resultados de la predicción óptima que alcanzó una mejor TAC para los dos estudios se muestran para su comparación en la **Figura 4.158**.

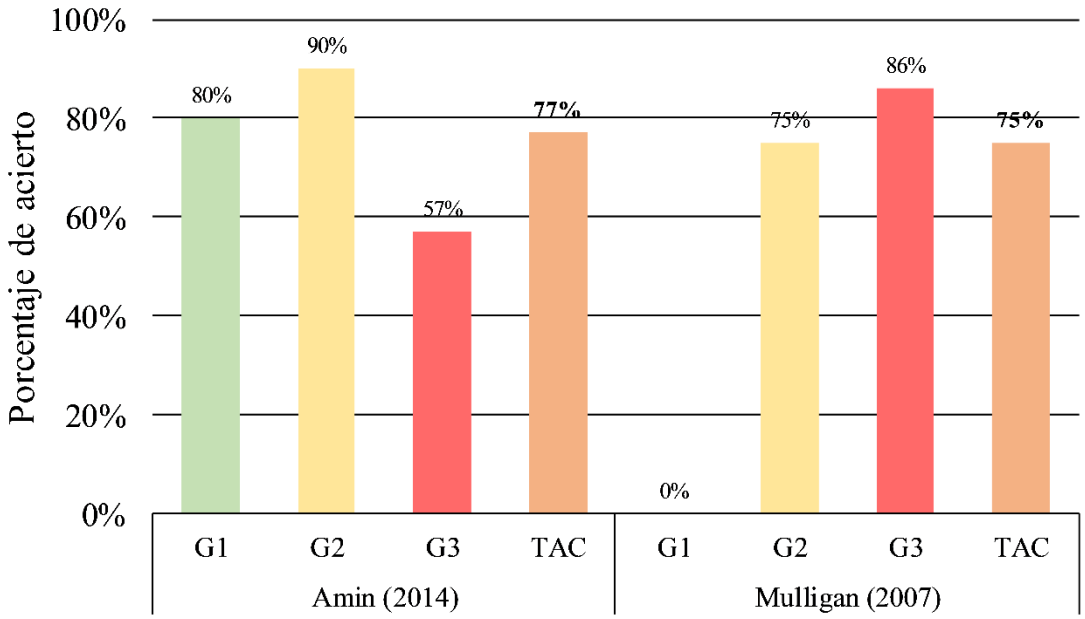



***Figura 4.158.*** *Resultados de la predicción óptima considerando la respuesta a bortezomib en monoterapia en tres subgrupos. El diagrama de barras recoge los valores de la tasa de acierto para los subgrupos G1, G2 y G3, y la tasa de acierto global (TAC) de los dos estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

A pesar del buen resultado obtenido en el estudio de Amin (2014) y que dos de los subgrupos del estudio de Mulligan (2007) obtuvieron buenos porcentajes de TAC, no se puede considerar esta como una buena aproximación, ya que en este segundo estudio no fue posible predecir de forma correcta la respuesta de los pacientes encuadrados en el subgrupo G1. Se recomienda que en futuros trabajos se amplíe el número de estudios recogidos en esta aproximación para comprobar si el estudio de Mulligan (2007) se trata de un caso singular, o si, por el contrario, ciertamente no es posible la predicción del subgrupo G1.

***Análisis predictivo de respuestas múltiples***

Este análisis predictivo se realizó sin llevar a cabo ningún tipo de agrupamiento de los niveles de respuesta a tratamiento. Los estudios seleccionados para en este análisis fueron los estudios de Amin (2014) y Mulligan (2007). En este caso, tanto la selección inicial de genes como las predicciones serán llevadas a cabo de manera independiente sobre cada uno de los dos estudios seleccionados, ya que los grupos de respuesta no son directamente comparables.

***a) Estudio de Amin (2014)***

El estudio de Amin (2014) recoge 67 pacientes de MM agrupados en 7 grupos en función de su respuesta a bortezomib en monoterapia como se recoge en la **Tabla 4.28**.

***Tabla 4.28.*** *Estratificación de las muestras en el estudio de Amin (2014) de pacientes tratados con bortezomib en monoterapia. Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| Matriz | RC-IF | oRC | cRC | MBRP | RM | EE | EP |
|---|---|---|---|---|---|---|---|
| Entrenamiento | n = 7 | n = 4 | n = 5 | n = 10 | n = 4 | n = 7 | n = 8 |
| Validación | n = 3 | n = 2 | n = 3 | n = 5 | n = 2 | n = 3 | n = 4 |

*RC-IF: respuesta completa inmunofenotipica, oRC: otras respuestas completas, cRC: respuesta cercana a la respuesta completa, MBRP: muy buena respuesta parcial, RM: respuesta mínima, EE: enfermedad estable, EP: enfermedad progresiva.*

A continuación, se procedió a la determinación de la lista de genes inicial para realizar la predicción. Para ello se llevó a cabo un análisis con el algoritmo *limma* en su versión para múltiples clases con las muestras de la matriz de entrenamiento. Como resultado se obtuvieron 632 genes a *p*-valor < 0,05 que constituyeron la que fue la lista "inicial" de genes. Se generó una segunda lista a partir de la lista inicial mediante el filtrado con el algoritmo *Boruta* consistente en 15 genes. Con ambas listas se procedió al ajuste de una batería de cinco modelos de predicción, obteniendo que el modelo con mejor tasa de acierto global de predicción fue el PLS con cinco factores y 632 genes, alcanzando una TAC del 73%. Esta tasa global de acierto al ser desgranada por grupo de respuesta desveló un 100% de acierto sobre los grupos RM, EE y EP. Sin embargo, las respuestas RC-IF y cRC solamente alcanzaron un 33% de acierto. De los tres pacientes seleccionados para la validación en el grupo RC-IF, solamente uno fue clasificado correctamente, mientras que dos fueron clasificados como oRC y como MBRP. En el caso del grupo cRC, también fueron tres los pacientes seleccionados para formar parte de la serie de validación, y del mismo modo solamente uno fue correctamente clasificado. Los dos pacientes restantes fueron clasificados como RM. Los resultados al detalle de este modelo óptimo de predicción se recogen en la **Tabla 4.29** y los resultados completos para todos los métodos empleados pueden consultarse en el **Anexo 23.**

***Tabla 4.29.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio de Amin (2014) que obtuvo mejores tasas de acierto (PLS, cinco factores [NF]).*

| **PLS, NF = 5, 632 genes** | | Grupo real | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | RC-IF | oRC | cRC | MBRP | RM | EE | EP |
| Predicción | RC-IF | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| | oRC | 1 | **1** | 0 | 0 | 0 | 0 | 0 |
| | cRC | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| | MBRP | 1 | 0 | 0 | **4** | 0 | 0 | 0 |
| | RM | 0 | 0 | 2 | 0 | **2** | 0 | 0 |
| | EE | 0 | 0 | 0 | 1 | 0 | **3** | 0 |
| | EP | 0 | 1 | 0 | 0 | 0 | 0 | **4** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

### *b) Estudio de Mulligan (2007)*

El estudio de Mulligan (2007) por su parte, recoge los datos de expresión génica de 169 pacientes que se agrupan en función de su respuesta como se recoge en la **Tabla 4.30**.

***Tabla 4.30.*** *Estratificación de las muestras en el estudio de Mulligan (2007) de pacientes tratados con bortezomib en monoterapia. Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| **Matriz** | **RC** | **RP** | **RM** | **EE** | **EP** |
|---|---|---|---|---|---|
| Entrenamiento | n = 9 | n = 40 | n = 8 | n = 29 | n = 27 |
| Validación | n = 4 | n = 20 | n = 4 | n = 14 | n = 14 |

*RC: respuesta completa, RP: respuesta parcial, RM: respuesta mínima, EE: enfermedad estable, EP: enfermedad progresiva.*

El procedimiento para el cálculo de la lista "inicial" de genes se llevó a cabo de manera similar al estudio de Amin (2014). En este caso, la lista "inicial" contó con 2.663 genes, de los cuales 2.238 no estaban duplicados, mientras que la lista "boruta" recogió 38 genes únicos. De los 10 modelos predictivos ajustados, el que obtuvo una mejor TAC fue el PLS con 11 factores y 2.663 genes. La tasa de acierto de este modelo fue del 63%, siendo los grupos de respuesta mejor caracterizados, los grupos EE y EP con un 71% de tasa de acierto. La matriz de contingencia asociada a esta predicción se recoge en la **Tabla 4.31**, donde se muestra la clasificación de todos los pacientes de la matriz de validación por el modelo predictivo óptimo. Así, puede observarse que los pacientes mal clasificados del grupo RC, cuya tasa de acierto es del 50%, son clasificados como RP. En el caso del grupo RM, también con el 50% de tasa de acierto, los dos pacientes incorrectamente clasificados caen en el grupo de EE.

***Tabla 4.31.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio de Mulligan (2007) que obtuvo mejores tasas de acierto (PLS, 11 factores [NF]).*

| **PLS, NF = 11, 2663 genes** | | Grupo real | | | | |
|---|---|---|---|---|---|---|
| | | RC | RP | RM | EE | EP |
| Predicción | RC | **2** | 0 | 0 | 0 | 0 |
| | RP | 2 | **11** | 0 | 2 | 3 |
| | RM | 0 | 0 | **2** | 0 | 0 |
| | EE | 0 | 6 | 2 | **10** | 1 |
| | EP | 0 | 3 | 0 | 2 | **10** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

El resultado de los 10 modelos predictivos ajustados para la serie de Mulligan (2004) puede ser consultado en el **Anexo 43**.

## 4.5.2. Terapias basadas en el uso de bortezomib

El análisis de predicción sobre terapias basadas en el uso de bortezomib se llevó a cabo sobre los estudios de expresión génica para los que se dispuso de datos de respuesta al tratamiento con bortezomib en monoterapia o en combinación con otros compuestos, siempre que no fuesen IMiDs. En cada uno de los cuatro apartados recogidos a continuación, se detallarán los estudios utilizados en las respectivas aproximaciones analíticas.

***Pacientes respondedores versus no respondedores***

La predicción de la respuesta a terapias basadas en el uso de bortezomib se realizó sobre los estudios CoMMpass (2017), Amin (2014) y Mulligan (2007). En la **Tabla 4.32**, se recoge el número de muestras que presentó cada uno de estos tres estudios. En cuanto a las ratios OR/NR, el estudio CoMMpass (2017) presentó una ratio de 10,29, en el estudio de Amin (2014) la ratio fue de 1,39 y en el estudio de Mulligan (2007) la ratio fue de 0,76 (~1,32). Esto implicó que el estudio más descompensado entre los dos grupos de análisis fue el estudio CoMMpass (2017), mientras que los otros dos estudios mostraron unas ratios relativamente compensadas.

***Tabla 4.32.*** *Número de muestras en los grupos de pacientes respondedores y no respondedores en los estudios seleccionados para la predicción de la respuesta en pacientes tratados con bortezomib en monoterapia.*

| Tipo de respuesta | Matriz | CoMMpass (2017) | Amin (2014) | Mulligan (2007) |
|---|---|---|---|---|
| Responden | Entrenamiento | n = 96 | n = 26 | n = 49 |
| | Validación | n = 48 | n = 13 | n = 24 |
| No responden | Entrenamiento | n = 9 | n = 19 | n = 64 |
| | Validación | n = 5 | n = 9 | n = 32 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Para llevar a cabo los análisis de predicción sobre estos estudios se seleccionó como lista "inicial" de genes los 143 genes estadísticamente significativos a *p*-valor < 0,05 del metaanálisis del **Apartado 4.4.2.1**. De manera adicional, se generaron otras dos listas de genes sobre las que ejecutar los análisis predictivos. Para construir la que sería la segunda lista, se utilizó el algoritmo de filtrado de variables *Boruta*. Esta tercera lista al ser exclusiva de cada uno de los tres estudios tuvo dos genes en el estudio CoMMpass (2017), 13 genes en el estudio de Amin (2014) y 18 genes en el estudio de Mulligan (2007). La tercera lista fue común a todos los estudios y se construyó realizando la suma de los genes de las tres listas generadas a partir de *Boruta* eliminando los duplicados. Esta lista "sumatorio" tuvo un total de 31 genes. Con estas tres listas se procedió a los análisis de predicción.

En el estudio CoMMpass (2017) la mejor predicción se llevó a cabo con el método PLS con cuatro factores utilizando los 143 genes de la lista "inicial". La TAC que logró esta predicción fue del 92%, con una RVP del 100% y una RVN del 20%, lo que indica que el modelo no funcionó bien prediciendo la ausencia de respuesta. En cuanto a los valores de VPP y VPN, se situaron en el 92% y en el 100%, respectivamente. Esto significa que en el 92% de las veces que el modelo prediga que un paciente va a responder, el paciente realmente responderá, y que, si el modelo indica que un paciente no va a responder, el 100% de las veces no responderá.

En lo que se refiere al estudio de Amin (2014), la predicción óptima se consiguió con el método SVM utilizando un kernel sigmoidal a un coste de 0,1 con un factor gamma = 1. Para llevar a cabo esta predicción se utilizó la lista de 13 genes de Boruta, obteniendo una TAC del 91%, con unos valores de RVP y RVN del 92% y del 89%, siendo consideradas como elevadas estas tasas de acierto tanto para los pacientes respondedores como para los no respondedores. En la misma línea se situaron los valores de VPP y VPN, alcanzando un 92% y un 89%, respectivamente.

En último lugar, el estudio de Mulligan (2007) logró su predicción óptima utilizando el método PLS con tres factores y los 143 genes de la lista "inicial". La TAC para este estudio fue ligeramente inferior a los estudios anteriores, alcanzando un 73% de acierto.

El valor del RVP fue del 79% y el del RVN del 69%, situándose también en una cota inferior al estudio de Amin (2014). En cuanto al VPP y al VPN, los valores obtenidos fueron del 66% y del 81%, respectivamente.

Los resultados de la predicción óptima que alcanzó una mejor TAC para los tres estudios se muestran para su comparación en la **Figura 4.159**, mientras que los resultados al completo considerando todos los métodos de predicción y listas de genes pueden ser consultados en el **Anexo 43**.

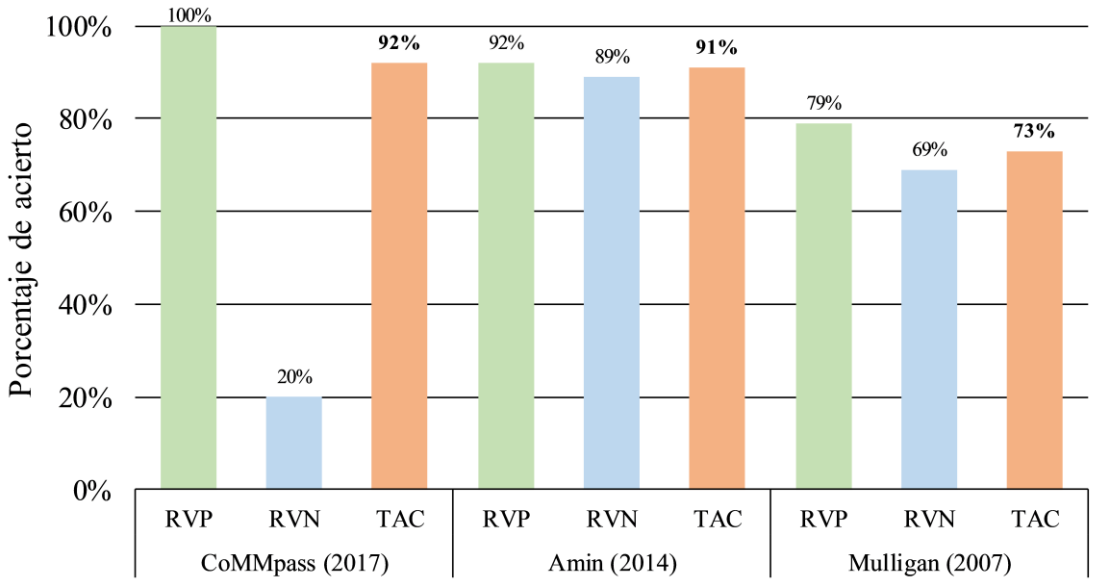



***Figura 4.159.*** *Resultados de la predicción óptima de la respuesta a regímenes de tratamiento basados en bortezomib. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

Los tres estudios cuentan con unas TAC aceptables, siendo el grupo de pacientes respondedores el que mejores tasas de acierto alcanza en los tres casos. El grupo de pacientes no respondedores también es predicho de manera satisfactoria en los estudios de Amin (2014) y Mulligan (2007), sin embargo, en el estudio CoMMpass solamente alcanza un 20% de aciertos. Una de las causas de esta baja tasa de acierto en este grupo podría ser la fuerte descompensación de los dos grupos de respuesta, provocando que ninguno de los métodos de predicción logre capturar las características de este grupo de pacientes. Otra posible causa es la alta heterogeneidad de los tratamientos aplicados a los pacientes de este estudio, ya que, mientras en los estudios de Amin (2014) y Mulligan (2007) todos los pacientes están sometidos al mismo régimen de tratamiento, en el estudio CoMMpass (2017) encontramos pacientes sometidos a regímenes dispares, como bortezomib en monoterapia, bortezomib en terapia combinada con dexametasona, e incluso bortezomib combinado con dexametasona y otros compuestos como

daratumumab, melfalán o ciclofosfamida, aunque todos ellos con el denominador común del bortezomib.

***Pacientes que alcanzan respuesta completa versus resto***

Fueron seleccionados cinco estudios para la predicción de la RC en regímenes de tratamiento basados en bortezomib: CoMMpass (2017), Amin (2014), López-Corral (2014), Gutiérrez (2010) y Mulligan (2007). El tamaño de estos cinco estudios, expresado como el número de pacientes por grupo de respuesta, aparece recogido en la **Tabla 4.33**. En cuanto a las ratios RC/resto de respuestas se detectó una alta heterogeneidad entre los cinco estudios. De esta manera, encontramos que los estudios CoMMpass (2017) y Mulligan (2007) presentaron unas ratios muy elevadas de 0,12 (~8,33) y 0,08 (~12,05), el estudio de Amin (2014) presentó una ratio intermedia de 0,31 (~3,19), y los estudios de López-Corral (2014) y Gutiérrez (2010) fueron los que presentaron una mayor homogeneidad en el tamaño de los dos grupos, con ratios de 1,4 y 1,17, respectivamente.

***Tabla 4.33.*** *Número de muestras en los grupos de pacientes que alcanzaron respuesta completa (RC) y del resto de pacientes en los estudios seleccionados para la predicción de la RC en pacientes tratados con regímenes basados en bortezomib.*

| Respuesta | Matriz | CoMMpass (2017) | Amin (2014) | López-Corral (2014) | Gutiérrez (2010) | Mulligan (2007) |
|---|---|---|---|---|---|---|
| RC | Entrenamiento | n = 11 | n = 11 | n = 5 | n = 5 | n = 9 |
| | Validación | n = 6 | n = 5 | n = 2 | n = 2 | n = 4 |
| Resto | Entrenamiento | n = 94 | n = 34 | n = 3 | n = 4 | n = 101 |
| | Validación | n = 47 | n = 17 | n = 2 | n = 2 | n = 55 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Para realizar los análisis de predicción se seleccionaron los 76 genes estadísticamente significativos a *p*-valor < 0,05 del metaanálisis del **Apartado 4.4.2.2**. A partir de esta lista "inicial" de genes se generaron otras dos listas adicionales. La primera de estas listas fue construida mediante el filtrado de los genes de la lista "inicial" utilizando el algoritmo *Boruta*. Esta lista fue exclusiva para cada uno de los cuatro estudios en los que fue posible aplicar el algoritmo, de manera que el tamaño de esta lista fue de tres genes en el estudio CoMMpass (2017), 13 genes en el estudio de Amin (2014), 8 genes en el estudio de Mulligan (2007) y tres genes en el estudio de Gutiérrez (2010). No fue posible aplicar el filtrado sobre el estudio de López-Corral (2014) debido a su bajo tamaño muestral. La segunda de las listas adicionales fue establecida mediante la suma de los genes de las cuatro listas de *Boruta* eliminando los elementos duplicados. Esta lista constó de 25 genes y se utilizó de manera común para las predicciones en los cinco estudios seleccionados.

El primero de los estudios analizados fue el CoMMpass (2017). La mejor predicción para este estudio alcanzó una TAC del 92%, con una RVP del 33% y una RVN del 100%, lo que indica que mientras que la ausencia de RC se predice de manera correcta en todos los casos, la RC no puede ser predicha de una forma adecuada. Sin embargo, el VPP y el VPN alcanzaron unos valores de 100% y 92 %. Esto implica que, aunque el modelo no prediga la RC de forma correcta, si este indica que un paciente va a hacer RC, el paciente la alcanzará realmente el 100% de las veces, mientras que, si el modelo indica que el paciente no alcanzará RC, el 92% de las veces el modelo acertará. Este modelo fue ajustado utilizando el método PLS con tres factores y los 76 genes de la lista "inicial".

En cuanto al estudio de Amin (2014), la mejor predicción se logró mediante el método PLS con tres factores y los 25 genes de la lista "sumatorio". Esta predicción alcanzó una TAC del 86%, con una RVP del 60% y una RVN del 94%, con lo que la predicción tanto de pacientes que alcanzan RC como del resto de respuestas estuvo más compensada que en el estudio CoMMpass (2017). En cuanto a los valores de VPP y VPN, el modelo consiguió unos valores de 75% y 89%, respectivamente.

La mejor predicción para el estudio de Mulligan (2007) también se llevó a cabo utilizando el método PLS, en este caso con dos factores y utilizando los 76 genes de la lista "inicial". Esta predicción alcanzó una TAC del 97%, sin embargo, solo fue capaz de predecir correctamente un 50% de los pacientes que alcanzaron RC, por lo que esta alta TAC se consiguió al predecir correctamente la respuesta del 100% de pacientes no respondedores. El VPP para este estudio se situó en el 100%, mientras que el VPN obtuvo un 97%, por tanto, si el modelo indica que el paciente alcanza RC o que no lo alcanza, e habrá un 100% o un 97% de posibilidades, respectivamente, de que así ocurra realmente.

El estudio de López-Corral (2014) alcanzó su mejor TAC indistintamente del método predictivo utilizado con los 76 genes de la lista "inicial". La TAC alcanzado en todos los casos fue del 100%, por lo que también los valores de RVP y RVN fueron del 100%.

De manera similar, el estudio de Gutiérrez (2010) también alcanzo un 100% de TAC con los 76 genes de la lista "inicial", sin embargo, esta no fue la predicción óptima al alcanzar la predicción con la lista "boruta", con un menor número de genes, llevada a cabo mediante PLS con un factor, el mismo valor de TAC (100%).

El resumen del análisis de predicción para la RC se recoge en la **Figura 4.160**, mientras que los resultados detallados de las predicciones para cada uno de los estudios pueden ser consultados en el **Anexo 43.**

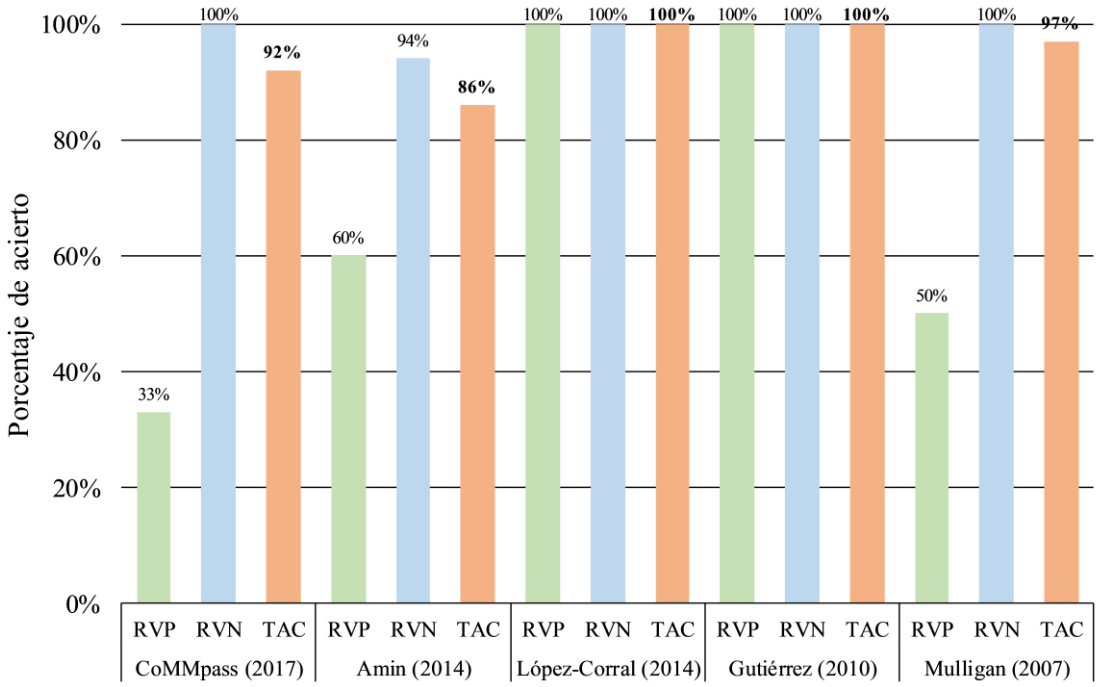



***Figura 4.160.*** *Resultados de la predicción óptima de la respuesta completa (RC) a regímenes de tratamiento basados en bortezomib. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

Los resultados de este análisis de predicción han supuesto para todos los estudios valores altos de TAC, mayores al 85% en todos los casos. Esto se une a que la predicción del grupo de no respondedores ha sido siempre superior al 90% y a que en la mayor parte de los estudios se predice correctamente la respuesta de los pacientes que alcanzan RC. Por tanto, estos resultados indican que es factible la predicción de estos grupos de respuesta a partir de las listas de genes propuestas. Sin embargo, hay que advertir que habría que trabajar con estudios en los que el tamaño muestral de los dos grupos de respuesta sea lo más homogéneo posible, ya que puede observarse un mayor error de la predicción de la RC a medida que la ratio entre ambos grupos está más descompensada.

***Respuesta codificada en tres subgrupos***

Para este análisis de predicción fueron seleccionados los tres estudios que presentaron un tamaño muestral suficiente para la estratificación en tres subgrupos de respuesta: CoMMpass (2017), Amin (2014) y Mulligan (2007). La conformación de los tres subgrupos de respuesta se detalla en el **Apartado 4.5.1**. El número de muestras en cada uno de los subgrupos para cada estudio, detallado en función de la selección para las matrices de entrenamiento y validación, aparece recogido en la **Tabla 4.34**.

***Tabla 4.34.*** *Número de muestras en los tres subgrupos de pacientes para los estudios seleccionados en la predicción de la respuesta en pacientes con regímenes de tratamiento basados en bortezomib.*

| Respuesta | Matriz | CoMMpass (2017) | Amin (2014) | Mulligan (2007) |
|---|---|---|---|---|
| G1 | Entrenamiento | n = 11 | n = 11 | n = 9 |
| | Validación | n = 6 | n = 5 | n = 4 |
| G2 | Entrenamiento | n = 85 | n = 19 | n = 48 |
| | Validación | n = 42 | n = 10 | n = 24 |
| G3 | Entrenamiento | n = 9 | n = 15 | n = 56 |
| | Validación | n = 5 | n = 7 | n = 28 |

*En cada subgrupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Se procedió en primer lugar a la determinación de los genes necesarios para realizar el análisis predictivo. Para ello, se llevó a cabo un análisis de ED para tres grupos en cada uno de los tres estudios. Este análisis resultó en la detección de 1.724 genes en el estudio CoMMpass (2017), 950 genes en el estudio de Amin (2014) y 3.012 genes en el estudio de Mulligan (2007), en todos los casos el *p*-valor de estos genes fue inferior a 0,05. A continuación se seleccionaron los genes comunes a un mínimo de dos de los estudios para generar la lista "inicial" de 600 genes con la que se ejecutaron los análisis predictivos. Además, se construyeron dos listas adicionales de genes. La primera de estas listas fue exclusiva para cada uno de los tres estudios, siendo construida mediante el filtrado de los genes iniciales con el algoritmo *Boruta*. Este proceso resultó en la selección de 11 genes en el estudio CoMMpass (2017), 24 genes en el estudio de Amin (2014) y 28 genes en el estudio de Mulligan (2007). Por último, se llevó a cabo la suma de los genes de las tres listas de *Boruta* para confeccionar la tercera lista "sumatorio", común a todos los estudios. Con las listas "inicial", "boruta" y "sumatorio" se procedió a los análisis predictivos.

En lo que respecta al estudio CoMMpass (2017), la mejor predicción se realizó utilizando el método PLS con 10 factores y 600 genes. Esta predicción alcanzó una TAC del 91%, mientras que las tasas de acierto en los tres subgrupos fueron del 67% para G1, 100% para G2 y 40% para G3. Las dos muestras mal predichas del G1 fueron dos RC que fueron clasificadas en el G2, mientras que las tres muestras en las que el modelo erró del G3, fueron tres EE, clasificadas en el G2. Los resultados completos de esta predicción óptima aparecen recogidos en la **Tabla 4.35**.

***Tabla 4.35.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio CoMMpass (2017) que obtuvo mejores tasas de acierto (PLS, 10 factores [NF] y 600 genes).*

| **PLS, NF = 10, 600 genes** | | Subgrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **3** | 0 | 0 |
| | G2 | 2 (2 RC) | **42** | 3 (3 EE) |
| | G3 | 0 | 0 | **2** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes*

En lo que respecta al estudio de Amin (2014) la mejor predicción se logró con el método PLS utilizando dos factores y los 600 genes de la lista inicial. Esta predicción alcanzó una TAC del 73%, con un 80% de acierto en G1, 60% en G2 y 86% de respuestas predichas correctamente en G3. Los resultados detallados de esta predicción se recogen en la **Tabla 4.36**, donde puede observarse además la respuesta real de los pacientes mal clasificados en los tres subgrupos. El buen balance en el número de muestras en los tres subgrupos es una de las causas que ha podido propiciar las buenas tasas de acierto obtenidas en esta predicción.

***Tabla 4.36.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de Amin (2014) que obtuvo mejores tasas de acierto (PLS, dos factores [NF] y 600 genes).*

| **PLS, NF = 2, 600 genes** | | Grupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **4** | 3 (1 cRC, 2 MBRP) | 0 |
| | G2 | 0 | **6** | 1 (1 EP) |
| | G3 | 1 (1 RC-IF) | 1 (1 RM) | **6** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

El tercero de los estudios fue el de Mulligan (2007). La mejor predicción para este estudio alcanzó una TAC del 71%, con tasas de acierto en la predicción buenas para los subgrupos G2 y G3 superiores al 70% en ambos casos. Sin embargo, la tasa de acierto para el subgrupo G1 fue del 0%, siendo las cuatro RC utilizadas para la validación en este subgrupo predichas como G2 (**Tabla 4.37**). Como se viene indicando, la principal causa que se baraja para explicar este error de predicción es la falta de balance en el tamaño muestral del G1 con respecto a G2 y G3. Así podemos ver que las ratios G2/G1 y G3/G1 fueron de 5,54 y 6,46, respectivamente, mientras que la ratio G3/G2 fue 1,17, con lo que claramente G1 apareció muy infrarrepresentado respecto a los otros dos subgrupos en este estudio.

***Tabla 4.37.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de Mulligan (2007) que obtuvo mejores tasas de acierto (PLS, 8 factores [NF] y 600 genes).*

| **PLS, NF = 8, 600 genes*** | | Subgrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **0** | 0 | 0 |
| | G2 | 4 (4 RC) | **17** | 5 (2 EE, 3 EP) |
| | G3 | 0 | 7 (6 RP, 1 RM) | **23** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes. *Genes no duplicados.*

La estratificación en tres subgrupos de respuesta ha supuesto en los tres estudios analizados unas TAC superiores al 70%. Sin embargo, las tasas de acierto sobre los distintos subgrupos no han sido adecuadas en todos los estudios, barajando como una de las causas de este hecho la falta de balance en el tamaño muestral entre los tres subgrupos. El resumen de los resultados de este análisis predictivo se recoge en la **Figura 4.161**, mientras que los resultados al detalle pueden ser consultados en el **Anexo 43**.

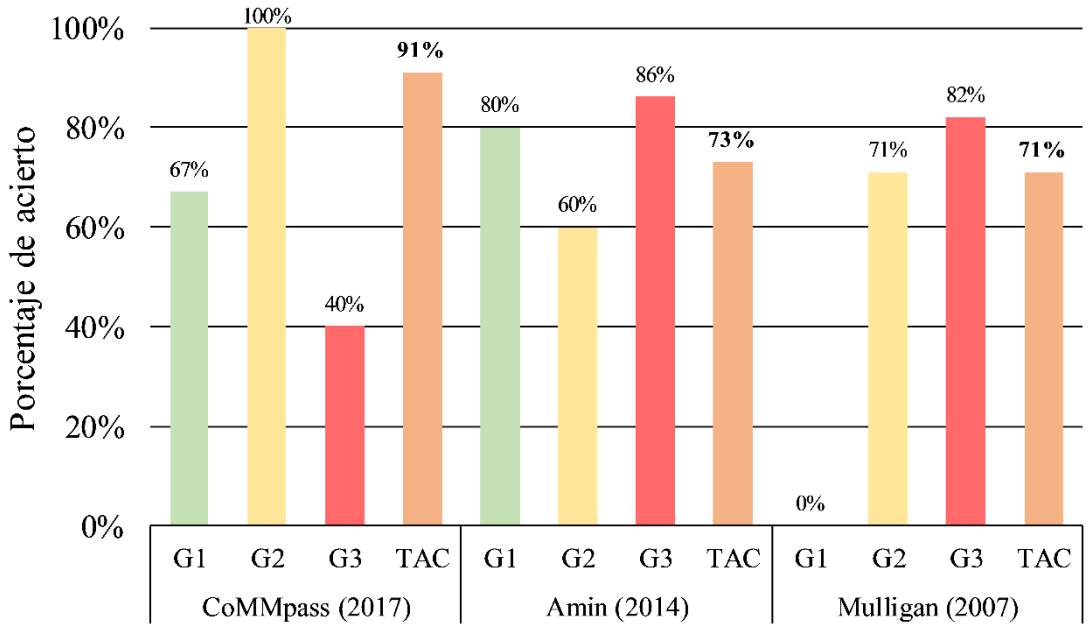



***Figura 4.161.*** *Resultados de la predicción óptima considerando la respuesta a regímenes de tratamiento basados en bortezomib en tres subgrupos. El diagrama de barras recoge los valores de la tasa de acierto para los subgrupos G1, G2 y G3, y la tasa de acierto global (TAC) de los dos estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

***Análisis predictivo de respuestas múltiples***

En este análisis predictivo se consideraron todos los grupos de respuesta disponibles en los tres estudios seleccionados: CoMMpass (2017), Amin (2014) y Mulligan (2007). Tanto la selección de los genes para realizar los análisis predictivos, como la propia predicción, fueron conducidos de forma independiente en cada uno de los estudios, al no ser los grupos de respuesta comparables entre los tres estudios.

***a) Estudio CoMMpass (2017)***

El estudio CoMMpass (2017) consta de 158 pacientes con MM agrupados en cinco grupos de respuesta como se recoge en la **Tabla 4.38**.

***Tabla 4.38****. Estratificación de la respuesta para las muestras del estudio CoMMpass (2017). Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| Matriz | RC | RCs | MBRP | RP | EE |
|---|---|---|---|---|---|
| Entrenamiento | n = 6 | n = 5 | n = 52 | n = 33 | n = 9 |
| Validación | n = 4 | n = 2 | n = 26 | n = 16 | n = 5 |

*RC: respuesta completa, RCs: respuesta completa estricta, MBRP: muy buena respuesta parcial, RP: respuesta parcial, EE: enfermedad estable.*

Una vez definidos los grupos de respuesta se procedió a la determinación de la lista "inicial" de genes, necesaria para realizar el análisis predictivo. Mediante el uso del algoritmo *edgeR* se llevó a cabo un análisis de ED considerando cinco grupos con la matriz de entrenamiento. El resultado fue la obtención de 2.494 genes diferencialmente expresados considerando un $p$-valor < 0,05. Estos 2.494 constituyeron la llamada lista "inicial". A partir de estos genes se generó una segunda lista de 13 genes mediante el filtrado con el algoritmo *Boruta*. Con estas dos listas se procedió al análisis de predicción, de manera que el modelo que obtuvo una mayor TAC (70%) fue el ajustado con el método PLS con 9 factores y utilizando los genes de la lista "inicial". Los grupos de respuesta que alcanzaron mayores tasas de acierto fueron también los más representados en pacientes, MBRP y RP, ambos con un 81% de acierto. Por el contrario, el grupo con menor tasa de acierto fue también el de menor tamaño muestral, RCs, para el que el modelo no consiguió predecir correctamente la respuesta de ninguno de los pacientes, clasificando los dos pacientes con RCs de la matriz de validación como RP. El resultado completo de la predicción con este modelo se recoge en la **Tabla 4.39**, mientras que los resultados completos del análisis de predicción sobre este estudio pueden ser consultados en el **Anexo 43**.

***Tabla 4.39.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio CoMMpass (2017) que obtuvo mejores tasas de acierto (PLS, 9 factores [NF] y 2494 genes).*

| **PLS, NF = 9, 2494 genes** | | Grupo real | | | | |
|---|---|---|---|---|---|---|
| | | RC | RCs | MBRP | RP | EE |
| Predicción | RC | **1** | 0 | 0 | 0 | 0 |
| | RCs | 0 | **0** | 0 | 0 | 0 |
| | MBRP | 2 | 0 | **21** | 3 | 2 |
| | RP | 1 | 2 | 5 | **13** | 1 |
| | EE | 0 | 0 | 0 | 0 | **2** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

### *b) Estudio de Amin (2014)*

En el estudio de Amin (2014) se reunieron los datos de expresión génica y de la respuesta de 67 pacientes con MM, cuya clasificación en función de su respuesta máxima al tratamiento se recoge en la **Tabla 4.40**.

***Tabla 4.40.*** *Estratificación de la respuesta de los pacientes en el estudio de Amin (2014). Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| **Matriz** | **RC-IF** | **oRC** | **cRC** | **MBRP** | **RM** | **EE** | **EP** |
|---|---|---|---|---|---|---|---|
| Entrenamiento | n = 7 | n = 4 | n = 5 | n = 10 | n = 4 | n = 7 | n = 8 |
| Validación | n = 3 | n = 2 | n = 3 | n = 5 | n = 2 | n = 3 | n = 4 |

*RC-IF: respuesta completa inmunofenotipica, oRC: otras respuestas completas, cRC: respuesta cercana a la respuesta completa, MBRP: muy buena respuesta parcial, RM: respuesta mínima, EE: enfermedad estable, EP: enfermedad progresiva.*

Aunque el análisis sobre esta serie fue similar al llevado a cabo en el **Apartado 4.5.1.4.** hay ligeras diferencias en el número de genes considerados de manera previa al análisis, ya que, para este régimen de tratamiento, se consideraron dos nuevos estudios, y la matriz de genes de entrada se generó con los genes comúnmente interrogados en todos los estudios de cada régimen de tratamiento. Así, se procedió con esta matriz a confeccionar la lista de genes inicial para el análisis predictivo. Para ello se llevó a cabo un análisis de ED con el algoritmo *limma* considerando 7 grupos. Con este análisis se determinó la ED de 597 genes considerando un *p*-valor < 0,05. El filtrado de esta lista "inicial" mediante *Boruta* obtuvo la que fue la segunda lista de 23 genes necesaria para el análisis predictivo.

El mejor modelo predictivo para el estudio de Amin (2014) fue el llevado a cabo con el método PLS con seis factores y los 597 genes de la lista "inicial". La TAC en este

caso fue relativamente baja, alcanzando un 55% de acierto. Dos de los grupos de respuesta, cRC y EE, obtuvieron una tasa de acierto del 100%, mientras que para el resto de los grupos la tasa de acierto fue del 50% o menor. En este estudio los grupos de respuesta estuvieron relativamente bien equilibrados en cuanto a tamaño muestral, ya que se recogieron 15 muestras en el grupo mayor (MBRP), y 6 muestras en los grupos de menor tamaño (oRC y RM). En este caso, la baja tasa de acierto podría ser debida al bajo tamaño muestral que tienen los 7 grupos de respuesta. El resultado detallado del modelo con mejor TAC se recoge en la **Tabla 4.41**, mientras que los resultados completos del análisis de predicción pueden ser consultados en el **Anexo 43**.

***Tabla 4.41.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio de Amin (2014) que obtuvo mejores tasas de acierto (PLS, seis factores [NF] y 597 genes).*

| **PLS, NF = 6, 597 genes** | | Grupo real | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | RC-IF | oRC | cRC | MBRP | RM | EE | EP |
| Predicción | RC-IF | **1** | 1 | 0 | 2 | 0 | 0 | 0 |
| | oRC | 0 | **1** | 0 | 0 | 0 | 0 | 1 |
| | cRC | 0 | 0 | **3** | 1 | 0 | 0 | 1 |
| | MBRP | 0 | 0 | 0 | **2** | 0 | 0 | 0 |
| | RM | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| | EE | 2 | 0 | 0 | 0 | 0 | **3** | 1 |
| | EP | 0 | 0 | 0 | 0 | 1 | 0 | **1** |

*La predicción que obtuvo mejores resultados aparece resaltada en negrita.*

### *c) Estudio de Mulligan (2007)*

El estudio de Mulligan (2007) recoge los datos de expresión génica y la respuesta al tratamiento de 169 pacientes con MM. Los grupos de respuesta para estos pacientes se recogen en la **Tabla 4.42**.

***Tabla 4.42.*** *Estratificación de la respuesta de los pacientes del estudio de Mulligan (2007). Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| **Matriz** | **RC** | **RP** | **RM** | **EE** | **EP** |
|---|---|---|---|---|---|
| Entrenamiento | n = 9 | n = 40 | n = 8 | n = 29 | n = 27 |
| Validación | n = 4 | n = 20 | n = 4 | n = 14 | n = 14 |

*RC-IF: respuesta completa inmunofenotipica, RC: respuesta completa, cRC: respuesta cercana a la respuesta completa, MBRP: muy buena respuesta parcial, RM: respuesta mínima, EE: enfermedad estable, EP: enfermedad progresiva.*

Al igual que se indicó en la serie de Amin (2014), los resultados de este análisis podrían ser similares a los recogidos en el **Apartado 4.5.1.4**, no obstante, se procederá

con el análisis predictivo debido a las ligeras diferencias que presentan las matrices de entrada en ambos casos. Con la matriz de entrada se procedió a confeccionar la lista de genes "inicial" para el análisis predictivo. Para ello se realizó un análisis de ED con el algoritmo *limma* para cinco grupos, siendo seleccionados 2.594 genes (2.173 genes sin duplicados) diferencialmente expresados a *p*-valor < 0,05. A partir de estos genes se generó una segunda lista mediante el filtrado con *Boruta*, lo que llevó a la selección de 32 genes. Con estas dos listas se llevaron a cabo los análisis predictivos. El modelo que alcanzó una mejor TAC fue el ajustado mediante el método PLS con 11 factores y 2.173 genes, alcanzando una TAC del 77%. En todos los grupos de respuesta la tasa de acierto fue mayor o igual al 50%, siendo la RP la que obtuvo la menor tasa de acierto (50%) y la EE la mayor tasa de acierto (93%). Los resultados detallados de la predicción utilizando este modelo se recogen en la **Tabla 4.43**, mientras que los resultados completos de todos los análisis predictivos pueden ser consultados en el **Anexo 43**.

***Tabla 4.43.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio de Mulligan (2007) que obtuvo mejores tasas de acierto (PLS, 11 factores [NF] y 2.173 genes sin duplicados).*

| **PLS, NF = 11, 2173 genes*** | | Grupo real | | | | |
|---|---|---|---|---|---|---|
| | | RC | RP | RM | EE | EP |
| Predicción | RC | **3** | 0 | 0 | 0 | 0 |
| | RP | 0 | **15** | 0 | 1 | 3 |
| | RM | 0 | 0 | **2** | 0 | 0 |
| | EE | 0 | 2 | 0 | **13** | 1 |
| | EP | 1 | 3 | 2 | 0 | **10** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes. * Genes no duplicados.*

## 4.5.3. Terapia combinada de bortezomib e IMiDs

Para llevar a cabo este análisis de predicción se seleccionaron los estudios en pacientes con MM cuyo tratamiento consistió en bortezomib en combinación con IMiDs, como la talidomida, la lenalidomida o la pomalidomida. En cada una de las cuatro aproximaciones analísticas realizadas a continuación se detallarán los estudios que han sido seleccionados para realizar el análisis predictivo.

***Pacientes respondedores versus no respondedores***

Para la predicción de la respuesta al tratamiento frente a la no respuesta se dispuso de dos estudios: CoMMpass (2017) y Terragna (2016). El número de muestras recogidas en cada uno de los dos estudios se recoge en la **Tabla 4.44**. En esta aproximación, los dos grupos de respuesta en los dos estudios aparecen fuertemente desequilibrados con unas

ratios OR/NR de 36,67 en el caso de CoMMpass (2017) y 15,86 en el caso de Terragna (2016) debido a que la mayor parte de los pacientes responde a esta combinación de fármacos.

***Tabla 4.44**. Número de muestras en los grupos de pacientes respondedores y no respondedores en los estudios seleccionados para la predicción de la respuesta en pacientes tratados con bortezomib en combinación con IMiDs.*

| Respuesta | Matriz | CoMMpass (2017) | Terragna (2016) |
|---|---|---|---|
| Responden | Entrenamiento | n = 220 | n = 74 |
| | Validación | n = 110 | n = 37 |
| No responden | Entrenamiento | n = 9 | n = 5 |
| | Validación | n = 4 | n = 2 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Los análisis de predicción se realizaron utilizando como lista "inicial" de genes los 44 genes estadísticamente significativos a *p*-valor < 0,05 del metaanálisis del **Apartado 4.4.3.1**. A partir de esta lista "inicial" se generaron dos listas adicionales de genes. La primera de estas listas se construyó mediante el filtrado de la lista "inicial" aplicando el algoritmo *Boruta*. Este procedimiento se realizó de forma individual para los dos estudios obteniendo una lista de 11 genes en el caso de CoMMpass (2017) y 8 genes en el caso del estudio de Terragna (2016). La segunda de las dos listas fue construida sumando los genes de las listas "Boruta" de los dos estudios y eliminando los elementos duplicados. Esta lista fue llamada "sumatorio" y constó de 19 genes. Con las tres listas se procedió al análisis predictivo.

En el caso del estudio CoMMpass (2017) no hubo un método vencedor a la hora de ajustar el modelo predictivo. Tanto wSVM como PLS, SVM, KNN y RF lograron un 96% de TAC. Esto fue debido a que en todos los casos el RVP fue del 100% y, sin embargo, ningún método logró una correcta predicción de la respuesta de las muestras del grupo de no respondedores (RVN = 0%). Como se indicó anteriormente, esta ausencia de acierto de predicción probablemente fue debida a la fuerte descompensación del número de muestras en los dos grupos de respuesta, imposibilitando un correcto funcionamiento de los métodos predictivos. Una segunda causa que se baraja es la heterogeneidad en los tratamientos aplicados en este estudio, ya que no solo se utilizan diferentes IMiDs (lenalidomida o talidomida), sino que además se combinan con otros fármacos como la dexametasona o la ciclofosfamida.

Un caso similar es el estudio de Terragna (2016), ya que, pese a que el mejor modelo logró un 97% de TAC y un 100% de RVP y 50% de RVN, este resultado no puede ser considerado aceptable, ya que la validación en el grupo de pacientes no respondedores solamente se llevó a cabo con dos pacientes frente a los 37 del grupo de respondedores.

El resumen de los resultados para esta aproximación se recoge mediante diagrama de barras en la **Figura 4.162**, mientras que los resultados completos pueden ser consultados a través del **Anexo 43**.

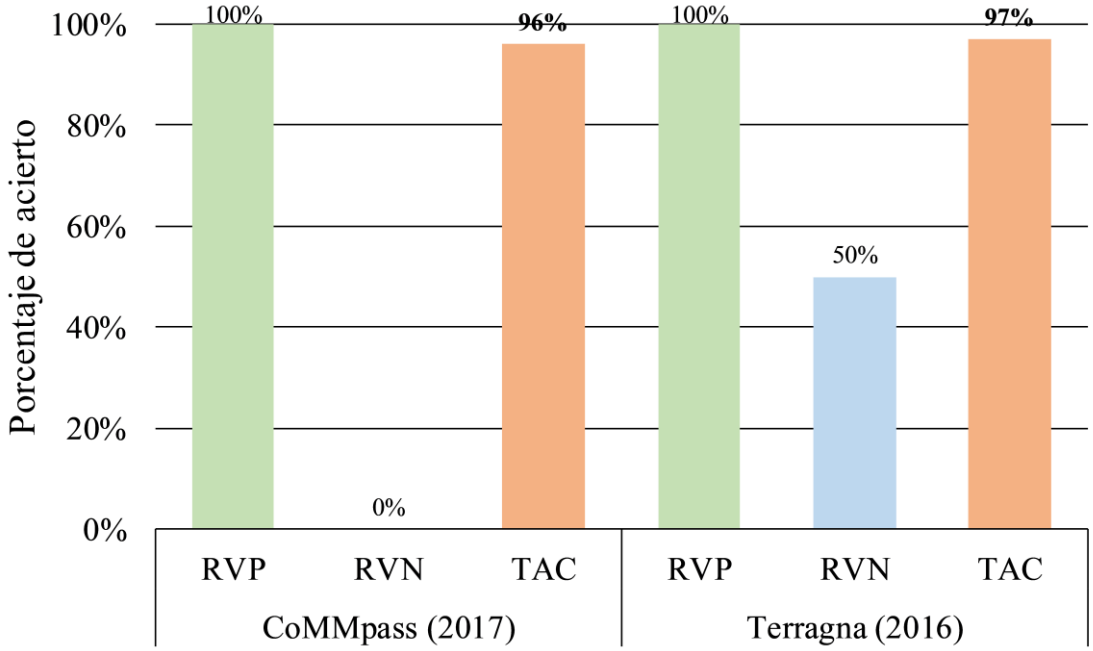



***Figura 4.162.*** *Resultados de la predicción óptima de la respuesta a bortezomib en combinación con IMiDs. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

En esta aproximación, pese a los altos porcentajes que alcanza la TAC en los dos estudios analizados, las predicciones no fueron óptimas en ninguno de los casos, ya que ningún modelo logró predecir de manera satisfactoria la respuesta del grupo de pacientes NR, quizá por las elevadas diferencias en el tamaño muestral entre los dos grupos de respuesta analizados, quizá, en el caso del estudio CoMMpass (2017), por la elevada heterogeneidad en los tratamientos que recibieron los pacientes.

***Pacientes que alcanzan respuesta completa versus resto***

La predicción de la RC en pacientes tratados con bortezomib e IMiDs se realizó sobre los estudios CoMMpass (2017), Terragna (2016), López-Corral (2014) y Gutiérrez (2010). El número de muestras recogido en cada uno de estos estudios estratificadas por el grupo de respuesta, en función de la matriz a la que se asignaron para la predicción se recoge en la **Tabla 4.45**. En lo que respecta a las ratios RC/resto de respuestas en los cuatro estudios, el estudio CoMMpass (2017) presentó una ratio de 0,25 (~3,97), el estudio de Terragna (2016) de 0,15 (~6,87), el estudio de López-Corral (2014) de 0,82 (~1,22) y el estudio de Gutiérrez (2010) de 1,5. Por tanto, los dos primeros estudios

estuvieron ligeramente desequilibrados, en cambio los dos últimos presentaron un buen balance en el tamaño de los dos grupos de respuesta, aunque, en el caso del estudio de Gutiérrez (2010) el tamaño muestral fue muy pequeño.

***Tabla 4.45**. Número de muestras en los grupos de pacientes que alcanzaron respuesta completa (RC) y del resto de pacientes en los estudios seleccionados para la predicción de la RC en pacientes tratados con bortezomib e IMiDs.*

| Respuesta | Matriz | CoMMpass (2017) | Terragna (2016) | López-Corral (2014) | Gutiérrez (2010) |
|---|---|---|---|---|---|
| RC | Entrenamiento | n = 46 | n = 10 | n = 6 | n = 4 |
| | Validación | n = 23 | n = 5 | n = 3 | n = 2 |
| Resto | Entrenamiento | n = 183 | n = 69 | n = 7 | n = 3 |
| | Validación | n = 91 | n = 34 | n = 4 | n = 1 |

*En cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

Como listado "inicial" de genes para este análisis se utilizaron los 54 genes estadísticamente significativos a p-valor < 0,05 del metaanálisis del **Apartado 4.4.3.2**. Con la lista "inicial" como referencia fueron generadas otras dos nuevas listas de genes: la lista "boruta" y la lista "sumatorio". La lista "boruta" fue exclusiva de cada estudio, siendo confeccionada mediante el filtrado de la lista "inicial" con el algoritmo *Boruta*. De este modo, esta lista consistió en tres genes en el estudio CoMMpass (2017), 11 genes en el estudio de Terragna (2016) y seis genes en el estudio de López-Corral (2014). El algoritmo *Boruta* no pudo ser aplicado en el estudio de Gutiérrez (2010) debido a su bajo tamaño muestral. En cuanto a la lista "sumatorio", fue generada sumando los genes de las tres listas "boruta" y eliminando los genes duplicados. Esta lista consistió en 20 genes y fue utilizada para el análisis predictivo en los cuatro estudios seleccionados. Con estas tres listas preparadas, se procedió al análisis de predicción.

La predicción sobre el estudio CoMMpass (2017) logró una TAC del 81% mediante el método PLS con un factor y utilizando los 20 genes de la lista "sumatorio". Pese a que con este modelo predictivo se alcanzó un 100% de RVN, el RVP solamente alcanzó el 4% con lo que la predicción de la RC no fue adecuada.

En el estudio de Terragna (2016) se obtuvieron mejores tasas de acierto a todos los niveles, ya que se alcanzó un 92% de TAC unido a un 40% de RVP y un 100% de RVN. Esta predicción se llevó a cabo utilizando el método PLS con cuatro factores y los 54 genes de la lista "inicial". De manera adicional, este modelo logró un VPP del 100% y un VPN del 92%, lo que indica que, si el modelo predice que un paciente alcanzará RC, el paciente la logrará el 100% de las veces, y si el modelo indica que un paciente no va a alcanzar RC, el paciente no lo hará el 92% de las veces.

En lo que respecta al estudio de López-Corral (2014), se obtuvo una mejor tasa de acierto del 86% mediante el método KNN considerando los 9 vecinos más cercanos y los seis genes de la lista "boruta". El RVP para este modelo fue del 100% mientras que el RVN fue del 75%, fallando en este último caso en la predicción de la respuesta de un único paciente.

Finalmente, el estudio de Gutiérrez (2010) logró su mejor TAC independientemente del método de predicción empleado, considerando en todos los casos la lista "inicial" de 54 genes. Esta mejor predicción alcanzó una TAC del 100%, sin embargo, hay que tratar este alto porcentaje de acierto con cautela debido al bajo tamaño muestral con el que contó este estudio, siendo el grupo de pacientes no respondedores validado con una única muestra.

El resumen de los resultados de la predicción para esta aproximación se recoge en la **Figura 4.163**, mientras que los resultados detallados pueden consultarse en el **Anexo 43**.

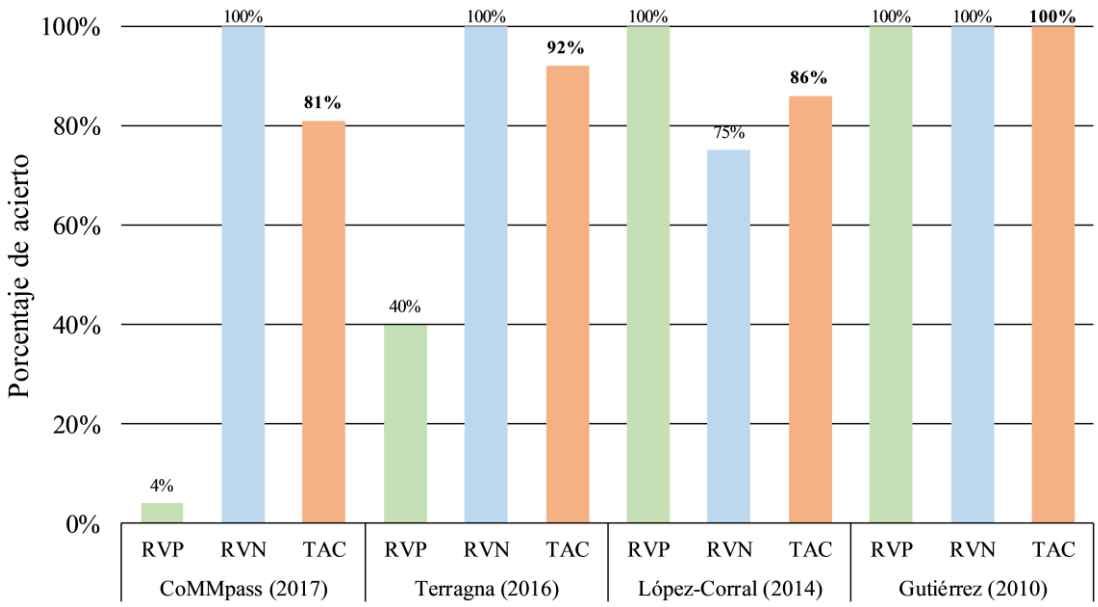



***Figura 4.163.*** *Resultados de la predicción óptima de la respuesta completa (RC) a bortezomib en combinación con IMiDs. El diagrama de barras recoge los valores de la razón de verdaderos positivos (RVP), razón de verdaderos negativos (RVN) y la tasa de acierto global (TAC) de los tres estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

El análisis de predicción de la RC en pacientes tratados con bortezomib e IMiDs mostró resultados dispares. Mientras los estudios de Terragna (2016), López-Corral (2014) y Gutiérrez (2010) mostraron altas TAC y elevados valores de RVN, junto con una buena clasificación también en el RVP, el estudio CoMMpass (2017), por su parte no logró una predicción satisfactoria del grupo de pacientes con RC, lo que pudo ser debido a la elevada heterogeneidad de los tratamientos aplicados en este último estudio,

como ya se viene indicando anteriormente. Esto, unido al bajo tamaño muestral que presentaron estudios como los de López-Corral (2014) y Gutiérrez (2010), con tres y dos pacientes en la serie de validación, respectivamente, hace que los resultados de esta aproximación deban ser tomados con precaución a la hora de sacar conclusiones sobre la capacidad predictora de la expresión génica en la RC del régimen de tratamiento analizado en este apartado.

***Respuesta codificada en tres subgrupos***

El análisis predictivo de la respuesta codificada en tres subgrupos solamente pudo realizarse en los estudios CoMMpass (2017) y Terragna (2016), ya que el resto de los estudios no tuvo un tamaño muestral suficiente para proceder con esta estratificación de la respuesta. Esta estratificación en tres subgrupos ya fue descrita en el **Apartado 4.5.1**, siendo recogido el número de muestras incluidas en cada subgrupo para cada uno de los dos estudios, particularmente para el régimen de tratamiento con bortezomib en combinación con IMiDs, en la **Tabla 4.46.**

***Tabla 4.46.*** *Número de muestras en los tres subgrupos de pacientes para los estudios seleccionados en la predicción de la respuesta en pacientes tratados con bortezomib e IMiDs.*

| Subgrupo de respuesta | Matriz | CoMMpass (2017) | Terragna (2016) |
|---|---|---|---|
| G1 | Entrenamiento | n = 46 | n = 10 |
| | Validación | n = 23 | n = 5 |
| G2 | Entrenamiento | n = 174 | n = 64 |
| | Validación | n = 87 | n = 32 |
| G3 | Entrenamiento | n = 9 | n = 5 |
| | Validación | n = 4 | n = 2 |

*En cada subgrupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación*

En un primer paso, se llevó a cabo un análisis de ED considerando las muestras de las matrices de entrenamiento en cada estudio con el objetivo de generar una lista "inicial" de genes para proceder con el análisis predictivo. Este análisis de ED se realizó con el algoritmo *edgeR* en el estudio CoMMpass (2017) y con *limma* sobre los datos de microarray del estudio de Terragna (2016), en ambos casos utilizando su versión para la comparación de tres grupos. El resultado de este análisis fue la obtención de 2.450 genes diferencialmente expresados en el estudio CoMMpass (2017) y 957 genes diferencialmente expresados en el estudio de Terragna (2016), en ambos casos se eligió como punto de corte para la selección de genes un $p$-valor < 0,05. La lista "inicial" de genes para la predicción fue generada seleccionando los 130 genes comunes a estas dos listas de genes diferencialmente expresados. Adicionalmente fueron construidas dos listas más a partir de esta lista inicial. La primera de las listas se generó mediante el filtrado de la lista "inicial" con el algoritmo *Boruta*, de manera que fueron seleccionados 9 genes en

el estudio CoMMpass (2017) y 14 genes en el estudio de Terragna (2016). Las listas "boruta" fueron exclusivas de cada estudio, por esto, a partir de estas dos listas, se generó una lista de genes común sumando los elementos de ambas listas y eliminando los genes duplicados. Tras la confección de estas tres listas se procedió en un siguiente paso al análisis predictivo.

El análisis de predicción en el estudio CoMMpass (2017) alcanzó una mayor TAC mediante el método PLS con dos factores y los 130 genes de la lista "inicial". Esta TAC fue del 80%, y las tasas de acierto por subgrupo fueron del 22% para G1, 99% para G2 y 0% para G3 (**Figura 4.164**). Por tanto, pese a la elevada TAC, el modelo predictor no mostró un comportamiento adecuado en la predicción de dos de los tres subgrupos, lo cual es reflejo del fuerte desequilibrio en el tamaño muestral de los tres subgrupos: G2/G1 = 3,78 y G2/G3 = 20,08. Estas ratios justificarían el sesgo de la predicción hacia el G2. Además, se comprobó la respuesta de las muestras mal asignadas por si podría justificarse el error de la predicción por la cercanía con la respuesta del grupo mal asignado, no encontrándose ningún tipo de asociación (**Tabla 4.47**).

***Tabla 4.47.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de CoMMpass (2017) que obtuvo mejores tasas de acierto (PLS, dos factores [NF] y 130 genes).*

| **PLS, NF = 2, 130 genes** | | Subgrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **5** | 1 (1 MBRP) | 0 |
| | G2 | 18 (17 RC, 1 RCs) | **86** | 4 (4 EE) |
| | G3 | 0 | 0 | **0** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

En lo que respecta al estudio de Terragna (2016), la mayor TAC (90%) se alcanzó, al igual que en el estudio anterior, con el método PLS y los 130 genes de la lista "inicial". En este caso el modelo requirió de dos factores para llevar a cabo esta predicción. Las tasas de acierto sobre los subgrupos G1, G2 y G3 fueron del 20%, 100% y 100%, respectivamente (**Figura 4.164**). De esta manera, las predicciones sobre G2 y G3 fueron buenas, mientras que G1 no fue predicho de manera satisfactoria. En este estudio también se observó una fuerte falta de equilibrio del subgrupo G2 frente a G1 y G3 (G2/G1 = 6,4 y G2/G3 = 13,71). Como ocurriese en el estudio CoMMpass (2017), las muestras mal clasificadas de G1 tienden a ser clasificadas como G2 (**Tabla 4.48**), sin embargo, pese a que el G3 es el subgrupo que tiene una ratio más desequilibrada respecto al G2, las dos muestras de este subgrupo son clasificadas correctamente, lo que hace pensar en la existencia de otros factores en el estudio CoMMpass (2017) que puedan estar distorsionando el resultado de la predicción, como puede ser la heterogeneidad de los tratamientos aplicados en ese estudio.

***Tabla 4.48.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta en tres subgrupos para el estudio de Terragna (2016) que obtuvo mejores tasas de acierto (PLS, cuatro factores [NF] y 130 genes).*

| **PLS, NF = 4, 130 genes** | | Subgrupo real | | |
|---|---|---|---|---|
| | | G1 | G2 | G3 |
| Predicción | G1 | **1** | 0 | 0 |
| | G2 | 4 (4 RC) | **32** | 0 |
| | G3 | 0 | 0 | **2** |

*En verde se muestran los pacientes a los que se asignó correctamente el subgrupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

Con todos estos resultados, parece que la predicción del régimen de tratamiento con bortezomib en combinación con IMiDs no obtuvo en ninguno de los casos unos resultados óptimos, al no ser predicho de forma correcta ni en el estudio CoMMpas (2017), ni en el estudio Terragna (2018) el subgrupo G1 correspondiente a los pacientes que alcanzaron RC- Adicionalmente, no fue posible la correcta clasificación del subgrupo G3 en el estudio CoMMpass (2017), probablemente, tal y como se viene indicando en este **Apartado 4.5.3**, debido a la elevada heterogenidad que presentó este estudio en cuanto al tratamiento aplicado a los pacientes. Por tanto, para elucidar las posibles causas de las bajas tasas de acierto en los diferentres subgrupos de respuesta se recomienda incrementar el número de estudios en futuros análisis, y, en la medida de lo posible, trabajar con estudios bien homogeneizados en cuanto a los tratamientos utilizados.

Los resultados completos del análisis de predicción para esta aproximación aparecen resumidos en la **Figura 4.164** y adicionalmente pueden ser consultados en el **Anexo 43**.

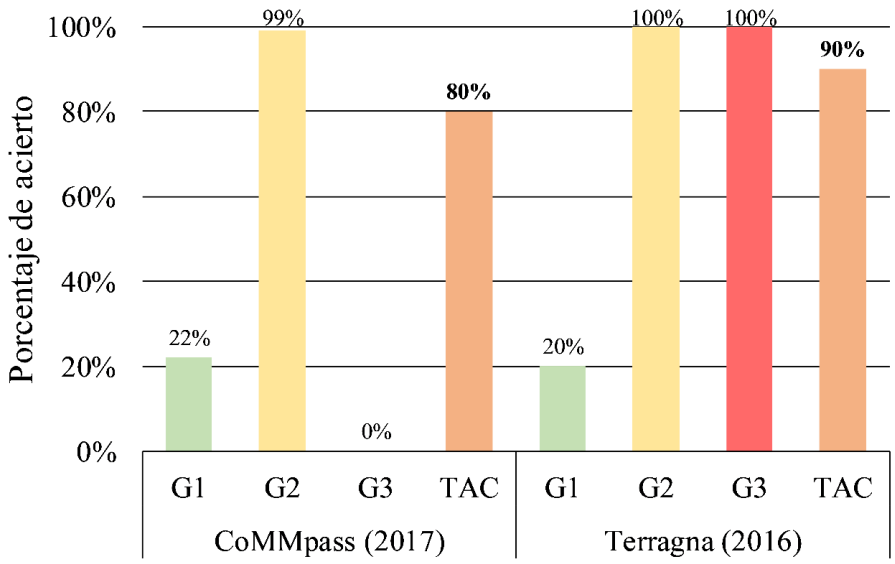



***Figura 4.164.*** *Resultados de la predicción óptima considerando la respuesta a bortezomib en combinación con IMiDs en tres subgrupos. El diagrama de barras recoge los valores de la tasa de acierto para los subgrupos G1, G2 y G3, y la tasa de acierto global (TAC) de los dos estudios seleccionados para este análisis. Los resultados se muestran como porcentaje.*

***Análisis predictivo de respuestas múltiples***

Para la aproximación de la predicción utilizando respuestas múltiples se consideraron todos los grupos de respuesta determinados por los autores de cada uno de los estudios seleccionados. Este análisis se realizó sobre los estudios CoMMPass (2017) y Terragna (2016), ya que fueron los únicos con un tamaño muestral suficiente para realizar la predicción en todos los grupos. Todos los análisis realizados en esta aproximación fueron llevados a cabo de manera independiente sobre cada uno de los dos estudios seleccionados.

***a) Estudio CoMMpass (2017)***

El estudio CoMMpass (2017) recopila datos de expresión génica de 343 pacientes con MM agrupados en cinco grupos de respuesta, tal y como se recoge en la **Tabla 4.49**.

***Tabla 4.49.*** *Estratificación de las muestras en el estudio de CoMMpass (2007) de pacientes tratados con bortezomib en combinación con IMiDs. Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| Matriz | RC | RCs | MBRP | RP | EE |
|---|---|---|---|---|---|
| Entrenamiento | n = 39 | n = 7 | n = 127 | n = 47 | n = 9 |
| Validación | n = 19 | n = 4 | n = 64 | n = 23 | n = 4 |

*RC: respuesta completa, RCs: respuesta completa estricta, MBRP: muy buena respuesta parcial, RP: respuesta parcial, EE: enfermedad estable.*

En un primer paso se procedió a la selección de los genes que conformaron la lista "inicial" para el análisis de predicción. Para ello se llevó a cabo un análisis de ED utilizando las muestras de entrenamiento mediante el algoritmo *edgeR* en su versión de múltiples grupos. Como resultado se obtuvo una lista de 3.930 genes diferencialmente expresados a *p*-valor < 0,05 que constituyeron la lista "inicial". A partir de esta lista, se construyó una segunda lista de 13 genes mediante el filtrado con el algoritmo *Boruta*. Con estas dos listas se realizaron los análisis de predicción.

La TAC en este estudio alcanzó unos valores máximos relativamente bajos, con un 56% de acierto. Esto estuvo condicionado por la falta de acierto (0%) en tres de los grupos de respuesta: RC, RCs y EE. De nuevo, estos resultados aparecen sesgados hacia los grupos de respuesta más representados, como se observa en la **Tabla 4.50**, lo que conduce a considerar como inadecuado este modelo de predicción de múltiples respuestas sobre este estudio.

***Tabla 4.50**. Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio CoMMpass (2017) que obtuvo mejores tasas de acierto (RF, 313 árboles y 3.930 genes).*

| **RF, Nº árboles = 313, 3930 genes** | | Grupo real | | | | |
|---|---|---|---|---|---|---|
| | | RC | RP | MBRP | RP | EE |
| Predicción | RC | **0** | 0 | 0 | 0 | 0 |
| | RCs | 0 | **0** | 0 | 0 | 0 |
| | MBRP | 18 | 4 | **61** | 20 | 4 |
| | RP | 1 | 0 | 3 | **3** | 0 |
| | EE | 0 | 0 | 0 | 0 | **0** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

### *b) Estudio de Terragna (2016)*

El estudio de Terragna (2016) recoge la expresión génica y la respuesta de 118 pacientes con MM, clasificados en cinco grupos de respuesta como se muestra en la **Tabla 4.51**.

***Tabla 4.51.** Estratificación de las muestras en el estudio de Terragna (2016) de pacientes tratados con bortezomib en combinación con IMiDs. Para cada grupo de respuesta, aparece el número de muestras seleccionadas para entrenar el modelo predictivo y para su validación.*

| Matriz | RC | cRC | MBRP | RP | EE |
|---|---|---|---|---|---|
| Entrenamiento | n = 10 | n = 9 | n = 27 | n = 28 | n = 5 |
| Validación | n = 5 | n = 5 | n = 13 | n = 14 | n = 2 |

*RC: respuesta completa, cRC: respuesta cercana a la respuesta completa, MBRP: muy buena respuesta parcial, RP: respuesta parcial, EE: enfermedad estable.*

Como lista "inicial" de genes se utilizaron los 1.438 genes diferencialmente expresados entre los distintos grupos de respuesta a *p*-valor < 0,05 obtenidos con el algoritmo *limma* para múltiples grupos. A partir de esta lista se generó una segunda lista mediante el filtrado de los genes con el algoritmo *Boruta*. Esta segunda lista consistió en los 28 genes con un mayor valor de "importancia" considerando los grupos de respuesta de este estudio. Con las dos listas ya definidas, se procedió al análisis predictivo.

El mejor modelo de predicción fue ajustado utilizando el método PLS con 10 factores y los 1.438 genes de la lista "inicial". Esta predicción logró una TAC del 72% con buenas tasas de acierto en la mayoría de los grupos de respuesta. Los grupos que presentaron una menor tasa de acierto fueron cRC, con un 40% de acierto y tres muestras clasificadas como RP y el grupo EE, con un 50% de acierto, aunque únicamente con dos muestras utilizadas en la validación. Los resultados de esta predicción se recogen en la **Tabla 4.52**.

***Tabla 4.52.*** *Matriz de contingencia correspondiente al modelo de predicción de la respuesta por grupo para el estudio de Terragna (2016) que obtuvo mejores tasas de acierto (PLS, 10 factores [NF] y 1.438 genes).*

| **PLS, NF = 10, 1438 genes** | | Grupo real | | | | |
|---|---|---|---|---|---|---|
| | | RC | cRC | MBRP | RP | EE |
| Predicción | RC | **4** | 0 | 0 | 0 | 0 |
| | cRC | 0 | **2** | 1 | 0 | 0 |
| | MBRP | 0 | 0 | **9** | 2 | 0 |
| | RP | 1 | 3 | 3 | **12** | 1 |
| | EE | 0 | 0 | 0 | 0 | **1** |

*En verde se muestran los pacientes a los que se asignó correctamente el grupo de respuesta tras la predicción, en rojo se muestran los errores junto con la respuesta que alcanzan dichos pacientes.*

### 4.5.4. Consideraciones generales del análisis de predicción

En general, se podría afirmar que los análisis de predicción realizados en este capítulo han tenido un rendimiento relativamente bueno, principalmente cuando se han llevado a cabo las predicciones considerando una clasificación binaria de la respuesta, como fueron los casos de la predicción R *vs.* NR, y sobre todo en la predicción de la RC, donde en los tres regímenes de tratamiento analizados la TAC mediana ha rondado el 90% de aciertos (**Figura 4.165**).

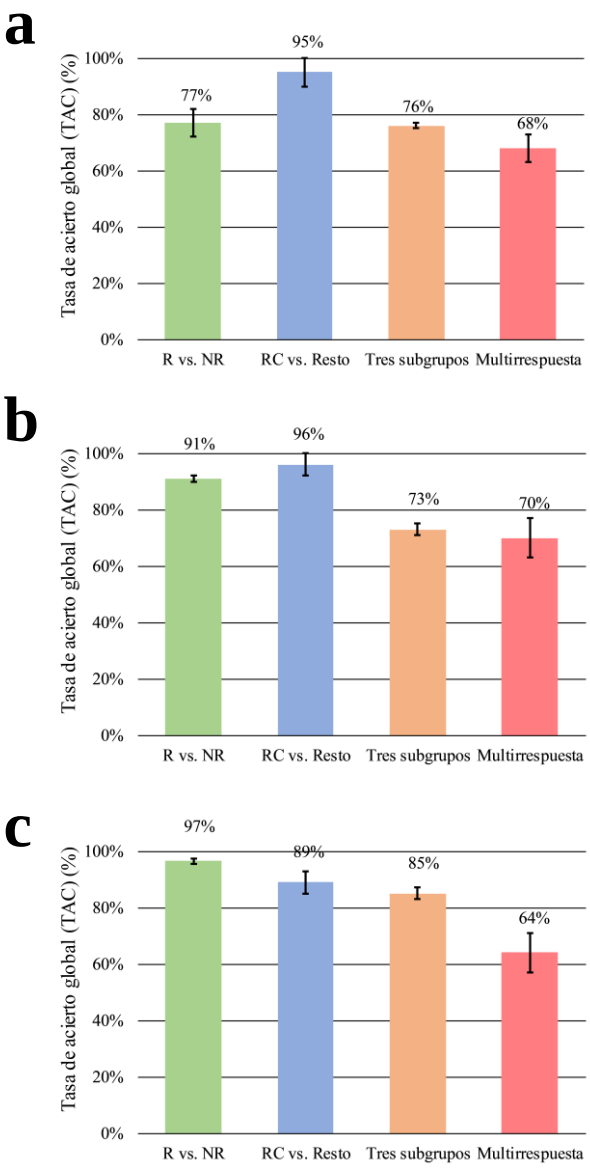



***Figura 4.165**. Tasa de acierto global (TAC) mediana de los estudios seleccionados para la predicción de la respuesta por régimen de tratamiento y aproximación analítica. Los paneles corresponden a los regímenes de tratamiento **a)** bortezomib en monoterapia, **b)** bortezomib en terapia combinada con otros fármacos excepto IMiDs y **c)** bortezomib en terapia combinada con IMiDs Los resultados se muestran como porcentaje y las barras de error corresponden a las MAD de los estudios recogidos en los estudios de las diferentes aproximaciones analíticas. OR = pacientes respondedores, NR = pacientes no respondedores, RC = respuesta completa.*

Sin embargo, esta consideración puede ser valorada como demasiado reduccionista, ya que el hecho de que una predicción logre buenos índices de TAC no significa necesariamente que se trate de una predicción adecuada. De hecho, factores como el tamaño muestral o la ratio entre los grupos analizados pueden ser determinantes sobre el rendimiento de un modelo de predicción, conduciendo a que su comportamiento pueda ser considerado como óptimo o no. Un caso extremo que demuestra esta circunstancia es la predicción OR *vs.* NR del **Apartado 4.5.3**, donde en el estudio CoMMpass (2017), el modelo de predicción no es capaz de predecir correctamente la respuesta de ninguno de

los pacientes del grupo NR a pesar de obtener una TAC del 96%. En este caso, una ratio de 36,6 a favor del grupo OR fue lo que propició la imposibilidad de una predicción satisfactoria del grupo NR, ya que la firma de expresión génica que podría caracterizar a este segundo grupo quedó totalmente infraestimada debido a su reducido tamaño muestral en comparación con el primer grupo. No obstante, teniendo en cuenta todas estas particularidades, es posible resaltar el buen desempeño de los métodos predictivos en los análisis en las que la respuesta estuvo codificada de forma binaria, y particularmente en aquellas en que la predicción fue RC *vs.* Resto de respuestas, donde, además, en la mayor parte de los estudios analizados se obtuvieron altos valores tanto para el RVP como para el RVN. Este resultado contradice en buena medida lo expuesto en el trabajo de Zhang y colaboradores[198] quienes descartaron la codificación binaria al considerarla, por un lado, arbitraria a la hora de la confección de los dos grupos y, por otro lado, arriesgada debido la pérdida de información que supone este modo de agrupamiento de los pacientes. De esta manera, el estudio predictivo que condujeron estos autores se realizó considerando en primer lugar los múltiples niveles de respuesta provistos en cada uno de los estudios que reanalizaron, y, en segundo lugar, llevando a cabo una recodificación de la respuesta en tres grupos. Ambas codificaciones de la respuesta fueron implementadas en este trabajo de tesis doctoral con fines comparativos. Así, las TAC obtenidas por Zhang sobre el estudio de Mulligan y colaboradores[528], fueron de 0,81 en el caso de la codificación en múltiples respuestas, y de 0,73, en el caso de los tres grupos de respuesta. De manera análoga, llevaron a cabo el análisis del estudio de Terragna y colaboradores[194], logrando unos TAC de 0,87 y 0,77 para la aproximación multirrespuesta y en tres grupos, respectivamente. Estos dos trabajos fueron también analizados en nuestro trabajo bajo los mismos criterios de codificación de la respuesta, obteniendo en el caso del estudio de Mulligan y colaboradores, unos TAC de 0,63/0,77 y 0,75/0,71 para las aproximaciones multirrespuesta y en tres grupos, respectivamente, considerando los dos regímenes de tratamiento en los que se analizó este estudio. Por su parte, para el estudio de Terragna y colaboradores, obtuvimos unos TAC de 0,72 y de 0,90 para las aproximaciones multirrespuesta y en tres grupos, respectivamente. Aunque los TAC obtenidos tanto en el trabajo de Zhang y colaboradores, como en este trabajo, pueden ser comparables y además relativamente altos, al situarse por encima de 0,7, como se viene indicando, el TAC no debe ser la única medida a tener en cuenta a la hora de considerar el rendimiento de un análisis predictivo. Así, en el caso del estudio de Mulligan y colaboradores, observamos en nuestro trabajo que, aunque los resultados de la predicción multirrespuesta sí fueron adecuados prediciendo correctamente la mayor parte de los grupos, en el caso de la respuesta codificada en tres grupos no se logró la correcta clasificación de ninguna respuesta del G1. Por su parte, en el estudio de Terragna y colaboradores, el análisis multirrespuesta no logró capturar de forma satisfactoria los grupos de respuesta menos representados (cRC y EE), y en el caso del análisis de tres grupos de respuesta solamente logró un 20% de clasificaciones correctas en el G1. Además, las predicciones de respuestas múltiples tuvieron un problema adicional y es que el modelo óptimo utilizó

listas con un gran número de genes, lo que podría reducir el interés clínico de estos modelos con vistas a la aplicabilidad en la realidad asistencial.

Con todo esto, como se venía indicando, los modelos predictivos que mostraron un mejor rendimiento fueron los aplicados sobre una codificación binaria de la respuesta, principalmente en la aproximación RC *vs.* Resto de respuestas, obteniendo altas tasas de acierto en los análisis predictivos y con un número razonable de genes. Este resultado, no obstante, estaría en discordancia con el obtenido por Amin y colaboradores[197], quienes proponen que la expresión génica por sí sola no es capaz de producir una predicción adecuada de la RC, ya que a través de una batería de 7 métodos predictivos sobre cuatro estudios de expresión génica, entre los que se encuentra el estudio de Mulligan, logran unas TAC que oscilan en el rango de 0,44 a 0,78. Estas diferencias en cuanto al poder predictivo entre el trabajo de Amin y nuestro trabajo pueden ser debidas a factores como la referencia utilizada en la normalización de los microarrays, al método de selección de genes o a los métodos empleados para la predicción.

Así, en lo relativo a la referencia de análisis utilizada, en nuestro trabajo se optó por utilizar la referencia provista por BrainArray, buscando de esta manera eliminar los elementos que pudieran generar ambigüedad a la hora de definir la expresión de los genes analizados, como puede ser el hecho de trabajar con varios *probesets* que interroguen un mismo gen. En lo que respecta al método de selección de genes, en nuestro análisis se optó por seleccionar los genes en base a su estudio previo por metaanálisis, seguido de un filtrado de genes por técnicas de RF con el paquete *Boruta* en R, mientras que en el estudio de Amin la selección fue llevada a cabo a través de su varianza, por métodos LASSO o bien a través de regresión contraída (*ridge regression*). Finalmente, en lo concerniente a los métodos predictores empleados, a pesar de que tanto en este trabajo como en el trabajo de Amin y colaboradores se emplearon en común métodos como KNN y SVM, los resultados obtenidos en nuestro trabajo demostraron, que estos dos métodos fueron superados en rendimiento claramente por PLS (**Anexo 43)**, obteniendo los mayores valores de TAC en la mayoría de las predicciones de la aproximación RC *vs.* Resto de pacientes.

Sin embargo, a pesar de los buenos resultados en la aproximación RC vs. Resto de pacientes, el modelo de predicción en el estudio CoMMpass (2017) bajo el régimen de tratamiento con bortezomib en combinación con IMiDS no funcionó correctamente, ya que este modelo solo logró un RVP = 0,04. La causa de esta baja tasa de acierto en el grupo de RC podría estar asociada a la variabilidad en los tratamientos que recibieron los pacientes de este estudio, ya que, pese a estar todos ellos tratados con bortezomib e IMiDs, se emplearon diferentes compuestos dentro de este segundo grupo, como pomalidomida, lenalidomida o talidomida. Además, muchos de estos pacientes recibieron fármacos adicionales como la ciclofosfamida, el daratumumab o el melfalán. Esto condujo a que no solo en la aproximación RC vs. Resto de pacientes fallasen los modelos de predicción en este estudio, sino que dichos fallos se produjeron en todas las

aproximaciones realizadas con este estudio asociadas al régimen de tratamiento mencionado.

Por tanto, a la vista de los resultados obtenidos, podría indicarse que los modelos predictivos utilizando la expresión génica para la predicción de la respuesta funcionan adecuadamente, principalmente para la predicción de la RC, aunque es conveniente tener bajo control variables como la heterogeneidad de los tratamientos aplicados para tener un resultado óptimo.

No obstante, con la vista puesta en futuros análisis de predicción se recomendaría en primer lugar ampliar todas las series con nuevos estudios de respuesta en MM y, en segundo lugar, y siempre que el número de muestras lo permita, llevar a cabo una estratificación de los pacientes de MM en función de los subgrupos citogenéticos asociados a esta enfermedad, para reducir de este modo la heterogeneidad intrínseca de esta patología. Finalmente, tal y como recomiendan Amin y colaboradores[197], se propone también la realización de estas predicciones integrando datos de distintas fuentes ómicas, como SNP, microARN o proteínas, entre otros, lo que podría mejorar los resultados de la predicción de la respuesta al añadir información biológica que la expresión génica por sí sola no puede aportar.

# 5. Conclusiones

## 5.1. Conclusiones relativas al diseño de un *pipeline* para el análisis de datos de RNA-seq.

**1)** El *pipeline* con una mejor puntuación para el análisis de la expresión génica con datos en crudo, en cuanto a su precisión y exactitud, ha sido el que utiliza como algoritmo de recortado *Trimmomatic*, como algoritmo de alineamiento *RUM*, como método de contaje *HTSeq Union* y como método de normalización TMM.

**2)** Las etapas más críticas en el proceso de análisis de la expresión génica en crudo son el contaje y la normalización, y los métodos de análisis que presentan un mejor rendimiento en estas dos etapas han sido *HTSeq Union* y TMM, respectivamente.

**3)** Los algoritmos de pseudoalineamiento son una excelente alternativa a la metodología tradicional de análisis de datos de RNA-seq, gracias a su gran velocidad de ejecución, su elevada precisión y al hecho de que ahorran al investigador mucho esfuerzo computacional.

**4)** El método de análisis de la expresión génica diferencial más equilibrado, considerando todos los criterios de bondad planteados en este trabajo, ha sido *limma trend*.

**5)** El rendimiento de los métodos de expresión génica diferencial, depende en mayor medida de las características del propio método que de la distribución estadística asumida *a priori* para los datos.

## 5.2. Conclusiones relativas a la comparación de la técnica RNA-seq Illumina 2500 frente al microarray HTA2.0 de Affymetrix.

**1)** La RNA-seq muestra superioridad sobre el microarray en la determinación de la expresión génica en crudo, si bien el microarray presenta un rendimiento similar a la RNA-seq cuando se eliminan los genes con valores atípicos.

**2)** La técnica RNA-seq presenta cierta ventaja sobre el microarray a la hora de la detección, de forma precisa y exacta, de la expresión génica diferencial. Sin embargo, cuando se trata de escenarios de análisis con un número intermedio de cambios de expresión génica, ambas tecnologías tienen un rendimiento similar.

## 5.3. Conclusiones relativas a la determinación de perfiles de expresión génica asociados a fármacos antimieloma mediante metaanálisis en líneas celulares de mieloma múltiple.

***1) Aspectos generales:***

**a)** Se logró identificar una firma de expresión génica para cada uno de los fármacos analizados para el tratamiento del mieloma mediante técnicas de metaanálisis.

**b)** Existe una elevada similitud entre las firmas génicas de los agentes inmunomoduladores lenalidomida y pomalidomida. Por el contrario, las firmas génicas de los agentes hipometilantes azacitidina y decitabina, son completamente diferentes entre sí.

***2) Aspectos específicos:***

**a) Melfalán**

**i)** El aumento del tiempo de tratamiento con melfalán produce un incremento de la expresión génica en las células tratadas con este fármaco.

**ii)** El melfalán produce una modificación estadísticamente significativa de la expresión de genes implicados en el ciclo celular y de la vía de p53.

**b) Dexametasona**

**i)** El aumento necesario del tiempo de tratamiento con dexametasona conduce a una disminución del efecto producido por este fármaco sobre la expresión génica.

**ii)** La dexametasona produce la desregulación de genes de unión a citocinas, que podrían asociarse en unos casos con mecanismos de resistencia a este compuesto, y en otros podrían tener un potencial interés terapéutico.

**c) Bortezomib**

**i)** El efecto del bortezomib sobre la expresión génica se incrementa a tiempos de tratamiento largos y concentraciones mayores de fármaco.

**ii)** El tratamiento con bortezomib modifica la expresión de genes de la vía del proteasoma, incrementando la expresión de genes que codifican diferentes subunidades del proteasoma, lo que podría estar asociado a un mecanismo de quimiorresistencia.

**iii)** El bortezomib altera la expresión de genes implicados en el procesamiento de proteínas en el retículo endoplásmico, lo que llevaría a la generación de estrés en este orgánulo desencadenando finalmente la apoptosis celular.

**d) Lenalidomida**

**i)** La lenalidomida actúa reduciendo la expresión de los genes de la vía de biosíntesis de ribosomas y el protooncogén *MYC*, lo que podría conducir a la reducción de la síntesis de proteínas y al bloqueo de la proliferación celular.

**ii)** La lenalidomida actúa sobre la vía de presentación y procesamiento de antígenos, sobreexpresando los genes que codifican los MHC tipo I y tipo II, lo que conduciría a la modulación de la respuesta inmune.

**e) Pomalidomida**

**i)** La pomalidomida actúa sobreexpresando diferentes componentes del MHC tipo II, y el componente *HSPA1A* del MCH tipo I, llevando a cabo un mecanismo de modulación de la respuesta inmune muy similar al de la lenalidomida.

**ii)** El tratamiento con pomalidomida conduce a la sobreexpresión de la interleucina *IL7*, que estimula la proliferación de las células del linaje linfoide.

**f) Panobinostat**

**i)** El panobinostat modifica la expresión de un gran número de genes implicados en la replicación del ADN, sin embargo, esto podría no ser suficiente para producir la detención del crecimiento celular.

**ii)** El panobinostat desregula la expresión de genes implicados en el ciclo celular como *PLK1*, que es un inductor de la apoptosis en células tumorales.

**g) Azacitidina**

**i)** Los tiempos cortos de tratamiento podrían favorecer el mecanismo de desregulación de la expresión génica de la azacitidina.

**ii)** La azacitidina actúa sobre los genes de la vía de biosíntesis de esteroides, produciendo una regulación negativa de la biosíntesis de colesterol, lo que podría tener un efecto citotóxico sobre las células cancerígenas.

**iii)** La azacitidina sobreexpresa los genes *CYP27B1*, implicado en el desarrollo del tejido biomineral, y *NBR1*, implicado en el mantenimiento óseo, lo que podría conducir a la mejoría en el proceso de osificación en pacientes tratados con este compuesto.

**h) Decitabina**

**i)** La decitabina actúa sobre la vía de diferenciación de osteoclastos sobreexpresado genes como *JUNB*, cuya función debe se esclarecida en futuros estudios.

**ii)** El tratamiento con decitabina actúa sobre la vía de diferenciación de células Th1 y Th2 desregulado la expresión de genes que codifican distintas subunidades de los receptores de IFN-γ e *IL2* lo que podría favorecer un efecto antimieloma en los pacientes tratados con este compuesto.

**i) JQ1**

**i)** La desregulación de la expresión génica producida por JQ1 se incrementa a medida que aumenta el tiempo de tratamiento y la concentración aplicada de fármaco.

**ii)** JQ1 reduce la expresión de *MYC* y de un gran número de genes implicados en el procesamiento del ARNr y en la biogénesis de complejos de ribonucleoproteínas, lo que podría conducir a la inhibición del programa traduccional de la célula tumoral en MM.

**j) Amilorida**

**i)** La amilorida desregula la vía de señalización HIF-1 y produce la infraexpresión del gen *LDHA*, por lo que el uso terapéutico de este compuesto podría ser beneficioso como terapia de rescate en pacientes resistentes a bortezomib.

**ii)** La amilorida produce la sobreexpresión de los componentes del espliceosoma *DHX16* y *PRPF3*.

**k) TG003**

El TG003 presenta una desregulación de los procesos metabólicos al ser aplicado sobre células de MM muy similar a la amilorida.

**l) Interferón γ**

El tratamiento con INF-γ conduce a la sobreexpresión de los genes *IRF1* y *STAT1*, cruciales para la respuesta a esta molécula. Además, también produce la sobreexpresión del gen *STAT3*, pudiendo tratarse de un mecanismo de regulación cruzada para hacer frente a la actividad supresora tumoral de *STAT1*.

## 5.4. Conclusiones relativas a la determinación de perfiles de expresión génica asociados a la respuesta en pacientes con mieloma múltiple.

**1)** Se logró identificar un perfil de expresión génica asociado a los diferentes tipos de respuesta analizados en todos los regímenes de tratamiento estudiados en pacientes con mieloma múltiple.

**2)** Los pacientes respondedores a bortezomib en monoterapia presentan una regulación negativa de la iniciación de la traducción, mostrando infraexpresión en genes como *LAMTOR1* y *LAMTOR4*, lo que puede producir la inactivación del complejo mTORC1 inhibiendo la traducción de proteínas.

**3)** La sobreexpresión de *CYLD* y la infraexpresión de *NFKB1* en pacientes respondedores a bortezomib en combinación con otros fármacos (salvo agentes inmunomoduladores) podría indicar la activación de la vía de señalización de NF-κB, lo que favorecería la respuesta al tratamiento de estos pacientes.

**4)** Los pacientes con respuesta completa a bortezomib en combinación con agentes inmunomoduladores presentan una desregulación significativa de genes implicados en transporte celular dependiente de citoesqueleto, lo que podría sugerir la importancia de este proceso en la respuesta a estos fármacos.

## 5.5. Conclusiones relativas a la predicción de la respuesta al tratamiento en pacientes con mieloma múltiple.

**1)** Los análisis de predicción de la respuesta utilizando datos de expresión génica consiguen buenos resultados en la mayoría de los casos. Fueron especialmente efectivas las predicciones realizadas sobre los grupos de respuesta completa.

**2)** En general, las tasas globales de acierto de los análisis predictivos de tres grupos y multirrespuesta son notablemente inferiores que las alcanzadas en los análisis predictivos de respuesta dicotómica.

**3)** El método predictivo con mejores tasas globales de acierto en la mayor parte de los estudios realizados es la técnica PLS.

**4)** La predicción de la respuesta en los estudios en los que se combinan múltiples fármacos no son adecuadas ya que los modelos fallaron en la predicción de, al menos, uno de los grupos de respuesta, con lo que es recomendable su estratificación por grupos de tratamiento siempre que el tamaño muestral lo permita.

# 6. Bibliografía

1. Palumbo A and Anderson K. Multiple myeloma. N Engl J Med. 2011;364(11):1046-1060.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7-30.

3. Rajkumar SV and Kumar S. Multiple Myeloma: Diagnosis and Treatment. Mayo Clin Proc. 2016;91(1):101-119.

4. Howlader N, Noone A, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD. 2019:https://seer.cancer.gov/csr/1975_2007/.

5. Boyd KD, Ross FM, Chiecchio L, et al. Gender disparities in the tumor genetics and clinical outcome of multiple myeloma. Cancer Epidemiol Biomarkers Prev. 2011;20(8):1703-1707.

6. Landgren O and Weiss BM. Patterns of monoclonal gammopathy of undetermined significance and multiple myeloma in various ethnic/racial groups: support for genetic factors in pathogenesis. Leukemia. 2009;23(10):1691-1697.

7. Kristinsson SY, Landgren O, Dickman PW, Derolf AR, Bjorkholm M. Patterns of survival in multiple myeloma: a population-based study of patients diagnosed in Sweden from 1973 to 2003. J Clin Oncol. 2007;25(15):1993-1999.

8. Brenner H, Gondos A, Pulte D. Recent major improvement in long-term survival of younger patients with multiple myeloma. Blood. 2008;111(5):2521-2526.

9. Anaya A, Ramón y Cajal Junquera, S., Langa Langa M. Las células cianófilas de Cajal. Rev Esp Patol. 2002;35(2):233-237.

10. Merino C and Fernández Ruíz B. Ramón y Cajal y la ciencia española. 2005.

11. Ortiz Hidalgo C. De las células plasmáticas al mieloma múltiple. Una breve perspectiva histórica. Patologia. 2011;49(2):120-131.

12. Kyle RA. Multiple myeloma: an odyssey of discovery. Br J Haematol. 2000;111(4):1035-1044.

13. Clamp JR. Some aspects of the first recorded case of multiple myeloma. Lancet. 1967;2(7530):1354-1356.

14. Heller J. Die mikroskopisch-chemisch-pathologische untersuchung. Braumüller & Seidel. 1846:576.

15. Fleischer R. Ueber das Vorkommen des sogenannetn Bence Jones' schen Eiwisskörpers im normalen Knochenmark. Archiv für Pathologie Anatomie und Physiologie. 1880;80:482-489.

16. Weber FP and Ledingham JC. A Note on the Histology of a Case of Myelomatosis (Multiple Myeloma) with Bence-Jones Protein in the Urine (Myelopathic Albumosuria). Proc R Soc Med. 1909;2(Pathol Sect):193-206.

17. von Behring E and Kitasato S. Ueber das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. Deutsche Medicinische Wochenschrift. 1890;49:1113-1114.

18. Tiselius A. Electrophoresis of serum globulin: Electrophoretic analysis of normal and immune sera. Biochem J. 1937;31(9):1464-1477.

19. Tiselius A and Kabat EA. An Electrophoretic Study of Immune Sera and Purified Antibody Preparations. J Exp Med. 1939;69(1):119-131.

20. Longsworth LG, Shedlovsky T, Macinnes DA. Electrophoretic Patterns of Normal and Pathological Human Blood Serum and Plasma. J Exp Med. 1939;70(4):399-413.

21. Grabar P and Williams C. Méthode permettant l'étude conjuguée des propriétés électrophorétiques et immunochimiques d'un mélange de protéines. Application au sérum sanguin. . Biochim Biophys Acta. 1953;10:193-194.

22. KORNGOLD L and LIPARI R. Multiple-myeloma proteins. III. The antigenic relationship of Bence Jones proteins to normal gammaglobulin and multiple-myeloma serum proteins. Cancer. 1956;9(2):262-272.

23. Anirkin M. Die Intravitale Untersuchungsmethodik des Knochenmarks. Folia Haematologica. 1929;38:233-240.

24. Bersagel DE, Sprague CC, Austin C, Griffith KM. Evaluation of new chemotherapeutic agents in the treatment of multiple myeloma. IV. L-Phenylalanine mustard (NSC-8806). Cancer Chemother Rep. 1962;21:87-99.

25. Alexanian R, Haut A, Khan AU, et al. Treatment for multiple myeloma. Combination chemotherapy with different melphalan dose regimens. JAMA. 1969;208(9):1680-1685.

26. Gregory WM, Richards MA, Malpas JS. Combination chemotherapy versus melphalan and prednisolone in the treatment of multiple myeloma: an overview of published trials. J Clin Oncol. 1992;10(2):334-342.

27. Barlogie B, Smith L, Alexanian R. Effective treatment of advanced multiple myeloma refractory to alkylating agents. N Engl J Med. 1984;310(21):1353-1356.

28. Attal M, Harousseau JL, Stoppa AM, et al. A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. Intergroupe Francais du Myelome. N Engl J Med. 1996;335(2):91-97.

29. Chng WJ, Glebov O, Bergsagel PL, Kuehl WM. Genetic events in the pathogenesis of multiple myeloma. Best Pract Res Clin Haematol. 2007;20(4):571-596.

30. Ribatti D. A historical perspective on milestones in multiple myeloma research. Eur J Haematol. 2018;100(3):221-228.

31. Alexander DD, Mink PJ, Adami HO, et al. Multiple myeloma: a review of the epidemiologic literature. Int J Cancer. 2007;120 Suppl 12:40-61.

32. Nieters A, Deeg E, Becker N. Tobacco and alcohol consumption and risk of lymphoma: results of a population-based case-control study in Germany. Int J Cancer. 2006;118(2):422-430.

33. Bergstrom A, Pisani P, Tenet V, Wolk A, Adami HO. Overweight as an avoidable cause of cancer in Europe. Int J Cancer. 2001;91(3):421-430.

34. Larsson SC and Wolk A. Body mass index and risk of multiple myeloma: a meta-analysis. Int J Cancer. 2007;121(11):2512-2516.

35. Britton JA, Khan AE, Rohrmann S, et al. Anthropometric characteristics and non-Hodgkin's lymphoma and multiple myeloma risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). Haematologica. 2008;93(11):1666-1677.

36. Birmann BM, Giovannucci E, Rosner B, Anderson KC, Colditz GA. Body mass index, physical activity, and risk of multiple myeloma. Cancer Epidemiol Biomarkers Prev. 2007;16(7):1474-1478.

37. Reeves GK, Pirie K, Beral V, et al. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. BMJ. 2007;335(7630):1134.

38. Vlajinac HD, Pekmezovic TD, Adanja BJ, et al. Case-control study of multiple myeloma with special reference to diet as risk factor. Neoplasma. 2003;50(1):79-83.

39. Tavani A, Pregnolato A, Negri E, et al. Diet and risk of lymphoid neoplasms and soft tissue sarcomas. Nutr Cancer. 1997;27(3):256-260.

40. Brown LM, Gridley G, Pottern LM, et al. Diet and nutrition as risk factors for multiple myeloma among blacks and whites in the United States. Cancer Causes Control. 2001;12(2):117-125.

41. Becker N. Epidemiology of multiple myeloma. Recent Results Cancer Res. 2011;183:25-35.

42. Landgren O, Zeig-Owens R, Giricz O, et al. Multiple Myeloma and Its Precursor Disease Among Firefighters Exposed to the World Trade Center Disaster. JAMA Oncol. 2018;4(6):821-827.

43. Koura DT and Langston AA. Inherited predisposition to multiple myeloma. Ther Adv Hematol. 2013;4(4):291-297.

44. Broderick P, Chubb D, Johnson DC, et al. Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. Nat Genet. 2011;44(1):58-61.

45. Martino A, Campa D, Jamroziak K, et al. Impact of polymorphic variation at 7p15.3, 3p22.1 and 2p23.3 loci on risk of multiple myeloma. Br J Haematol. 2012;158(6):805-809.

46. Grass S, Preuss KD, Ahlgrimm M, et al. Association of a dominantly inherited hyperphosphorylated paraprotein target with sporadic and familial multiple myeloma and monoclonal gammopathy of undetermined significance: a case-control study. Lancet Oncol. 2009;10(10):950-956.

47. Grass S, Preuss KD, Thome S, et al. Paraproteins of familial MGUS/multiple myeloma target family-typical antigens: hyperphosphorylation of autoantigens is a consistent finding in familial and sporadic MGUS/MM. Blood. 2011;118(3):635-637.

48. Fernandez de Larrea C, Kyle RA, Durie BG, et al. Plasma cell leukemia: consensus statement on diagnostic requirements, response criteria and treatment recommendations by the International Myeloma Working Group. Leukemia. 2013;27(4):780-791.

49. Kyle RA and Rajkumar SV. Multiple myeloma. N Engl J Med. 2004;351(18):1860-1873.

50. van Nieuwenhuijzen N, Spaan I, Raymakers R, Peperzak V. From MGUS to Multiple Myeloma, a Paradigm for Clonal Evolution of Premalignant Cells. Cancer Res. 2018;78(10):2449-2456.

51. Bergsagel PL. Impressions of the myeloma landscape. Blood. 2010;116(14):2403-2404.

52. Pawlyn C and Morgan GJ. Evolutionary biology of high-risk multiple myeloma. Nat Rev Cancer. 2017;17(9):543-556.

53. Manier S, Salem KZ, Park J, Landau DA, Getz G, Ghobrial IM. Genomic complexity of multiple myeloma and its clinical implications. Nat Rev Clin Oncol. 2017;14(2):100-113.

54. Kumar SK, Rajkumar V, Kyle RA, et al. Multiple myeloma. Nat Rev Dis Primers. 2017;3:17046.

55. Morgan GJ, Walker BA, Davies FE. The genetic architecture of multiple myeloma. Nat Rev Cancer. 2012;12(5):335-348.

56. Chng WJ, Kumar S, Vanwier S, et al. Molecular dissection of hyperdiploid multiple myeloma by gene expression profiling. Cancer Res. 2007;67(7):2982-2989.

57. Magrath I. The lymphoid neoplasms. London: Hodder Arnold; 2010.

58. Anderson KC. Progress and Paradigms in Multiple Myeloma. Clin Cancer Res. 2016;22(22):5419-5427.

59. Kumar SK, Rajkumar SV, Dispenzieri A, et al. Improved survival in multiple myeloma and the impact of novel therapies. Blood. 2008;111(5):2516-2520.

60. Kyle RA and Rajkumar SV. An overview of the progress in the treatment of multiple myeloma. Expert Rev Hematol. 2014;7(1):5-7.

61. Falco P, Bringhen S, Avonto I, et al. Melphalan and its role in the management of patients with multiple myeloma. Expert Rev Anticancer Ther. 2007;7(7):945-957.

62. Bergel F and Stock J. Cytotoxic alpha amino acids and peptides. Ann Rep Brit Empire Cancer Campgn. 1953;31:6-7.

63. Kondo N, Takahashi A, Ono K, Ohnishi T. DNA damage induced by alkylating agents and repair pathways. J Nucleic Acids. 2010;2010:543531.

64. Pieper RO and Erickson LC. DNA adenine adducts induced by nitrogen mustards and their role in transcription termination in vitro. Carcinogenesis. 1990;11(10):1739-1746.

65. Polavarapu A, Stillabower JA, Stubblefield SG, Taylor WM, Baik MH. The mechanism of guanine alkylation by nitrogen mustards: a computational study. J Org Chem. 2012;77(14):5914-5921.

66. Fernberg JO, Lewensohn R, Skog S. Cell cycle arrest and DNA damage after melphalan treatment of the human myeloma cell line RPMI 8226. Eur J Haematol. 1991;47(3):161-167.

67. Dorr RT. Cancer chemotherapy handbook. London: Kimpton; 1980.

68. Tsukidate K, Yamamoto K, Snyder JW, Farber JL. Microtubule antagonists activate programmed cell death (apoptosis) in cultured rat hepatocytes. Am J Pathol. 1993;143(3):918-925.

69. Harmon BV, Takano YS, Winterford CM, Potten CS. Cell death induced by vincristine in the intestinal crypts of mice and in a human Burkitt's lymphoma cell line. Cell Prolif. 1992;25(6):523-536.

70. Cline MJ. Effect of vincristine on synthesis of ribonucleic acid and protein in leukaemic leucocytes. Br J Haematol. 1968;14(1):21-29.

71. Creasey W. Vinca alkaloids and colchicine. In Sarotelli, A.C., Johns, D.G. (Eds.): Antineoplastic and Immunosuppresive Agents, Handbook of experimental Pharmacology, Part II. Springer. 1975:670-694.

72. Watanabe K and West WL. Calmodulin, activated cyclic nucleotide phosphodiesterase, microtubules, and vinca alkaloids. Fed Proc. 1982;41(7):2292-2299.

73. Giraud B, Hebert G, Deroussent A, Veal GJ, Vassal G, Paci A. Oxazaphosphorines: new therapeutic strategies for an old class of drugs. Expert Opin Drug Metab Toxicol. 2010;6(8):919-938.

74. Ahlmann M and Hempel G. The effect of cyclophosphamide on the immune system: implications for clinical cancer therapy. Cancer Chemother Pharmacol. 2016;78(4):661-671.

75. Fleer R and Brendel M. Toxicity, interstrand cross-links and DNA fragmentation induced by 'activated' cyclophosphamide in yeast. Chem Biol Interact. 1981;37(1-2):123-140.

76. Schmoll H. Review of etoposide single-agent activity. Cancer Treat Rev. 1982;9 Suppl:21-30.

77. Pommier Y, Leo E, Zhang H, Marchand C. DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chem Biol. 2010;17(5):421-433.

78. Hande KR. Etoposide: four decades of development of a topoisomerase II inhibitor. Eur J Cancer. 1998;34(10):1514-1521.

79. Keizer HG, Pinedo HM, Schuurhuis GJ, Joenje H. Doxorubicin (adriamycin): a critical review of free radical-dependent mechanisms of cytotoxicity. Pharmacol Ther. 1990;47(2):219-231.

80. Jung K and Reszka R. Mitochondria as subcellular targets for clinically useful anthracyclines. Adv Drug Deliv Rev. 2001;49(1-2):87-105.

81. Batist G, Ramakrishnan G, Rao CS, et al. Reduced cardiotoxicity and preserved antitumor efficacy of liposome-encapsulated doxorubicin and cyclophosphamide compared with conventional doxorubicin and cyclophosphamide in a randomized, multicenter trial of metastatic breast cancer. J Clin Oncol. 2001;19(5):1444-1454.

82. Gandhi V. Metabolism and mechanisms of action of bendamustine: rationales for combination therapies. Semin Oncol. 2002;29(4 Suppl 13):4-11.

83. Leoni LM and Hartley JA. Mechanism of action: the unique pattern of bendamustine-induced cytotoxicity. Semin Hematol. 2011;48 Suppl 1:S12-23.

84. Wooldridge JE, Anderson CM, Perry MC. Corticosteroids in advanced cancer. Oncology (Williston Park). 2001;15(2):225-34; discussion 234-6.

85. Bergsagel DE. Plasma cell myeloma. An interpretive review. Cancer. 1972;30(6):1588-1594.

86. Giles AJ, Hutchinson MND, Sonnemann HM, et al. Dexamethasone-induced immunosuppression: mechanisms and implications for immunotherapy. J Immunother Cancer. 2018;6(1):51-018-0371-5.

87. Newton R. Molecular mechanisms of glucocorticoid action: what is important? Thorax. 2000;55(7):603-613.

88. Anonymous Glucocorticoid Response Elements (GRE). In: Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Springer, Dordrecht. 2008.

89. Galustian C, Labarthe MC, Bartlett JB, Dalgleish AG. Thalidomide-derived immunomodulatory drugs as therapeutic agents. Expert Opin Biol Ther. 2004;4(12):1963-1970.

90. Quach H, Ritchie D, Stewart AK, et al. Mechanism of action of immunomodulatory drugs (IMiDS) in multiple myeloma. Leukemia. 2010;24(1):22-32.

91. Raza S, Safyan RA, Lentzsch S. Immunomodulatory Drugs (IMiDs) in Multiple Myeloma. Curr Cancer Drug Targets. 2017;17(9):846-857.

92. Powell RJ. New roles for thalidomide. BMJ. 1996;313(7054):377-378.

93. Corral LG and Kaplan G. Immunomodulation by thalidomide and thalidomide analogues. Ann Rheum Dis. 1999;58 Suppl 1:I107-13.

94. Barnhill RL, Doll NJ, Millikan LE, Hastings RC. Studies on the anti-inflammatory properties of thalidomide: effects on polymorphonuclear leukocytes and monocytes. J Am Acad Dermatol. 1984;11(5 Pt 1):814-819.

95. Kotla V, Goel S, Nischal S, et al. Mechanism of action of lenalidomide in hematological malignancies. J Hematol Oncol. 2009;2:36-8722-2-36.

96. Fink EC and Ebert BL. The novel mechanism of lenalidomide activity. Blood. 2015;126(21):2366-2369.

97. Bolzoni M, Storti P, Bonomini S, et al. Immunomodulatory drugs lenalidomide and pomalidomide inhibit multiple myeloma-induced osteoclast formation and the RANKL/OPG ratio in the myeloma microenvironment targeting the expression of adhesion molecules. Exp Hematol. 2013;41(4):387-97.e1.

98. Corral LG, Haslett PA, Muller GW, et al. Differential cytokine modulation and T cell activation by two distinct classes of thalidomide analogues that are potent inhibitors of TNF-alpha. J Immunol. 1999;163(1):380-386.

99. Reske T, Fulciniti M, Munshi NC. Mechanism of action of immunomodulatory agents in multiple myeloma. Med Oncol. 2010;27 Suppl 1:S7-13.

100. Syed YY. Lenalidomide: A Review in Newly Diagnosed Multiple Myeloma as Maintenance Therapy After ASCT. Drugs. 2017;77(13):1473-1480.

101. Hoy SM. Pomalidomide: A Review in Relapsed and Refractory Multiple Myeloma. Drugs. 2017;77(17):1897-1908.

102. McCurdy AR and Lacy MQ. Pomalidomide and its clinical potential for relapsed or refractory multiple myeloma: an update for the hematologist. Ther Adv Hematol. 2013;4(3):211-216.

103. Terpos E, Kanellias N, Christoulas D, Kastritis E, Dimopoulos MA. Pomalidomide: a novel drug to treat relapsed and refractory multiple myeloma. Onco Targets Ther. 2013;6:531-538.

104. Ferguson GD, Jensen-Pergakes K, Wilkey C, et al. Immunomodulatory drug CC-4047 is a cell-type and stimulus-selective transcriptional inhibitor of cyclooxygenase 2. J Clin Immunol. 2007;27(2):210-220.

105. Verhelle D, Corral LG, Wong K, et al. Lenalidomide and CC-4047 inhibit the proliferation of malignant B cells while expanding normal CD34+ progenitor cells. Cancer Res. 2007;67(2):746-755.

106. Manasanch EE and Orlowski RZ. Proteasome inhibitors in cancer therapy. Nat Rev Clin Oncol. 2017;14(7):417-433.

107. Dou QP and Li B. Proteasome inhibitors as potential novel anticancer agents. Drug Resist Updat. 1999;2(4):215-223.

108. Adams J. The proteasome: a suitable antineoplastic target. Nat Rev Cancer. 2004;4(5):349-360.

109. Obeng EA, Carlson LM, Gutman DM, Harrington WJ,Jr, Lee KP, Boise LH. Proteasome inhibitors induce a terminal unfolded protein response in multiple myeloma cells. Blood. 2006;107(12):4907-4916.

110. Kouroukis CT, Baldassarre FG, Haynes AE, et al. Bortezomib in multiple myeloma: a practice guideline. Clin Oncol (R Coll Radiol). 2014;26(2):110-119.

111. Merin NM and Kelly KR. Clinical use of proteasome inhibitors in the treatment of multiple myeloma. Pharmaceuticals (Basel). 2014;8(1):1-20.

112. Chen D, Frezza M, Schmitt S, Kanwar J, Dou QP. Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives. Curr Cancer Drug Targets. 2011;11(3):239-253.

113. Chen D and Dou QP. The ubiquitin-proteasome system as a prospective molecular target for cancer treatment and prevention. Curr Protein Pept Sci. 2010;11(6):459-470.

114. Bonvini P, Zorzi E, Basso G, Rosolen A. Bortezomib-mediated 26S proteasome inhibition causes cell-cycle arrest and induces apoptosis in CD-30+ anaplastic large cell lymphoma. Leukemia. 2007;21(4):838-842.

115. Moreau P, Richardson PG, Cavo M, et al. Proteasome inhibitors in multiple myeloma: 10 years later. Blood. 2012;120(5):947-959.

116. Kuhn DJ, Chen Q, Voorhees PM, et al. Potent activity of carfilzomib, a novel, irreversible inhibitor of the ubiquitin-proteasome pathway, against preclinical models of multiple myeloma. Blood. 2007;110(9):3281-3290.

117. Demo SD, Kirk CJ, Aujay MA, et al. Antitumor activity of PR-171, a novel irreversible inhibitor of the proteasome. Cancer Res. 2007;67(13):6383-6391.

118. Stapnes C, Doskeland AP, Hatfield K, et al. The proteasome inhibitors bortezomib and PR-171 have antiproliferative and proapoptotic effects on primary human acute myeloid leukaemia cells. Br J Haematol. 2007;136(6):814-828.

119. Raedler LA. Ninlaro (Ixazomib): First Oral Proteasome Inhibitor Approved for the Treatment of Patients with Relapsed or Refractory Multiple Myeloma. Am Health Drug Benefits. 2016;9(Spec Feature):102-105.

120. Offidani M, Corvatta L, Caraffa P, Gentili S, Maracci L, Leoni P. An evidence-based review of ixazomib citrate and its potential in the treatment of newly diagnosed multiple myeloma. Onco Targets Ther. 2014;7:1793-1800.

121. Chauhan D, Tian Z, Zhou B, et al. In vitro and in vivo selective antitumor activity of a novel orally bioavailable proteasome inhibitor MLN9708 against multiple myeloma cells. Clin Cancer Res. 2011;17(16):5311-5321.

122. Copeland A, Buglio D, Younes A. Histone deacetylase inhibitors in lymphoma. Curr Opin Oncol. 2010;22(5):431-436.

123. Archer SY and Hodin RA. Histone acetylation and cancer. Curr Opin Genet Dev. 1999;9(2):171-174.

124. Harada T, Hideshima T, Anderson KC. Histone deacetylase inhibitors in multiple myeloma: from bench to bedside. Int J Hematol. 2016;104(3):300-309.

125. Eckschlager T, Plch J, Stiborova M, Hrabeta J. Histone Deacetylase Inhibitors as Anticancer Drugs. Int J Mol Sci. 2017;18(7):10.3390/ijms18071414.

126. Andreu-Vieyra CV and Berenson JR. The potential of panobinostat as a treatment option in patients with relapsed and refractory multiple myeloma. Ther Adv Hematol. 2014;5(6):197-210.

127. Tzogani K, van Hennik P, Walsh I, et al. EMA Review of Panobinostat (Farydak) for the Treatment of Adult Patients with Relapsed and/or Refractory Multiple Myeloma. Oncologist. 2018;23(5):631-636.

128. Moore D. Panobinostat (Farydak): A Novel Option for the Treatment of Relapsed Or Relapsed and Refractory Multiple Myeloma. P T. 2016;41(5):296-300.

129. Pei XY, Dai Y, Grant S. Synergistic induction of oxidative injury and apoptosis in human multiple myeloma cells by the proteasome inhibitor bortezomib and histone deacetylase inhibitors. Clin Cancer Res. 2004;10(11):3839-3852.

130. Richardson PG, Mitsiades CS, Laubach JP, et al. Preclinical data and early clinical experience supporting the use of histone deacetylase inhibitors in multiple myeloma. Leuk Res. 2013;37(7):829-837.

131. Brody T. Clinical trials : study design, endpoints and biomarkers, drug safety, FDA and ICH guidelines. ; 2016.

132. Campestri R and Iastrebner C. Agentes hipometilantes. Hematología. 2010;14(3):135-140.

133. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429(6990):457-463.

134. Taylor SM and Jones PA. Mechanism of action of eukaryotic DNA methyltransferase. Use of 5-azacytosine-containing DNA. J Mol Biol. 1982;162(3):679-692.

135. Bender CM, Zingg JM, Jones PA. DNA methylation as a target for drug design. Pharm Res. 1998;15(2):175-187.

136. Leone G, Voso MT, Teofili L, Lubbert M. Inhibitors of DNA methylation in the treatment of hematological malignancies and MDS. Clin Immunol. 2003;109(1):89-102.

137. Derissen EJ, Beijnen JH, Schellens JH. Concise drug review: azacitidine and decitabine. Oncologist. 2013;18(5):619-624.

138. Jones PA. Altering gene expression with 5-azacytidine. Cell. 1985;40(3):485-486.

139. Hollenbach PW, Nguyen AN, Brady H, et al. A comparison of azacitidine and decitabine activities in acute myeloid leukemia cell lines. PLoS One. 2010;5(2):e9001.

140. Stresemann C and Lyko F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. Int J Cancer. 2008;123(1):8-13.

141. Zhang W, Chen Y, Pei X, Zang Y, Han S. Effects of Decitabine on the proliferation of K562 cells and the expression of DR4 gene. Saudi J Biol Sci. 2018;25(2):242-247.

142. Nguyen AN, Hollenbach PW, Richard N, et al. Azacitidine and decitabine have different mechanisms of action in non-small cell lung cancer cell lines. Lung Cancer (Auckl). 2010;1:119-140.

143. Lonial S, Durie B, Palumbo A, San-Miguel J. Monoclonal antibodies in the treatment of multiple myeloma: current status and future perspectives. Leukemia. 2016;30(3):526-535.

144. Sherbenou DW, Mark TM, Forsberg P. Monoclonal Antibodies in Multiple Myeloma: A New Wave of the Future. Clin Lymphoma Myeloma Leuk. 2017;17(9):545-554.

145. de Weers M, Tai YT, van der Veer MS, et al. Daratumumab, a novel therapeutic human CD38 monoclonal antibody, induces killing of multiple myeloma and other hematological tumors. J Immunol. 2011;186(3):1840-1848.

146. Overdijk MB, Verploegen S, Bogels M, et al. Antibody-mediated phagocytosis contributes to the anti-tumor activity of the therapeutic antibody daratumumab in lymphoma and multiple myeloma. MAbs. 2015;7(2):311-321.

147. Lammerts van Bueren J, A1 Jakobs D, A1 Kaldenhoven N, et al. Direct in vitro comparison of daratumumab with surrogate analogs of CD38 antibodies MOR03087, SAR650984 and Ab79. Presented at the 56th ASH Annual Meeting and Exposition, San Francisco. 2014.

148. Malaer JD and Mathew PA. CS1 (SLAMF7, CD319) is an effective immunotherapeutic target for multiple myeloma. Am J Cancer Res. 2017;7(8):1637-1641.

149. Collins SM, Bakan CE, Swartzel GD, et al. Elotuzumab directly enhances NK cell cytotoxicity against myeloma via CS1 ligation: evidence for augmented NK cell function complementing ADCC. Cancer Immunol Immunother. 2013;62(12):1841-1849.

150. Kaminska T, Dmoszynska A, Cioch M, et al. Interferon gamma as immunomodulator in a patient with multiple myeloma. Arch Immunol Ther Exp (Warsz). 1999;47(2):107-112.

151. Gado K, Domjan G, Hegyesi H, Falus A. Role of INTERLEUKIN-6 in the pathogenesis of multiple myeloma. Cell Biol Int. 2000;24(4):195-209.

152. Jernberg-Wiklund H, Pettersson M, Nilsson K. Recombinant interferon-gamma inhibits the growth of IL-6-dependent human multiple myeloma cell lines in vitro. Eur J Haematol. 1991;46(4):231-239.

153. Dyson MH, Rose S, Mahadevan LC. Acetyllysine-binding and function of bromodomain-containing proteins in chromatin. Front Biosci. 2001;6:D853-65.

154. Winston F and Allis CD. The bromodomain: a chromatin-targeting module? Nat Struct Biol. 1999;6(7):601-604.

155. Delmore JE, Issa GC, Lemieux ME, et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Cell. 2011;146(6):904-917.

156. Holien T, Vatsveen TK, Hella H, Waage A, Sundan A. Addiction to c-MYC in multiple myeloma. Blood. 2012;120(12):2450-2453.

157. Filippakopoulos P, Qi J, Picaud S, et al. Selective inhibition of BET bromodomains. Nature. 2010;468(7327):1067-1073.

158. Ghurye RR, Stewart HJ, Chevassut TJ. Bromodomain inhibition by JQ1 suppresses lipopolysaccharide-stimulated interleukin-6 secretion in multiple myeloma cells. Cytokine. 2015;71(2):415-417.

159. Norris AD and Calarco JA. Emerging Roles of Alternative Pre-mRNA Splicing Regulation in Neuronal Development and Function. Front Neurosci. 2012;6:122.

160. Chabot B and Shkreta L. Defective control of pre-messenger RNA splicing in human disease. J Cell Biol. 2016;212(1):13-27.

161. Aktas Samur A, Samur M, Minvielle S, et al. A Detailed Alternate Splicing Landscape in Multiple Myeloma with Significant Potential Biological and Clinical Implications. Blood. 2016;128(22):356.

162. Rashid N, Minvielle S, Magrangeas F, et al. Alternative Splicing Is a Frequent Event and Impacts Clinical Outcome in Myeloma: A Large RNA-Seq Data Analysis of Newly-Diagnosed Myeloma Patients. Blood. 2014;124(21):638.

163. Munshi N, Li C, Minvielle S, et al. Alternate Splicing Is a Frequent Event and Impacts Clinical Outcome in Myeloma: A High-Density Exon Array Analysis of Uniformly Treated Newly-Diagnosed Myeloma Patients. Blood. 2008;112(11):498.

164. Munshi NC and Avet-Loiseau H. Genomics in multiple myeloma. Clin Cancer Res. 2011;17(6):1234-1242.

165. Benos DJ. Amiloride: a molecular probe of sodium transport in tissues and cells. Am J Physiol. 1982;242(3):C131-45.

166. Kleyman TR and Cragoe EJ,Jr. Amiloride and its analogs as tools in the study of ion transport. J Membr Biol. 1988;105(1):1-21.

167. Chang JG, Yang DM, Chang WH, et al. Small molecule amiloride modulates oncogenic RNA alternative splicing to devitalize human cancer cells. PLoS One. 2011;6(6):e18643.

168. Slepkov ER, Rainey JK, Sykes BD, Fliegel L. Structural and functional analysis of the Na+/H+ exchanger. Biochem J. 2007;401(3):623-633.

169. Rojas EA, Corchete LA, San-Segundo L, et al. Amiloride, An Old Diuretic Drug, Is a Potential Therapeutic Agent for Multiple Myeloma. Clin Cancer Res. 2017;23(21):6602-6615.

170. Chang WH, Liu TC, Yang WK, et al. Amiloride modulates alternative splicing in leukemic cells and resensitizes Bcr-AblT315I mutant cells to imatinib. Cancer Res. 2011;71(2):383-392.

171. Muraki M, Ohkawara B, Hosoya T, et al. Manipulation of alternative splicing by a newly developed inhibitor of Clks. J Biol Chem. 2004;279(23):24246-24254.

172. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol. 2007;14(3):185-193.

173. Fulciniti M, Bhasin M, Samur MK, et al. Functional and Clinical Relevance of Splicing Factor SRSF1 in Multiple Myeloma (MM). Blood. 2014;124(21):3388.

174. Grunstein M and Hogness DS. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci U S A. 1975;72(10):3961-3965.

175. Gergen JP, Stern RH, Wensink PC. Filter replicas and permanent collections of recombinant DNA plasmids. Nucleic Acids Res. 1979;7(8):2115-2136.

176. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-470.

177. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A. 1997;94(24):13057-13062.

178. Schulze A and Downward J. Navigating gene expression using microarrays--a technology review. Nat Cell Biol. 2001;3(8):E190-5.

179. Anonymous Analysing gene expression : a handbook of methods: possibilities and pitfalls. Weinheim ; Great Britain: Wiley-VCH; 2003.

180. Knudsen S. Guide to analysis of DNA microarray data. Hoboken, N.J.: Wiley-Liss; 2004.

181. Yuryev A. Gene expression profiling for targeted cancer treatment. Expert Opin Drug Discov. 2015;10(1):91-99.

182. Bellman RE. Dynamic programming. Princeton University Press, Princeton. 1957.

183. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-537.

184. Szalat R, Avet-Loiseau H, Munshi NC. Gene Expression Profiles in Myeloma: Ready for the Real World? Clin Cancer Res. 2016;22(22):5434-5442.

185. Bergsagel PL and Kuehl WM. Molecular pathogenesis and a consequent classification of multiple myeloma. J Clin Oncol. 2005;23(26):6333-6338.

186. Bergsagel PL, Kuehl WM, Zhan F, Sawyer J, Barlogie B, Shaughnessy J,Jr. Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma. Blood. 2005;106(1):296-303.

187. Broyl A, Hose D, Lokhorst H, et al. Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. Blood. 2010;116(14):2543-2553.

188. Anguiano A, Tuchman SA, Acharya C, et al. Gene expression profiles of tumor biology provide a novel approach to prognosis and may guide the selection of therapeutic targets in multiple myeloma. J Clin Oncol. 2009;27(25):4197-4203.

189. Lopez-Corral L, Corchete LA, Sarasquete ME, et al. Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. Haematologica. 2014;99(8):1365-1372.

190. Gulla A, Di Martino MT, Gallo Cantafio ME, et al. A 13 mer LNA-i-miR-221 Inhibitor Restores Drug Sensitivity in Melphalan-Refractory Multiple Myeloma Cells. Clin Cancer Res. 2016;22(5):1222-1233.

191. Zub KA, Sousa MM, Sarno A, et al. Modulation of cell metabolic pathways and oxidative stress signaling contribute to acquired melphalan resistance in multiple myeloma cells. PLoS One. 2015;10(3):e0119857.

192. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153(2):320-334.

193. Wu P, Walker BA, Broyl A, et al. A gene expression based predictor for high risk myeloma treated with intensive therapy and autologous stem cell rescue. Leuk Lymphoma. 2015;56(3):594-601.

194. Terragna C, Remondini D, Martello M, et al. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomib, thalidomide and dexamethasone (VTD). Oncotarget. 2016;7(9):9666-9679.

195. Terragna C, Renzulli M, Remondini D, et al. Correlation between eight-gene expression profiling and response to therapy of newly diagnosed multiple myeloma patients treated with thalidomide-dexamethasone incorporated into double autologous transplantation. Ann Hematol. 2013;92(9):1271-1280.

196. Zhan F, Barlogie B, Mulligan G, Shaughnessy JD,Jr, Bryant B. High-risk myeloma: a gene expression based risk-stratification model for newly diagnosed multiple myeloma treated with

high-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone. Blood. 2008;111(2):968-969.

197. Amin SB, Yip WK, Minvielle S, et al. Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. Leukemia. 2014;28(11):2229-2234.

198. Zhang X, Li B, Han H, et al. Predicting multi-level drug response with gene expression profile in multiple myeloma using hierarchical ordinal regression. BMC Cancer. 2018;18(1):551-018-4483-6.

199. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523-536.

200. Sanger F and Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441-448.

201. Maxam AM and Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560-564.

202. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-5467.

203. Smith LM, Sanders JZ, Kaiser RJ, et al. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321(6071):674-679.

204. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-380.

205. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science. 1998;281(5375):363, 365.

206. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387-402.

207. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220-2010-11-12-220. Epub 2010 Dec 22.

208. Macmanes MD. On the optimal trimming of high-throughput mRNA sequence data. Front Genet. 2014;5:13.

209. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 2016;17:103-016-0956-2.

210. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med. 2014;9:8-0473-9-8. eCollection 2014.

211. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.

212. Garg R, Patel RK, Tyagi AK, Jain M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res. 2011;18(1):53-63.

213. Mbandi SK, Hesse U, Rees DJ, Christoffels A. A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Front Genet. 2014;5:17.

214. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36-2013-14-4-r36.

215. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

216. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS One. 2013;8(10):e76935.

217. Grant GR, Farkas MH, Pizarro AD, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics. 2011;27(18):2518-2528.

218. Yang C, Wu PY, Phan JH, Wang MD. **The impact of RNA-seq alignment pipeline on detection of differentially expressed genes**. IEEE Global Conference on Signal and Information Processing (GlobalSIP). 2014.

219. Yang C, Wu PY, Tong L, Phan JH, Wang MD. The impact of RNA-seq aligners on gene expression estimation. ACM BCB. 2015;2015:462-471.

220. Lindner R and Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNA-seq. PLoS One. 2012;7(12):e52403.

221. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079.

222. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-515.

223. Roberts A and Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013;10(1):71-73.

224. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.

225. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290-295.

226. Li B and Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323-2105-12-323.

227. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-527.

228. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509-1517.

229. Bohnert R and Ratsch G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. Nucleic Acids Res. 2010;38(Web Server issue):W348-51.

230. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12):e131.

231. Srivastava S and Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. 2010;38(17):e170.

232. Maza E, Frasse P, Senin P, Bouzayen M, Zouine M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. Commun Integr Biol. 2013;6(6):e25849.

233. Wu PY, Phan JH, Zhou F, Wang MD. Evaluation of Normalization Methods for RNA-Seq Gene Expression Estimation. IEEE Int Conf Bioinform Biomed Workshops. 2011;2011:50-57.

234. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. 2015;16:347-015-0778-7.

235. Lin Y, Golovnina K, Chen ZX, et al. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC Genomics. 2016;17:28-015-2353-z.

236. Li X, Brock GN, Rouchka EC, et al. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. PLoS One. 2017;12(5):e0176185.

237. Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671-683.

238. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462-464.

239. Soneson C and Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91-2105-14-91.

240. Hardcastle TJ and Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422-2105-11-422.

241. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46-53.

242. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550-014-0550-8.

243. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.

244. Leng N, Dawson JA, Thomson JA, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8):1035-1043.

245. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Flexible isoform-level differential expression analysis with Ballgown. bioRxiv. 2014.

246. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

247. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21(12):2213-2223.

248. Li J and Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22(5):519-536.

249. Rapaport F, Khanin R, Liang Y, et al. Erratum to: Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2015;16:261-015-0813-z.

250. Guo Y, Li CI, Ye F, Shyr Y. Evaluation of read count based RNAseq analysis methods. BMC Genomics. 2013;14 Suppl 8:S2-2164-14-S8-S2. Epub 2013 Dec 9.

251. Zhang Z, Zhang Y, Evans P, Chinwalla A, Taylor D. **RNA-seq 2G: online analysis of differential gene expression with comprehensive options of statistical methods**. bioRxiv. 2017.

252. Zhou X and Robinson MD. Do count-based differential expression methods perform poorly when genes are expressed in only one condition? Genome Biol. 2015;16:222-015-0781-3.

253. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16(1):59-70.

254. Gao D, Kim J, Kim H, et al. A survey of statistical software for analysing RNA-seq data. Hum Genomics. 2010;5(1):56-60.

255. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12(12):e0190152.

256. Moulos P and Hatzis P. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. Nucleic Acids Res. 2015;43(4):e25.

257. Lyu Y and Li Q. A semi-parametric statistical model for integrating gene expression profiles across different platforms. BMC Bioinformatics. 2016;17 Suppl 1:5-015-0847-y.

258. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot. 2012;99(2):248-256.

259. Nookaew I, Papini M, Pornputtapong N, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. 2012;40(20):10084-10097.

260. Teng M, Love MI, Davis CA, et al. Erratum to: A benchmark for RNA-seq quantification pipelines. Genome Biol. 2016;17(1):203-016-1060-7.

261. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics. 2017;18(1):38-016-1457-z.

262. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13-016-0881-8.

263. Saif M and Parveen A. Computational Methods for prediction of miRNA with RNA-seq Analysis: Review. Journal of Applied Computing. 2016;1(1):2-13.

264. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.

265. Sinicropi D, Qu K, Collin F, et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS One. 2012;7(7):e40092.

266. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817-2826.

267. Atak ZK, Gianfelici V, Hulselmans G, et al. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. PLoS Genet. 2013;9(12):e1003997.

268. Ohguchi H, Harada T, Sagawa M, et al. KDM6B modulates MAPK pathway mediating multiple myeloma cell growth and survival. Leukemia. 2017;31(12):2661-2669.

269. Chapman MA, Sive J, Ambrose J, et al. RNA-seq of newly diagnosed patients in the PADIMAC study leads to a bortezomib/lenalidomide decision signature. Blood. 2018;132(20):2154-2165.

270. Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol. 2001;62(10):1311-1336.

271. Cooper CS, Campbell C, Jhavar S. Mechanisms of Disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer. Nat Clin Pract Urol. 2007;4(12):677-687.

272. Adib TR, Henderson S, Perrett C, et al. Predicting biomarkers for ovarian cancer using gene-expression microarrays. Br J Cancer. 2004;90(3):686-692.

273. Zhan F, Hardin J, Kordsmeier B, et al. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. Blood. 2002;99(5):1745-1757.

274. Binder H and Preibisch S. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. Biophys J. 2005;89(1):337-352.

275. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22-2011-12-3-r22. Epub 2011 Mar 16.

276. Sun Z, Asmann YW, Nair A, et al. Impact of library preparation on downstream analysis and interpretation of RNA-Seq data: comparison between Illumina PolyA and NuGEN Ovation protocol. PLoS One. 2013;8(8):e71745.

277. Grabowiecka E, Martin D, Crozier L, Holden N. Escherichia coli O157:H7 transcriptome datasets for comparison of RNA-seq and microarray platforms. Data Brief. 2018;22:126-131.

278. Wolff A, Bayerlova M, Gaedcke J, Kube D, Beissbarth T. A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells. PLoS One. 2018;13(5):e0197162.

279. Keck KJ, Breheny P, Braun TA, et al. Changes in gene expression in small bowel neuroendocrine tumors associated with progression to metastases. Surgery. 2018;163(1):232-239.

280. Romero JP, Ortiz-Estevez M, Muniategui A, et al. Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm. BMC Genomics. 2018;19(1):703-018-5082-2.

281. Chen L, Sun F, Yang X, et al. Correlation between RNA-Seq and microarrays results using TCGA data. Gene. 2017;628:200-204.

282. Nazarov PV, Muller A, Kaoma T, et al. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. BMC Genomics. 2017;18(1):443-017-3819-y.

283. Dapas M, Kandpal M, Bi Y, Davuluri RV. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. Brief Bioinform. 2017;18(2):260-269.

284. Li J, Hou R, Niu X, et al. Comparison of microarray and RNA-Seq analysis of mRNA expression in dermal mesenchymal stem cells. Biotechnol Lett. 2016;38(1):33-41.

285. Zhang Y, Akintola OS, Liu KJA, Sun B. Membrane gene ontology bias in sequencing and microarray obtained by housekeeping-gene analysis. Gene. 2016;575(2 Pt 2):559-566.

286. Yu J, Cliften PF, Juehne TI, et al. Multi-platform assessment of transcriptional profiling technologies utilizing a precise probe mapping methodology. BMC Genomics. 2015;16:710-015-1913-6.

287. Zhang W, Yu Y, Hertwig F, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol. 2015;16:133-015-0694-1.

288. Robinson DG, Wang JY, Storey JD. A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. Nucleic Acids Res. 2015;43(20):e131.

289. Nault R, Fader KA, Zacharewski T. RNA-Seq versus oligonucleotide array assessment of dose-dependent TCDD-elicited hepatic gene expression in mice. BMC Genomics. 2015;16:373-015-1527-z.

290. Zhang ZH, Jhaveri DJ, Marshall VM, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One. 2014;9(8):e103207.

291. Fumagalli D, Blanchet-Cohen A, Brown D, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. BMC Genomics. 2014;15:1008-2164-15-1008.

292. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;15:419-2164-15-419.

293. Perkins JR, Antunes-Martins A, Calvo M, et al. A comparison of RNA-seq and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat. Mol Pain. 2014;10:7-8069-10-7.

294. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.

295. Black MB, Parks BB, Pluta L, et al. Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments. Toxicol Sci. 2014;137(2):385-403.

296. Zwemer LM, Hui L, Wick HC, Bianchi DW. RNA-Seq and expression microarray highlight different aspects of the fetal amniotic fluid transcriptome. Prenat Diagn. 2014;34(10):1006-1014.

297. Xu X, Zhang Y, Williams J, et al. Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. BMC Bioinformatics. 2013;14 Suppl 9:S1-2105-14-S9-S1. Epub 2013 Jun 28.

298. Sekhon RS, Briskine R, Hirsch CN, et al. Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. PLoS One. 2013;8(4):e61005.

299. Mooney M, Bond J, Monks N, et al. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. PLoS One. 2013;8(4):e61088.

300. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana. Bioinformatics. 2013;29(6):717-724.

301. Raghavachari N, Barb J, Yang Y, et al. A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Med Genomics. 2012;5:28-8794-5-28.

302. Kogenaru S, Qing Y, Guo Y, Wang N. RNA-seq and microarray complement each other in transcriptome profiling. BMC Genomics. 2012;13:629-2164-13-629.

303. Sirbu A, Kerr G, Crane M, Ruskin HJ. RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. PLoS One. 2012;7(12):e50986.

304. van Delft J, Gaj S, Lienhard M, et al. RNA-Seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene. Toxicol Sci. 2012;130(2):427-439.

305. Bottomly D, Walter NA, Hunter JE, et al. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One. 2011;6(3):e17820.

306. Toung JM, Morley M, Li M, Cheung VG. RNA-sequence analysis of human B-cells. Genome Res. 2011;21(6):991-998.

307. Su Z, Li Z, Chen T, et al. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chem Res Toxicol. 2011;24(9):1486-1493.

308. Malone JH and Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011;9:34-7007-9-34.

309. Liu S, Lin L, Jiang P, Wang D, Xing Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res. 2011;39(2):578-588.

310. Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics. 2010;11:282-2164-11-282.

311. Griffith M, Griffith OL, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing. Nat Methods. 2010;7(10):843-847.

312. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94-2105-11-94.

313. Agarwal A, Koppstein D, Rozowsky J, et al. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics. 2010;11:383-2164-11-383.

314. Fu X, Fu N, Guo S, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics. 2009;10:161-2164-10-161.

315. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics. 2009;10:221-2164-10-221.

316. Anonymous Encyclopedia of genetics. San Diego, Calif. ; London: Academic; 2001.

317. Carrel A. On the Permanent Life of Tissues Outside of the Organism. J Exp Med. 1912;15(5):516-528.

318. Rodríguez-Hernández CO, Torres-García SE, Olvera-Sandoval C, et al. Cell Culture: History, Development and Prospects. Int J Curr Res Aca Rev. 2014;2(12):188-200.

319. Earle WR, Schilling EL, Stark TH, Straus NP, Brown MF, Shelton E. Production of Malignancy in Vitro. IV. The Mouse Fibroblast Cultures and Changes Seen in the Living Cells. JNCI: Journal of the National Cancer Institute. 1943;4(2):165-212.

320. Gey GO, Coffman WD, Kubicek MT. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res. 1952;12:264-265.

321. Kaur G and Dufour JM. Cell lines: Valuable tools or useless artifacts. Spermatogenesis. 2012;2(1):1-5.

322. Mitra A, Mishra L, Li S. Technologies for deriving primary tumor cells for use in personalized cancer therapy. Trends Biotechnol. 2013;31(6):347-354.

323. Masters JR. Human cancer cell lines: fact and fantasy. Nat Rev Mol Cell Biol. 2000;1(3):233-236.

324. Niu N and Wang L. In vitro human cell line models to predict clinical response to anticancer drugs. Pharmacogenomics. 2015;16(3):273-285.

325. Matsuoka Y, Moore GE, Yagi Y, Pressman D. Production of free light chains of immunoglobulin by a hematopoietic cell line derived from a patient with multiple myeloma. Proc Soc Exp Biol Med. 1967;125(4):1246-1250.

326. Hideshima T, Richardson P, Chauhan D, et al. The proteasome inhibitor PS-341 inhibits growth, induces apoptosis, and overcomes drug resistance in human multiple myeloma cells. Cancer Res. 2001;61(7):3071-3076.

327. Maiso P, Carvajal-Vergara X, Ocio EM, et al. The histone deacetylase inhibitor LBH589 is a potent antimyeloma agent that overcomes drug resistance. Cancer Res. 2006;66(11):5781-5789.

328. Hideshima T, Chauhan D, Shima Y, et al. Thalidomide and its analogs overcome drug resistance of human multiple myeloma cells to conventional therapy. Blood. 2000;96(9):2943-2950.

329. Ri M, Iida S, Nakashima T, et al. Bortezomib-resistant myeloma cell lines: a role for mutated PSMB5 in preventing the accumulation of unfolded proteins and fatal ER stress. Leukemia. 2010;24(8):1506-1512.

330. Zhu YX, Braggio E, Shi CX, et al. Cereblon expression is required for the antimyeloma activity of lenalidomide and pomalidomide. Blood. 2011;118(18):4771-4779.

331. Laubach J, Hideshima T, Richardson P, Anderson K. Clinical translation in multiple myeloma: from bench to bedside. Semin Oncol. 2013;40(5):549-553.

332. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175-182.

333. Khan K, Kunz R, Kleijnen J, Antes G. Systematic reviews to support evidence-based medicine, 2nd edition. CRC Press. 2011.

334. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. **Introduction to Meta-Analysis**. John Wiley & Sons, Ltd; 2009.

335. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1-12.

336. Viechtbauer W. **Conducting Meta-Analyses in R with the metafor Package**. J Stat Softw. 2010;36(3).

337. Bardsley WG. SIMFIT. A computer package for simulation, curve fitting and statistical analysis using life science models. In: Schuster S, Rigoulet M, Ouhabi R and Mazat JP, eds. Modern trends in Biothermokinetics. New York: Plenum Publishing Corporation; 1993:455-458.

338. Borenstein M, Hedges L, Higgins J, Rothstein H. Comprehensive Meta-Analysis Version 3. Biostat, Englewood, NJ. 2013.

339. Uman LS. Systematic reviews and meta-analyses. J Can Acad Child Adolesc Psychiatry. 2011;20(1):57-59.

340. O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. J R Soc Med. 2007;100(12):579-582.

341. Pearson K. Report on Certain Enteric Fever Inoculation Statistics. Br Med J. 1904;2(2288):1243-1246.

342. Cochran WG. The Combination of Estimates from Different Experiments. Biometrics. 1954;10(1):101-129.

343. DerSimonian R and Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-188.

344. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-634.

345. Higgins JP and Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-1558.

346. Esterhuizen TM and Thabane L. Con: Meta-analysis: some key limitations and potential solutions. Nephrol Dial Transplant. 2016;31(6):882-885.

347. Qadir XV, Clyne M, Lam TK, Khoury MJ, Schully SD. Trends in published meta-analyses in cancer research, 2008-2013. Cancer Causes Control. 2017;28(1):5-12.

348. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. 2002;62(15):4427-4433.

349. Leng D, Miao R, Huang X, Wang Y. In silico analysis identifies CRISP3 as a potential peripheral blood biomarker for multiple myeloma: From data modeling to validation with RT-PCR. Oncol Lett. 2018;15(4):5167-5174.

350. Dayde D, Tanaka I, Jain R, Tai MC, Taguchi A. Predictive and Prognostic Molecular Biomarkers for Response to Neoadjuvant Chemoradiation in Rectal Cancer. Int J Mol Sci. 2017;18(3):10.3390/ijms18030573.

351. Dalton WS and Friend SH. Cancer biomarkers--an invitation to the table. Science. 2006;312(5777):1165-1168.

352. Nimse SB, Sonawane MD, Song KS, Kim T. Biomarker detection technologies and future directions. Analyst. 2016;141(3):740-755.

353. Henry NL and Hayes DF. Cancer biomarkers. Mol Oncol. 2012;6(2):140-146.

354. Kamel HFM and Al-Amodi HSAB. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics Proteomics Bioinformatics. 2017;15(4):220-235.

355. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-536.

356. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671-679.

357. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A. 2001;98(20):11462-11467.

358. Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011;29(1):17-24.

359. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002;1(2):203-209.

360. Reme T, Hose D, De Vos J, et al. A new method for class prediction based on signed-rank algorithms applied to Affymetrix microarray experiments. BMC Bioinformatics. 2008;9:16-2105-9-16.

361. Zhou Y, Zhang Q, Stephens O, et al. Prediction of cytogenetic abnormalities with gene expression profiles. Blood. 2012;119(21):e148-50.

362. Biju SM. Analyzing the predictive capacity of various machine learning algorithms. International Journal of Engineering & Technology. 2017;5.

363. Vapnik V. **Estimation of Dependences Based on Empirical Data: Springer Series in Statistics** . Berlin, Heidelberg: Springer-Verlag; 1982.

364. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician. 1992;46(3):175-185.

365. Breiman L. Random Forests. Mach Learn. 2001;45(1):5-32.

366. Burguillo FJ, Corchete LA, Martín J, Barrera I, Bardsley WG. A Partial Least Squares Algorithm for Microarray Data Analysis Using the VIP Statistic for Gene Selection and Binary Classification. Curr Bioinform. 2014;9(3):348-359.

367. Pedregosa F, Varoquaux G, Gramfort A, et al. **Scikit-learn: Machine Learning in Python**. J Mach Learn Res. 2011;12:2825-2830.

368. Andrews S. *FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc*. 2010.

369. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120.

370. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10-12.

371. Bushnell B. BBMap, sourceforge.net/projects/bbmap/. 2015.

372. Dinno A. dunn.test: Dunn's test of multiple comparisons using Rank Sums. 2017;1.3.5:https://CRAN.R-project.org/package=dunn.test.

373. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018:https://www.R-project.org/.

374. Aken BL, Achuthan P, Akanni W, et al. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635-D642.

375. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357-360.

376. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-359.

377. Zyprych-Walczak J, Szabelska A, Handschuh L, et al. he Impact of Normalization Methods on RNA-Seq Data Analysis. BioMed Research International. 2015;2015(621690).

378. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-628.

379. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;26(4):493-500.

380. Robinson MD and Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25-2010-11-3-r25. Epub 2010 Mar 2.

381. Maza E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Front Genet. 2016;7:164.

382. Reddy R. A Comparison of Methods: Normalizing High-Throughput RNA Sequencing Data. bioRxiv. 2015.

383. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-419.

384. Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics. 2012;28(14):1838-1844.

385. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29-2014-15-2-r29.

386. Anders S and Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106-2010-11-10-r106. Epub 2010 Oct 27.

387. Gierlinski M, Cole C, Schofield P, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31(22):3625-3630.

388. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31(22):3718-3720.

389. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-264.

390. Irizarry RA, Ooi SL, Wu Z, Boeke JD. Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants. Stat Appl Genet Mol Biol. 2003;2:Article1-6115.1002. Epub 2003 Mar 18.

391. Carvalho BS and Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26(19):2363-2367.

392. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33(20):e175.

393. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-315.

394. Chrominski K and Tkacz M. Comparison of High-Level Microarray Analysis Methods in the Context of Result Consistency. PLoS One. 2015;10(6):e0128845.

395. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116-5121.

396. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3-6115.1027. Epub 2004 Feb 12.

397. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. Biotechnol Lett. 2004;26(6):509-515.

398. Andersen CL, Jensen JL, Orntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 2004;64(15):5245-5250.

399. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3(7):RESEARCH0034.

400. Silver N, Best S, Jiang J, Thein SL. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. BMC Mol Biol. 2006;7:33-2199-7-33.

401. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.

402. Everaert C, Luypaert M, Maag JLV, et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep. 2017;7(1):1559-017-01617-3.

403. Filzmoser P and Gschwandtner M. mvoutlier: Multivariate outlier detection based on robust methods. 2018;2.0.9:https://CRAN.R-project.org/package=mvoutlier.

404. Gavaghan DJ, Moore RA, McQuay HJ. An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain. 2000;85(3):415-424.

405. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.

406. Clarke E and Sherrill-Mix S. ggbeeswarm: categorical scatter (violin point) plots. 2017;0.6.0:https://CRAN.R-project.org/package=ggbeeswarm.

407. Oommen T, Misra D, Twarakavi NKC, Prakash A, Sahoo B, Bandopadhyay S. An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing. Math Geosci. 2008;40(4):409-424.

408. Kursa MB and Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. J Stat Softw. 2010;36(11):1-13.

409. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. 2018;1.7-0:https://CRAN.R-project.org/package=e1071.

410. Pérez-Martín A, Pérez-Torregrosa A, Vaca-Lamata M, Verdú-Jover AJ. OptimClassifier: create the best train for classification models. 2018;0.1.4:https://CRAN.R-project.org/package=OptimClassifier.

411. Wold H. Nonlinear estimation by iterative least squates procedures. In: David FN, ed. Research papers in Statistics. Festschrift for J. Neyman. New York: Wiley; 1966:411-444.

412. Kuhn M, Wing J, Weston S, et al. caret: Classification and Regression Training. R package version. 2018;6.0-80:https://CRAN.R-project.org/package=caret.

413. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 2005;33(Web Server issue):W741-8.

414. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013;41(Web Server issue):W77-83.

415. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017;45(W1):W130-W137.

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

417. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62.

418. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D361.

419. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-29.

420. Luo W and Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830-1831.

421. Luo W, Pant G, Bhavnasi YK, Blanchard SG,Jr, Brouwer C. Pathview Web: user friendly pathway visualization and data integration. Nucleic Acids Res. 2017;45(W1):W501-W508.

422. Robert C and Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16:177-015-0734-x.

423. Paul S, Arlehamn CSL, Schulten V, et al. Experimental validation of the RATE tool for inferring HLA restrictions of T cell epitopes. BMC Immunol. 2017;18(Suppl 1):20-017-0204-1.

424. Marchesini M, Ogoti Y, Fiorini E, et al. ILF2 Is a Regulator of RNA Splicing and DNA Damage Response in 1q21-Amplified Multiple Myeloma. Cancer Cell. 2017;32(1):88-100.e6.

425. Gomez-Bougie P, Oliver L, Le Gouill S, Bataille R, Amiot M. Melphalan-induced apoptosis in multiple myeloma cells is associated with a cleavage of Mcl-1 and Bim and a decrease in the Mcl-1/Bim complex. Oncogene. 2005;24(54):8076-8079.

426. Mateyak MK, Obaya AJ, Sedivy JM. c-Myc regulates cyclin D-Cdk4 and -Cdk6 activity but affects cell cycle progression at multiple independent points. Mol Cell Biol. 1999;19(7):4672-4683.

427. Tigan AS, Bellutti F, Kollmann K, Tebb G, Sexl V. CDK6-a review of the past and a glimpse into the future: from cell-cycle control to transcriptional regulation. Oncogene. 2016;35(24):3083-3091.

428. Staller P, Peukert K, Kiermaier A, et al. Repression of p15INK4b expression by Myc through association with Miz-1. Nat Cell Biol. 2001;3(4):392-399.

429. Tadesse S, Yu M, Kumarasiri M, Le BT, Wang S. Targeting CDK6 in cancer: State of the art and new insights. Cell Cycle. 2015;14(20):3220-3230.

430. Stolz A, Ertych N, Bastians H. Tumor suppressor CHK2: regulator of DNA damage response and mediator of chromosomal stability. Clin Cancer Res. 2011;17(3):401-405.

431. Kulikov R, Letienne J, Kaur M, Grossman SR, Arts J, Blattner C. Mdm2 facilitates the association of p53 with the proteasome. Proc Natl Acad Sci U S A. 2010;107(22):10038-10043.

432. Secchiero P, Barbarotto E, Tiribelli M, et al. Functional integrity of the p53-mediated apoptotic pathway induced by the nongenotoxic agent nutlin-3 in B-cell chronic lymphocytic leukemia (B-CLL). Blood. 2006;107(10):4122-4129.

433. Li Y, Jenkins CW, Nichols MA, Xiong Y. Cell cycle expression and p53 regulation of the cyclin-dependent kinase inhibitor p21. Oncogene. 1994;9(8):2261-2268.

434. Lee CK, Wang S, Huang X, Ryder J, Liu B. HDAC inhibition synergistically enhances alkylator-induced DNA damage responses and apoptosis in multiple myeloma cells. Cancer Lett. 2010;296(2):233-240.

435. Tamura RE, de Vasconcellos JF, Sarkar D, Libermann TA, Fisher PB, Zerbini LF. GADD45 proteins: central players in tumorigenesis. Curr Mol Med. 2012;12(5):634-651.

436. Nurse P. The central role of a CDK in controlling the fission yeast cell cycle. Harvey Lect. 1996;92:55-64.

437. Milella M, Falcone I, Conciatori F, et al. PTEN: Multiple Functions in Human Malignant Tumors. Front Oncol. 2015;5:24.

438. Wee S, Wiederschain D, Maira SM, et al. PTEN-deficient cancers depend on PIK3CB. Proc Natl Acad Sci U S A. 2008;105(35):13057-13062.

439. Burwick N, Zhang MY, de la Puente P, et al. The eIF2-alpha kinase HRI is a novel therapeutic target in multiple myeloma. Leuk Res. 2017;55:23-32.

440. Thomas AL, Coarfa C, Qian J, et al. Identification of potential glucocorticoid receptor therapeutic targets in multiple myeloma. Nucl Recept Signal. 2015;13:e006.

441. Palagani A, Op de Beeck K, Naulaerts S, et al. Ectopic microRNA-150-5p transcription sensitizes glucocorticoid therapy response in MM1S multiple myeloma cells but fails to overcome hormone therapy resistance in MM1R cells. PLoS One. 2014;9(12):e113842.

442. Rickles RJ, Tam WF, Giordano TP,3rd, et al. Adenosine A2A and beta-2 adrenergic receptor agonists: novel selective and synergistic multiple myeloma targets discovered through systematic combination screening. Mol Cancer Ther. 2012;11(7):1432-1442.

443. Nojima M, Maruyama R, Yasui H, et al. Genomic screening for genes silenced by DNA methylation revealed an association between RASD1 inactivation and dexamethasone resistance in multiple myeloma. Clin Cancer Res. 2009;15(13):4356-4364.

444. Kim SW, Kim HY, Lee HJ, Yun HJ, Kim S, Jo DY. Dexamethasone and hypoxia upregulate CXCR4 expression in myeloma cells. Leuk Lymphoma. 2009;50(7):1163-1173.

445. Chauhan D, Auclair D, Robinson EK, et al. Identification of genes regulated by dexamethasone in multiple myeloma cells using oligonucleotide arrays. Oncogene. 2002;21(9):1346-1358.

446. Barash U, Zohar Y, Wildbaum G, et al. Heparanase enhances myeloma progression via CXCL10 downregulation. Leukemia. 2014;28(11):2178-2187.

447. Vallet S and Anderson KC. CCR1 as a target for multiple myeloma. Expert Opin Ther Targets. 2011;15(9):1037-1047.

448. Dairaghi DJ, Oyajobi BO, Gupta A, et al. CCR1 blockade reduces tumor burden and osteolysis in vivo in a mouse model of myeloma bone disease. Blood. 2012;120(7):1449-1457.

449. Starheim KK, Holien T, Misund K, et al. Intracellular glutathione determines bortezomib cytotoxicity in multiple myeloma cells. Blood Cancer J. 2016;6(7):e446.

450. Fristedt Duvefelt C, Lub S, Agarwal P, et al. Increased resistance to proteasome inhibitors in multiple myeloma mediated by cIAP2--implications for a combinatorial treatment. Oncotarget. 2015;6(24):20621-20635.

451. Wiita AP, Ziv E, Wiita PJ, et al. Global cellular response to chemotherapy-induced apoptosis. Elife. 2013;2:e01236.

452. Stessman HA, Baughn LB, Sarver A, et al. Profiling bortezomib resistance identifies secondary therapies in a mouse myeloma model. Mol Cancer Ther. 2013;12(6):1140-1150.

453. Tomasella A, Picco R, Ciotti S, et al. The isopeptidase inhibitor 2cPE triggers proteotoxic stress and ATM activation in chronic lymphocytic leukemia cells. Oncotarget. 2016;7(29):45429-45443.

454. Mitsiades N, Mitsiades CS, Poulaki V, et al. Molecular sequelae of proteasome inhibition in human multiple myeloma cells. Proc Natl Acad Sci U S A. 2002;99(22):14374-14379.

455. Oerlemans R, Franke NE, Assaraf YG, et al. Molecular basis of bortezomib resistance: proteasome subunit beta5 (PSMB5) gene mutation and overexpression of PSMB5 protein. Blood. 2008;112(6):2489-2499.

456. Ohoka N, Yoshii S, Hattori T, Onozaki K, Hayashi H. TRB3, a novel ER stress-inducible gene, is induced via ATF4-CHOP pathway and is involved in cell death. EMBO J. 2005;24(6):1243-1255.

457. Slaby J, Tagoug I, Neri P, et al. Lenalidomide Induces A Ribosomal Stress Response In Multiple Myeloma (MM) Cells. Blood. 2013;122(21):3161.

458. van Riggelen J, Yetil A, Felsher DW. MYC as a regulator of ribosome biogenesis and protein synthesis. Nat Rev Cancer. 2010;10(4):301-309.

459. Maggi LB,Jr, Kuchenruether M, Dadey DY, et al. Nucleophosmin serves as a rate-limiting nuclear export chaperone for the Mammalian ribosome. Mol Cell Biol. 2008;28(23):7050-7065.

460. Garrido F, Aptsiauri N, Doorduijn EM, Garcia Lora AM, van Hall T. The urgent need to recover MHC class I in cancers for effective immunotherapy. Curr Opin Immunol. 2016;39:44-51.

461. Zhao M, Flynt FL, Hong M, et al. MHC class II transactivator (CIITA) expression is upregulated in multiple myeloma cells by IFN-gamma. Mol Immunol. 2007;44(11):2923-2932.

462. Bartlett JB, Dredge K, Dalgleish AG. The evolution of thalidomide and its IMiD derivatives as anticancer agents. Nat Rev Cancer. 2004;4(4):314-322.

463. Chanan-Khan AA, Swaika A, Paulus A, et al. Pomalidomide: the new immunomodulatory agent for the treatment of multiple myeloma. Blood Cancer J. 2013;3:e143.

464. Henney CS. Interleukin 7: effects on early events in lymphopoiesis. Immunol Today. 1989;10(5):170-173.

465. Davies FE, Raje N, Hideshima T, et al. Thalidomide and immunomodulatory derivatives augment natural killer cell cytotoxicity in multiple myeloma. Blood. 2001;98(1):210-216.

466. Matthews GM, Lefebure M, Doyle MA, et al. Preclinical screening of histone deacetylase inhibitors combined with ABT-737, rhTRAIL/MD5-1 or 5-azacytidine using syngeneic Vk*MYC multiple myeloma. Cell Death Dis. 2013;4:e798.

467. Lemaire M, Fristedt C, Agarwal P, et al. The HDAC inhibitor LBH589 enhances the antimyeloma effects of the IGF-1RTK inhibitor picropodophyllin. Clin Cancer Res. 2012;18(8):2230-2239.

468. Ma Y, Liu W, Zhang L, Jia G. Effects of Histone Deacetylase Inhibitor Panobinostat (LBH589) on Bone Marrow Mononuclear Cells of Relapsed or Refractory Multiple Myeloma Patients and Its Mechanisms. Med Sci Monit. 2017;23:5150-5157.

469. Laubach JP, Moreau P, San-Miguel JF, Richardson PG. Panobinostat for the Treatment of Multiple Myeloma. Clin Cancer Res. 2015;21(21):4767-4773.

470. Bailey H, Stenehjem DD, Sharma S. Panobinostat for the treatment of multiple myeloma: the evidence to date. J Blood Med. 2015;6:269-276.

471. Wilson AJ, Sarfo-Kantanka K, Barrack T, et al. Panobinostat sensitizes cyclin E high, homologous recombination-proficient ovarian cancer to olaparib. Gynecol Oncol. 2016;143(1):143-151.

472. Di Fazio P, Schneider-Stock R, Neureiter D, et al. The pan-deacetylase inhibitor panobinostat inhibits growth of hepatocellular carcinoma models by alternative pathways of apoptosis. Cell Oncol. 2010;32(4):285-300.

473. Neri P, Bahlis NJ, Lonial S. Panobinostat for the treatment of multiple myeloma. Expert Opin Investig Drugs. 2012;21(5):733-747.

474. Mithraprabhu S, Khong T, Spencer A. Overcoming inherent resistance to histone deacetylase inhibitors in multiple myeloma cells by targeting pathways integral to the actin cytoskeleton. Cell Death Dis. 2014;5:e1134.

475. Prystowsky MB, Adomako A, Smith RV, et al. The histone deacetylase inhibitor LBH589 inhibits expression of mitotic genes causing G2/M arrest and cell death in head and neck squamous cell carcinoma cell lines. J Pathol. 2009;218(4):467-477.

476. Abbas T and Dutta A. P21 in Cancer: Intricate Networks and Multiple Activities. Nat Rev Cancer. 2009;9(6):400-414.

477. Prystowsky M, Feeney K, Kawachi N, et al. Inhibition of Plk1 and Cyclin B1 expression results in panobinostat-induced G(2) delay and mitotic defects. Sci Rep. 2013;3:2640.

478. Lin YC, Sun SH, Wang FF. Suppression of Polo like kinase 1 (PLK1) by p21(Waf1) mediates the p53-dependent prevention of caspase-independent mitotic death. Cell Signal. 2011;23(11):1816-1823.

479. Tategu M, Nakagawa H, Sasaki K, et al. Transcriptional regulation of human polo-like kinases and early mitotic inhibitor. J Genet Genomics. 2008;35(4):215-224.

480. Liu X and Erikson RL. Polo-like kinase (Plk)1 depletion induces apoptosis in cancer cells. Proc Natl Acad Sci U S A. 2003;100(10):5789-5794.

481. Spankuch B, Steinhauser I, Wartlick H, Kurunci-Csacsko E, Strebhardt KI, Langer K. Downregulation of Plk1 expression by receptor-mediated uptake of antisense oligonucleotide-loaded nanoparticles. Neoplasia. 2008;10(3):223-234.

482. Poirier S, Samami S, Mamarbachi M, et al. The epigenetic drug 5-azacytidine interferes with cholesterol and lipid metabolism. J Biol Chem. 2014;289(27):18736-18751.

483. Zhang F, Dai X, Wang Y. 5-Aza-2'-deoxycytidine induced growth inhibition of leukemia cells through modulating endogenous cholesterol biosynthesis. Mol Cell Proteomics. 2012;11(7):M111.016915.

484. Madden EA, Bishop EJ, Fiskin AM, Melnykovych G. Possible role of cholesterol in the susceptibility of a human acute lymphoblastic leukemia cell line to dexamethasone. Cancer Res. 1986;46(2):617-622.

485. Dang H, Liu Y, Pang W, et al. Suppression of 2,3-oxidosqualene cyclase by high fat diet contributes to liver X receptor-alpha-mediated improvement of hepatic lipid profile. J Biol Chem. 2009;284(10):6218-6226.

486. Brusselmans K, Timmermans L, Van de Sande T, et al. Squalene synthase, a determinant of Raft-associated cholesterol and modulator of cancer cell proliferation. J Biol Chem. 2007;282(26):18777-18785.

487. Tuzmen S, Hostetter G, Watanabe A, et al. Characterization of farnesyl diphosphate farnesyl transferase 1 (FDFT1) expression in cancer. Per Med. 2019;16(1):51-65.

488. Xu CZ, Shi RJ, Chen D, et al. Potential biomarkers for paclitaxel sensitivity in hypopharynx cancer cell. Int J Clin Exp Pathol. 2013;6(12):2745-2756.

489. Shi H, Guo J, Duff DJ, et al. Discovery of novel epigenetic markers in non-Hodgkin's lymphoma. Carcinogenesis. 2007;28(1):60-70.

490. Shi H, Yan PS, Chen CM, et al. Expressed CpG island sequence tag microarray for dual screening of DNA hypermethylation and gene silencing in cancer cells. Cancer Res. 2002;62(11):3214-3220.

491. Brozyna AA, Jozwicki W, Jochymski C, Slominski AT. Decreased expression of CYP27B1 correlates with the increased aggressiveness of ovarian carcinomas. Oncol Rep. 2015;33(2):599-606.

492. Deeb KK, Trump DL, Johnson CS. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. Nat Rev Cancer. 2007;7(9):684-700.

493. Feldman D, Krishnan AV, Swami S, Giovannucci E, Feldman BJ. The role of vitamin D in reducing cancer risk and progression. Nat Rev Cancer. 2014;14(5):342-357.

494. Panda DK, Miao D, Tremblay ML, et al. Targeted ablation of the 25-hydroxyvitamin D 1alpha -hydroxylase enzyme: evidence for skeletal, reproductive, and immune dysfunction. Proc Natl Acad Sci U S A. 2001;98(13):7498-7503.

495. Heller G, Schmidt WM, Ziegler B, et al. Genome-wide transcriptional response to 5-aza-2'-deoxycytidine and trichostatin a in multiple myeloma cells. Cancer Res. 2008;68(1):44-54.

496. Yang MY, Liu TC, Chang JG, Lin PM, Lin SF. JunB gene expression is inactivated by methylation in chronic myeloid leukemia. Blood. 2003;101(8):3205-3211.

497. Ott RG, Simma O, Kollmann K, et al. JunB is a gatekeeper for B-lymphoid leukemia. Oncogene. 2007;26(33):4863-4871.

498. Mathas S, Hinz M, Anagnostopoulos I, et al. Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma cells, stimulate proliferation and synergize with NF-kappa B. EMBO J. 2002;21(15):4104-4113.

499. Fan F, Bashari MH, Morelli E, et al. The AP-1 transcription factor JunB is essential for multiple myeloma cell proliferation and drug resistance in the bone marrow microenvironment. Leukemia. 2017;31(7):1570-1581.

500. Boyce BF, Xiu Y, Li J, Xing L, Yao Z. NF-kappaB-Mediated Regulation of Osteoclastogenesis. Endocrinol Metab (Seoul). 2015;30(1):35-44.

501. Vrabel D, Pour L, Sevcikova S. The impact of NF-kappaB signaling on pathogenesis and current treatment strategies in multiple myeloma. Blood Rev. 2019;34:56-66.

502. Guan H, Mi B, Li Y, et al. Decitabine represses osteoclastogenesis through inhibition of RANK and NF-kappaB. Cell Signal. 2015;27(5):969-977.

503. Yilmaz ZB, Weih DS, Sivakumar V, Weih F. RelB is required for Peyer's patch development: differential regulation of p52-RelB by lymphotoxin and TNF. EMBO J. 2003;22(1):121-130.

504. Liu X, Wang B, Ma X, Guo Y. NF-kappaB activation through the alternative pathway correlates with chemoresistance and poor survival in extranodal NK/T-cell lymphoma, nasal type. Jpn J Clin Oncol. 2009;39(7):418-424.

505. Li X, Zhang Y, Chen M, et al. Increased IFNgamma(+) T Cells Are Responsible for the Clinical Responses of Low-Dose DNA-Demethylating Agent Decitabine Antitumor Therapy. Clin Cancer Res. 2017;23(20):6031-6043.

506. Palumbo A, Bruno B, Boccadoro M, Pileri A. Interferon-gamma in multiple myeloma. Leuk Lymphoma. 1995;18(3-4):215-219.

507. Zhu J and Paul WE. CD4 T cells: fates, functions, and faults. Blood. 2008;112(5):1557-1569.

508. Cimino G, Avvisati G, Amadori S, et al. High serum IL-2 levels are predictive of prolonged survival in multiple myeloma. Br J Haematol. 1990;75(3):373-377.

509. Hogg SJ, Vervoort SJ, Deswal S, et al. BET-Bromodomain Inhibitors Engage the Host Immune System and Regulate Expression of the Immune Checkpoint Ligand PD-L1. Cell Rep. 2017;18(9):2162-2174.

510. Manier S, Huynh D, Shen YJ, et al. Inhibiting the oncogenic translation program is an effective therapeutic strategy in multiple myeloma. Sci Transl Med. 2017;9(389):10.1126/scitranslmed.aal2668.

511. Munoz-Pinedo C, El Mjiyad N, Ricci JE. Cancer metabolism: current perspectives and future directions. Cell Death Dis. 2012;3:e248.

512. El Arfani C, De Veirman K, Maes K, De Bruyne E, Menu E. Metabolic Features of Multiple Myeloma. Int J Mol Sci. 2018;19(4):10.3390/ijms19041200.

513. Maiso P, Huynh D, Moschetta M, et al. Metabolic signature identifies novel targets for drug resistance in multiple myeloma. Cancer Res. 2015;75(10):2071-2082.

514. Lubin M, Cahn F, Coutermarsh BA. Amiloride, protein synthesis, and activation of quiescent cells. J Cell Physiol. 1982;113(2):247-251.

515. Taub M and Saier MH,Jr. Amiloride-resistant Madin-Darby canine kidney (MDCK) cells exhibit decreased cation transport. J Cell Physiol. 1981;106(2):191-199.

516. Pathria G, Scott DA, Feng Y, et al. Targeting the Warburg effect via LDHA inhibition engages ATF4 signaling for cancer cell survival. EMBO J. 2018;37(20):10.15252/embj.201899735. Epub 2018 Sep 12.

517. Miao P, Sheng S, Sun X, Liu J, Huang G. Lactate dehydrogenase A in cancer: a promising target for diagnosis and therapy. IUBMB Life. 2013;65(11):904-910.

518. Escalante CR, Yie J, Thanos D, Aggarwal AK. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. Nature. 1998;391(6662):103-106.

519. Ramana CV, Gil MP, Schreiber RD, Stark GR. Stat1-dependent and -independent pathways in IFN-gamma-dependent signaling. Trends Immunol. 2002;23(2):96-101.

520. Meissl K, Macho-Maschler S, Muller M, Strobl B. The good and the bad faces of STAT1 in solid tumours. Cytokine. 2017;89:12-20.

521. Yu H and Jove R. The STATs of cancer--new molecular targets come of age. Nat Rev Cancer. 2004;4(2):97-105.

522. Pensa S, Regis G, Boselli D, Novelli F, Poli V. STAT1 and STAT3 in Tumorigenesis: Two Sides of the Same Coin? In: Anonymous Madame Curie Bioscience Database. Austin (TX): Landes Bioscience; 2013:https://www.ncbi.nlm.nih.gov/books/NBK6568/.

523. Zhao C, Li H, Lin HJ, Yang S, Lin J, Liang G. Feedback Activation of STAT3 as a Cancer Drug-Resistance Mechanism. Trends Pharmacol Sci. 2016;37(1):47-61.

524. Gorgun G, Calabrese E, Soydan E, et al. Immunomodulatory effects of lenalidomide and pomalidomide on interaction of tumor and bone marrow accessory cells in multiple myeloma. Blood. 2010;116(17):3227-3237.

525. Flotho C, Claus R, Batz C, et al. The DNA methyltransferase inhibitors azacitidine, decitabine and zebularine exert differential effects on cancer gene expression in acute myeloid leukemia cells. Leukemia. 2009;23(6):1019-1028.

526. Xie M, Jiang Q, Xie Y. Comparison between decitabine and azacitidine for the treatment of myelodysplastic syndrome: a meta-analysis with 1,392 participants. Clin Lymphoma Myeloma Leuk. 2015;15(1):22-28.

527. Stretch C, Khan S, Asgarian N, et al. Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. PLoS One. 2013;8(6):e65380.

528. Mulligan G, Mitsiades C, Bryant B, et al. Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. Blood. 2007;109(8):3177-3188.

529. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009;27(12):1135-1137.

530. Zhang T, Wang R, Wang Z, Wang X, Wang F, Ding J. Structural basis for Ragulator functioning as a scaffold in membrane-anchoring of Rag GTPases and mTORC1. Nat Commun. 2017;8(1):1394-017-01567-4.

531. Bar-Peled L, Schweitzer LD, Zoncu R, Sabatini DM. Ragulator is a GEF for the rag GTPases that signal amino acid levels to mTORC1. Cell. 2012;150(6):1196-1208.

532. Csibi A, Cornille K, Leibovitch MP, et al. The translation regulatory subunit eIF3f controls the kinase-dependent mTOR signaling required for muscle differentiation and hypertrophy in mouse. PLoS One. 2010;5(2):e8994.

533. Kerr KF. Comments on the analysis of unbalanced microarray data. Bioinformatics. 2009;25(16):2035-2041.

534. Barlund M, Monni O, Weaver JD, et al. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes Cancer. 2002;35(4):311-317.

535. Lee ST, Feng M, Wei Y, et al. Protein tyrosine phosphatase UBASH3B is overexpressed in triple-negative breast cancer and promotes invasion and metastasis. Proc Natl Acad Sci U S A. 2013;110(27):11121-11126.

536. Lai RH, Hsiao YW, Wang MJ, et al. SOCS6, down-regulated in gastric cancer, inhibits cell proliferation and colony formation. Cancer Lett. 2010;288(1):75-85.

537. Carroll RG and Martin SJ. Autophagy in multiple myeloma: what makes you stronger can also kill you. Cancer Cell. 2013;23(4):425-426.

538. Anonymous SPRED1 Is a Tumor Suppressor in Mucosal Melanoma. Cancer Discov. 2018;8(12):1507-8290.CD-RW2018-196. Epub 2018 Nov 9.

539. Das DS, Das A, Ray A, et al. Blockade of Deubiquitylating Enzyme USP1 Inhibits DNA Repair and Triggers Apoptosis in Multiple Myeloma Cells. Clin Cancer Res. 2017;23(15):4280-4289.

540. Biggs PJ, Wooster R, Ford D, et al. Familial cylindromatosis (turban tumour syndrome) gene localised to chromosome 16q12-q13: evidence for its role as a tumour suppressor gene. Nat Genet. 1995;11(4):441-443.

541. van Andel H, Kocemba KA, de Haan-Kramer A, et al. Loss of CYLD expression unleashes Wnt signaling in multiple myeloma and is associated with aggressive disease. Oncogene. 2017;36(15):2105-2115.

542. Markovina S, Callander NS, O'Connor SL, et al. Bortezomib-resistant nuclear factor-kappaB activity in multiple myeloma cells. Mol Cancer Res. 2008;6(8):1356-1364.

543. Emmons MF, Anreddy N, Cuevas J, et al. MTI-101 treatment inducing activation of Stim1 and TRPC1 expression is a determinant of response in multiple myeloma. Sci Rep. 2017;7(1):2685-017-02713-0.

544. Gutierrez NC, Sarasquete ME, Misiewicz-Krzeminska I, et al. Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. Leukemia. 2010;24(3):629-637.

545. Jones RJ, Gu D, Bjorklund CC, et al. The novel anticancer agent JNJ-26854165 induces cell death through inhibition of cholesterol transport and degradation of ABCA1. J Pharmacol Exp Ther. 2013;346(3):381-392.

546. Yang N, Chen J, Zhang H, et al. LncRNA OIP5-AS1 loss-induced microRNA-410 accumulation regulates cell proliferation and apoptosis by targeting KLF10 via activating PTEN/PI3K/AKT pathway in multiple myeloma. Cell Death Dis. 2017;8(8):e2975.

547. Wei XF, Chen QL, Fu Y, Zhang QK. Wnt and BMP signaling pathways co-operatively induce the differentiation of multiple myeloma mesenchymal stem cells into osteoblasts by upregulating EMX2. J Cell Biochem. 2019;120(4):6515-6527.

548. Lara R, Seckl MJ, Pardo OE. The p90 RSK family members: common functions and isoform specificity. Cancer Res. 2013;73(17):5301-5308.

549. Salhi A, Farhadian JA, Giles KM, et al. RSK1 activation promotes invasion in nodular melanoma. Am J Pathol. 2015;185(3):704-716.

550. Lara R, Mauri FA, Taylor H, et al. An siRNA screen identifies RSK1 as a key modulator of lung cancer metastasis. Oncogene. 2011;30(32):3513-3521.

551. Abramson HN. Kinase inhibitors as potential agents in the treatment of multiple myeloma. Oncotarget. 2016;7(49):81926-81968.

552. Hoepfner S, Severin F, Cabezas A, et al. Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. Cell. 2005;121(3):437-450.

553. Stenmark H and Olkkonen VM. The Rab GTPase family. Genome Biol. 2001;2(5):REVIEWS3007.

554. Zhu YX, Yin H, Bruins LA, et al. RNA interference screening identifies lenalidomide sensitizers in multiple myeloma, including RSK2. Blood. 2015;125(3):483-491.