Universidad de Valladolid

Facultad de Ciencias

Departamento de Estadística e Investigación Operativa

TESIS DOCTORAL:

Desarrollo y aplicación de métodos estadísticos basados en recortes imparciales a datos de expresión génica de alta dimensionalidad

Presentada por Itziar Fernández Martínez para optar al grado de doctor por la Universidad de Valladolid

Dirigida por:

Dr. Javier de las Rivas Sanz y Dr. Agustín Mayo Íscar

Junio 2012

Universidad de Valladolid

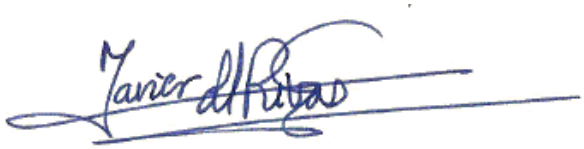
Impreso 2T

AUTORIZACIÓN DEL DIRECTOR DE TESIS
(Art. 21 del R.D. 1393/2007 de 29 de octubre y Art. 4 c) de la Normativa para la defensa de la Tesis Doctoral)

D. **Javier DE LAS RIVAS SANZ**, con D.N.I. nº **15949000H** profesor/investigador del departamento de **Bioinformática y Genómica Funcional del Centro de Investigación del Cáncer** (CiC-IBMCC) de la **Universidad de Salamanca y del Consejo Superior de Investigaciones Científicas** (USAL/CSIC), como Director de la Tesis Doctoral titulada "***Desarrollo* y *aplicación de métodos* estadísticos *basados en recortes imparciales* a *datos de expresión génica de alta dimensionalidad***" presentada por Dña. **Itziar Fernández Martínez**, alumna del programa de **Doctorado** impartido por el departamento de **Estadística e Investigación Operativa** de la **Universidad de Valladolid**, autoriza la presentación de la misma, considerando que **ha supervisado y dirigido el trabajo realizado en los últimos cuatro años en coordinación con el director y profesor del departamento Dr. Agustin Mayo Iscar.**

Valladolid, 31 de Mayo de 2012.

El Director de la Tesis,



Fdo.: Dr. Javier De Las Rivas Sanz
Investigador Científico del CSIC
Centro de Investigación del Cáncer (CiC-IBMCC. USAL/CSIC)
Salamanca, España

ILMO. SR. PRESIDENTE DE LA COMISIÓN DE DOCTORADO

Universidad de Valladolid

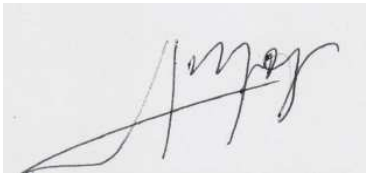
Impreso 2T

AUTORIZACIÓN DEL DIRECTOR DE TESIS
(Art. 21 del R.D. 1393/2007 de 29 de octubre y Art. 4 c) de la Normativa para la defensa de la Tesis Doctoral)

D. Agustín Mayo Iscar, con D.N.I. nº 9766882R
profesor del departamento de Estadística e Investigación Operativa, como Director de la Tesis Doctoral titulada *"Desarrollo* y *aplicación de métodos* estadísticos *basados en recortes imparciales* a *datos de expresión génica de alta dimensionalidad"* presentada por Dña. Itziar Fernández Martínez, alumna del programa de Doctorado impartido por el departamento de Estadística e Investigación Operativa de la Universidad de Valladolid, autoriza la presentación de la misma, considerando que ha supervisado y dirigido el trabajo realizado en los últimos cuatro años en coordinación con el director Dr. Javier de las Rivas Sanz.

Valladolid, 31 de Mayo de 2012.

El Director de la Tesis,



Fdo.: Agustín Mayo Iscar.

ILMO. SR. PRESIDENTE DE LA COMISIÓN DE DOCTORADO

Agradecimientos

"Sólo un exceso es recomendable en el mundo: el exceso de gratitud".

Jean de La Bruyere

Me resulta difícil agradecer a todas las personas que, de una forma u otra, han formado parte de este trabajo, no me alcanza el tiempo, y seguramente tampoco el papel. Quiero daros las gracias, y aunque es imposible nombraros a todos, sabéis que no me olvido de ninguno.

En primer lugar, quiero expresar mi gratitud a Agustín Mayo por su dedicación y por compartir conmigo su experiencia y su conocimiento. También a Javier de las Rivas, por su implicación y su entusiasmo. A los dos, muchas gracias por vuestra generosidad y por el esfuerzo y tiempo dedicado.

También me gustaría agradecer a mis compañeros del IOBA su comprensión, principalmente en esta última etapa en la que he descuidado algo nuestras tareas comunes. Especialmente quiero darle las gracias a mi jefe, José Carlos Pastor, por su apoyo, pero sobre todo por su insistencia, inagotable por cierto, y que de alguna manera, ha sido una parte fundamental en este trabajo.

A todos mis amigos que han estado siempre ahí, escuchándome y animándome. Sin excluir a ninguno, pero en especial a Maite, a Judit y a Jimena por haber tenido que sufrirme todos los días. Gracias por haber estado, pero sobre todo por seguir estando a pesar de todo.

Y finalmente quiero dedicar esta tesis a mi familia. A mis padres, mis hermanos, mi cuñada y mis sobrinos, por darme el cariño, la confianza y la seguridad para perseguir siempre mis metas. A Pedro por su optimismo, por conseguir que disfrute de todos los momentos, fáciles y difíciles, y por disfrutarlos conmigo. A todos, y en especial a Ismael, os pido perdón por el tiempo que no os he dedicado. Espero saber compensarlo.

Índice

Acrónimos

A	**A**denina
ANOVA	Análisis de la varianza (***AN****alysis* ***O****f* ***VA****riance*)
BBC	***B****ayesian* ***B****i****C****lustering model*
C	**C**itosina
cDNA	DNA codificante (***c****omplementary* **DNA**)
CGH	Hibridación genómica comparada (***C****omparative* ***G****enomic* ***H****ybridization*)
COPA	***C****ancer* ***O****utlier* ***P****rofile* ***A****nalysis*
DAVID	***D****atabase for* ***A****nnotation,* ***V****isualization and* ***I****ntegrated* ***D****iscovery*
DE	Gen diferencialmente expresado (***D****ifferentially* ***E****xpressed gene*)
DFW	***D****istribution* ***F****ree* ***W****eighted*
DNA	Ácido desoxirribonucleico (***D****eoxyribo****N****ucleic* ***A****cid*)
DLDA	***D****iagonal* ***L****inear* ***D****iscriminant* ***A****nalysis*
DT	**D**esviación **T**ípica
EE	Gen igualmente expresado (***E****qually* ***E****xpressed gene*)
EM	***E****xpectation –* ***M****aximization algorithm*
ER	Receptor de estrógeno (***E****strogen* ***R****eceptor*)
EST	Marcadores de secuencia expresada (***E****xpressed* ***S****equence* ***T****ag*)
FC	***F****old-****C****hange*
FDR	***F****alse* ***D****iscovery* ***R****ate*
FEA	Análisis de enriquecimiento funcional (***F****unctional* ***E****nrichment* ***A****nalysis*)
FP	**F**alsos **P**ositivos
FWER	***F****amily-****W****ise* ***E****rror* ***R****ate*
G	**G**uanina
GAD	***G****enetic* ***A****ssociation* ***D****atabase*
GATExplorer	***G****enomic* ***A****nd* ***T****ranscriptomic* ***Explorer***
GEO	***G****ene* ***E****xpression* ***O****mnibus*
GMS	***G****ene* ***M****atch* ***S****core*
GO	Ontología de genes (***G****ene* ***O****ntology*)
GO-CC	GO de componentes celulares (**GO**: ***C****ellular* ***C****omponent*)
GO-BP	GO de procesos biológicos (**GO**: ***B****iological* ***P****rocess*)
GO-MF	GO de funciones moleculares (**GO**: ***M****olecular* ***F****unction*)
GTI	***G****ene* ***T****issue* ***I****ndex*
HK	Genes constitutivos (***H****ouse****K****eeping*)

IM	***I**deal **M**ismatch*
ISA	***I**terative **S**ignature **A**lgorithm*
ISPY	***I**nvestigation of **S**erial Studies to **P**redict **Y**our therapeutic response with imaging and molecular analysis*
ITWC	***I**nterrelated **T**wo-**W**ay **C**lustering*
KEGG	***K**yoto **E**ncyclopedia of **G**enes and **G**enomes*
KNN	K vecinos más próximos (***K** **N**earest **N**eighbor*)
limma	***l**inear **m**odels for **m**icro**a**rray data*
log2	**log**aritmo en base **2**
MAD	Desviación media absoluta (***M**edian **A**bsolute **D**eviation*)
MAQC	***M**icro**A**rray **Q**uality **C**ontrol*
MCD	Estimador mínimo determinante de la matriz de covarianzas (***M**inimum **C**ovariance **D**eterminant*)
MDACC	**M**.**D**. **A**nderson ***C**ancer **C**enter*
MIAME	***M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment*
MM	***M**is**M**atch*
MMM	***M**ixture **M**odel **M**ethod*
MOST	***M**aximum **O**rdered **S**ubset **T**-statistics*
mRNA	RNA mensajero (***m**essenger* **RNA**)
MVE	Estimador elipsoide de mínimo volumen (***M**inimum **V**olumen **E**llipsoid*)
NCBI	***N**ational **C**enter for **B**iotechnology **I**nformation*
OHE	***O**utlier **H**igh **E**xpression* genes
ORT	***O**utlier **R**obust **T**-statistic*
OS	***O**utilier **S**um*
PAM	***P**rediction **A**nalysis for **M**icroarrays*
PCA	Análisis de componentes principales (***P**rincipal **C**omponents **A**nalysis*)
pcDNAs	Genes codificantes para proteínas (***p**rotein-**c**oding* **DNAs**)
pCR	Respuesta patológica completa (***p**athologic **C**omplete **R**esponse*)
PCR	***P**olymerase **C**hain **R**eaction*
PHE	***P**redominantly **H**igh **E**xpression* genes
PM	***P**erfect **M**atch*
PPST	***P**ermutation **P**ercentile **S**eparability **T**est*
PR	Receptor de progesterona (***P**rogesterone **R**eceptor*)
pre-mRNA	Precursor del mRNA (***pre**cursor* **mRNA**)
qRT-PCR	PCR cuantitativa tras transcripción inversa (***q**uantitative **R**everse **T**ranscription* **PCR**)
RD	Enfermedad residual *(**R**esidual **D**esease)*
RMA	***R**obust **M**ultichip **A**verage*
RNA	Ácido ribonucleico (***R**ibo**N**ucleic **A**cid*)
SAGE	Análisis en serie de expresión génica (**S**erial **A**nalysis of **G**ene **E**xpression)
SAM	***S**ignificance **A**nalysis of **M**icroarrays*
SAMBA	***S**tatistic-**A**lgorithmic **M**ethod for **B**icluster **A**nalysis*

smart-MCD	Estimador **smart** basado en el **MCD**
SMOB	***S**equential **M**ulti-**O**bjetive **B**iclustering*
SNP	Polimorfismo de un solo nucleótido (***S**ingle **N**ucleotide **P**olymorphism*)
SOM	***S**elf **O**rganizing **M**aps*
SVM	***S**upport **V**ector **M**achines*
T	**T**imina
TF	Factor de transcripción (***T**ranscription **F**actor*)
TS	Genes selectivos de tejido (***T**issue-**S**elective*)
TSp	Genes específicos de tejidos (***T**issue-**Sp**ecificity*)
U	**U**racilo
VP	**V**erdaderos **P**ositivos

Capítulo 1
Motivación y Objetivos

La última década ha estado marcada por una importante cantidad de logros en el campo de la genómica, como nueva ciencia derivada de proyectos de secuenciación de genomas completos. Iniciativas como el Proyecto Genoma Humano **[International Human Genome Consortium, 2001]**, han establecido las bases para el avance de la genética molecular hacia estudios globales capaces de medir señales de miles de genes a la vez, manteniendo la identificación de cada gen en base a su secuencia específica de DNA. El conocimiento y análisis de genomas completos ha propiciado la aparición de nuevas tecnologías que permiten extraer cantidades enormes de datos. Son las tecnologías de altas prestaciones o *high throughput*. La incorporación de estas tecnologías ha supuesto una revolución en la investigación biológica y biomédica, creando la necesidad de resolver nuevos problemas, necesidad que ha motivado la aparición y rápido crecimiento del campo de la bioinformática.

La bioinformática, definida como una nueva área de conocimiento que pretende el desarrollo y aplicación de técnicas y herramientas para el almacenamiento, organización y análisis de información biológica, es un campo de investigación multidisciplinar que incluye áreas tan diversas como la informática, la computación, la bioquímica o la estadística. Su objetivo fundamental es descubrir conocimiento a partir de las grandes cantidades de datos biológicos y biomoleculares proporcionadas por las tecnologías analíticas de alto rendimiento que permiten la producción masiva de datos. Este conocimiento debe ayudar a clarificar la regulación de los procesos celulares y los fundamentos que rigen el funcionamiento de los organismos vivos.

Uno de los objetivos más comunes dentro de la genómica es el estudio de la expresión de los genes, normalmente obtenidos como grandes matrices de datos con información cuantitativa relativa a los niveles de actividad de miles de genes. Típicamente, estas matrices contienen información de todos los genes conocidos de un organismo en distintas condiciones experimentales, en diferentes estados o en diferentes individuos (por ejemplo, sanos y enfermos). Estos volúmenes de datos tan grandes, junto con el progresivo abaratamiento de la tecnología y su uso cada vez más común, ha generado problemas relacionados con la capacidad de análisis, y la consecuente necesidad de desarrollar nuevas metodologías para solucionar estos problemas.

En el caso de los estudios genómicos que abordan datos de expresión génica, el análisis de los datos está dirigido a responder varias cuestiones como pueden ser:

- ¿Cómo varía el nivel de expresión genética entre los miles de genes que hay en un genoma?
- ¿Cómo varía el nivel de expresión genética en distintos tipos de células ó en distintos estados (sanos y patológicos/enfermos) ó en respuesta a diversos tratamientos?
- ¿Cómo están regulados los genes? ¿Qué genes interactúan entre sí? ¿En qué procesos?
- ¿Cuál es el papel funcional de determinados genes y en que procesos celulares participan?

Junto a la información derivada de los datos experimentales concretos, existe gran cantidad de información biológica disponible acerca de los genes y de los productos génicos (las proteínas), que puede ser utilizada para validar los distintos análisis e interpretar los resultados desde un punto de vista biológico. Este tipo de información se encuentra en numerosas bases de datos biológicas que incluyen miles de anotaciones y que almacenan desde el cromosoma en el que se localiza un gen, hasta las funciones moleculares en las que está implicado, o las vías de señalización, que relacionan los genes que trabajan juntos en ciertos procesos biológicos.

Una de las tecnologías genómicas más utilizadas para la medida experimental de expresión de los genes son los microarrays de alta densidad. Todos los métodos estadísticos propuestos en esta Tesis Doctoral se aplican a matrices de datos de expresión de alta dimensionalidad procedentes de esta tecnología. Sin embargo, los métodos desarrollados son perfectamente aplicables a matrices numéricas de características similares generadas a partir de otras tecnologías genómicas (ó de otras tecnologías *ómicas* en general).

1.1. Motivación

El análisis de datos de expresión génica responde a una multitud de objetivos en una amplia variedad de estudios biológicos. Hay disponible un amplio corpus de metodología estadística, incluso específicamente diseñada, para responder a ellos. Sin embargo, aunque se sabe que en muchos casos estos procedimientos funcionan razonablemente bien, hay muchos otros en los que es necesario seguir adaptando estos métodos para que respondan ante las características especiales de este tipo de datos.

Un aspecto relevante, en las matrices de expresión génica, es la presencia de mediciones atípicas. Por ello, algunos procedimientos utilizados en este ámbito nacieron con motivación robusta. La aplicación de este tipo de metodología está muy extendida en los análisis iniciales, de bajo nivel o de pre-procesado de los datos. En estos casos, la aparición de puntos atípicos puede estar relacionada con la tecnología de medida (por ejemplo, ciertas sondas de los microarrays que no son suficientemente reactivas). En el análisis de más alto

nivel, tras el pre-procesado, también aparece este problema, pero en un sentido más biológico. En muchos experimentos, en los que intervienen distintas condiciones, es difícil llevar a cabo una buena clasificación inicial de las muestras biológicas. Por ejemplo, se sabe que algunos tipos de cáncer presentan algunas sub-clases muy difíciles de distinguir morfológicamente, pero claramente diferentes en cuanto a expresión génica. Esto puede producir que las etiquetas de grupos estén mal asignadas y que necesitemos aplicar procedimientos robustos frente a este tipo de errores. También se sabe que determinadas enfermedades pueden presentar patrones de expresión heterogéneos, muy variables entre distintos individuos en algunos genes relacionados con la patología. Esto va a traducirse en la presencia, en algunos genes, de mezclas de individuos de distintos tipos. Estadísticos robustos basados en percentiles son muy habituales en el análisis de este tipo de datos, pero no tienen en cuenta la naturaleza asimétrica de la contaminación, estrategia necesaria porque en este ámbito es más frecuente la presencia de sobre-expresión que de infra-expresión.

La identificación de comportamientos atípicos en los niveles de expresión de cada uno de los genes analizados y, mejor todavía, de patrones de atipicidad ligados a un grupo de arrays para un mismo grupo de genes, permitirá identificar genes relacionados con patologías o, más generalmente, con tipologías relacionadas con la biología. Para la correcta identificación de comportamientos atípicos será necesario conocer el *core*, núcleo de expresión o nivel típico de cada gen en medidas de escala genómica (*genome-wide*) como el que aparece asociado a una mayoría de condiciones. De éste comportamiento típico, la Biología nos dice que para una mayoría de genes existe. Además, es esperable que los procedimientos robustos habituales, basados en percentiles y motivados por una idea de una contaminación simétrica, fracasen en su intento de determinarlo correctamente. En una muestra de diversos tejidos, el comportamiento relativo de tejidos sanos respecto del núcleo de expresión correspondiente a todos ellos servirá para caracterizar genes como *housekeeping* (porque muestran un comportamiento de expresión generalizado), como específicos de tejido o como predominantemente expresados.

Las técnicas de *biclustering* han experimentado un gran desarrollo en los últimos años, motivado principalmente, por su aplicación a las matrices de expresión génicas. Estaremos interesados en desarrollar herramientas de *biclustering* robustas para identificar grupos de genes que co-expresan.

En esta Tesis Doctoral proponemos la utilización de técnicas estadísticas basadas en recortes imparciales **[Gordaliza, 1991]** como estrategia para descubrir el nivel de expresión típico en un determinado gen e identificar grupos minoritarios de expresión. Esta identificación nos servirá de punto de partida para el desarrollo de nuevas herramientas estadísticas y para la adaptación de alguna metodología ya considerada clásica en este entorno.

1.2. Objetivos

El objetivo principal de esta Tesis es el desarrollo y la aplicación de herramientas estadísticas robustas en el análisis de datos expresión génica, incorporando el concepto de recorte imparcial como base de la metodología propuesta.

Las siguientes líneas contienen los objetivos que motivarán la metodología propuesta:

- Descubrimiento y caracterización del nivel de expresión típico, núcleo de expresión o *core* de un gen.
- Identificación de expresión diferencial, en experimentos con muchas condiciones biológicas distintas, respecto del núcleo de expresión. Esta identificación permitirá la determinación de genes que muestran comportamientos de expresión generalizada *(housekeeping),* específicos de tejido y de genes predominantemente expresados en un grupo de tejidos.
- Identificación de expresión diferencial entre dos condiciones biológicas o experimentales con propiedades de robustez frente a contaminaciones no conocidas a *priori* en uno o en ambos grupos.
- Identificación de genes para los que algunas condiciones muestran comportamientos atípicos o, incluso identificación de grupos de genes que comparten patrones de atipicidad.
- Obtención de agrupamientos, con propiedades de robustez, para los genes o simultáneos para los genes y los arrays, motivados por caracterizar co-expresión.

1.3. Estructura de la memoria

La memoria de esta Tesis se estructura en los siguientes capítulos,

1. Motivación y Objetivos. En este capítulo presentamos los problemas de análisis de datos que han motivado esta Tesis Doctoral y los objetivos que se persiguen con la metodología propuesta.

2. Antecedentes e introducción al tema. Resumimos los conceptos más relevantes relativos a la expresión génica, a los microarrays *y* al análisis e interpretación de este tipo de datos.

3. Análisis de expresión en múltiples muestras y múltiples clases: búsqueda del núcleo. Capítulo dedicado a la aplicación del recorte imparcial en conjuntos de datos de expresión génica con múltiples clases para identificación del núcleo de expresión y su utilidad para la identificación de genes *housekeeping*.

4. Expresión diferencial en múltiples clases independientes. En este capítulo proponemos una alternativa a los contrastes tipo ANOVA basada en la búsqueda de un núcleo de expresión y su aplicación a la identificación de genes específicos de tejido y predominantemente expresados en un grupo de tejidos.

5. Expresión diferencial en dos clases independientes con respuesta heterogénea: identificación de *outliers*. Capítulo dedicado al problema de la expresión diferencial entre dos clases con posible presencia de datos contaminantes. Proponemos metodología para detectar genes con comportamientos diferenciales, respecto de una muestra control, asociados a la mayor parte de la muestra de casos o a la activación de una fracción reducida de estos.

6. *Biclustering* para identificar patrones de co-expresión. En este capítulo se describen en detalle algoritmos de *clustering y biclustering* para identificar co-expresión y metodología orientada a la búsqueda de patrones de atipicidad comunes a grupos de genes.

7. Conclusiones

Apéndice A. Incluimos la demostración del Teorema enunciado en el capítulo 3.

Apéndice B. Se hace una pequeña descripción de los conjuntos de datos reales que se emplean en esta memoria para ilustrar la metodología propuesta.

Para todas las herramientas propuestas hemos desarrollado funciones, algoritmos y programas de R **[R Development Core Team, 2011]** que los implementan. Estos programas están disponibles en la dirección de internet,
http://www.eio.uva.es/~agustin/softmicroarrays.zip.

El funcionamiento de los métodos propuestos se ha ilustrado utilizando datos simulados y, sobre todo, utilizando varios conjuntos de datos reales correspondientes a experimentos y estudios biológicos con microarrays de oligonucleótidos de alta densidad, que es una de las tecnologías genómicas de gran escala más utilizadas para el análisis de datos de expresión génica.

Capítulo 2
Antecedentes e introducción al tema

En este capítulo se hace una revisión de algunos conceptos relacionados con la expresión génica y los microarrays como tecnología genómica utilizada para cuantificar esta expresión en muestras biológicas. Estos conceptos básicos son clave para entender la naturaleza de los datos que se van a manejar y analizar en esta Tesis. Inicialmente, se repasan las ideas y conceptos biológicos en los que se apoya este tipo de experimentos. Posteriormente, se profundiza en la tecnología en sí y en el proceso de análisis de datos, junto con las limitaciones y posibilidades que ofrece.

2.1. Antecedentes Biológicos

Expuestos de modo sencillo, en esta sección se presentan algunos conceptos biológicos referentes a los genes y el DNA –como pilares de la información biológica– cuya activación es clave para todo proceso celular y sistema vivo, y cuya medida a nivel global genómico (*genome-wide*) ha revolucionado la biología en la última década. Se puede consultar una revisión más detallada en **[Brazma et al, 2001a]**.

2.1.1. DNA, genes y proteínas: un dogma central en biología

Los pilares de la biología molecular son el DNA y las proteínas como biopolímeros que codifican la información biológica de todos los sistemas vivos conocidos, y lo hacen con dos lenguajes basados en secuencias: secuencias de nucleótidos –el lenguaje de los genes– y secuencias de aminoácidos –el lenguaje de las proteínas–. En su nivel de organización más sencillo, tanto el DNA como las proteínas se representan como secuencias lineales, es decir polímeros lineales –polinucleótidos y polipéptidos– en los que cada eslabón es un elemento: un nucleótido o un aminoácido, respectivamente. Dichos políperos lineales dan lugar a lo que se conoce como estructura primaria, que se suele identificar simplemente con la secuencia.

La información necesaria para especificar la secuencia, estructura y función de las proteínas de un organismo –es decir, su proteoma– está contenida en el genoma, que son moléculas

de DNA localizadas, en organismos eucariotas, en el núcleo de la célula, empaquetadas en forma de cromosomas durante el proceso de división celular. Los genes corresponden a las regiones codificantes del DNA, de modo que se da una perfecta coordinación entre el lenguaje del DNA y el lenguaje de las proteínas, con traducción de uno a otro mediante el código genético.

El DNA está organizado como una cadena de subunidades moleculares llamadas nucleótidos. Cada nucleótido consta de una base nitrogenada, un fosfato y un azúcar pentosa, la desoxiribosa. En el DNA existen cuatro bases diferentes: Adenina (A), Guanina (G), Citosina (C) y Timina (T).

Según el modelo de doble hélice propuesto por Watson y Crick **[Watson y Crick, 1953]**, la estructura de cada molécula de DNA consiste en dos cadenas de nucleótidos que se enrollan entre sí adquiriendo una configuración de doble hélice. Esta configuración es posible porque entre las bases de cada cadena se forman puentes de hidrógeno entre pares complementarias: frente a una A siempre se sitúa una T (A-T) y frente a una G, una C (G-C) (figura 2.1). Este principio de complementariedad es una característica muy importante del DNA, que permite los procesos de replicación y transcripción.

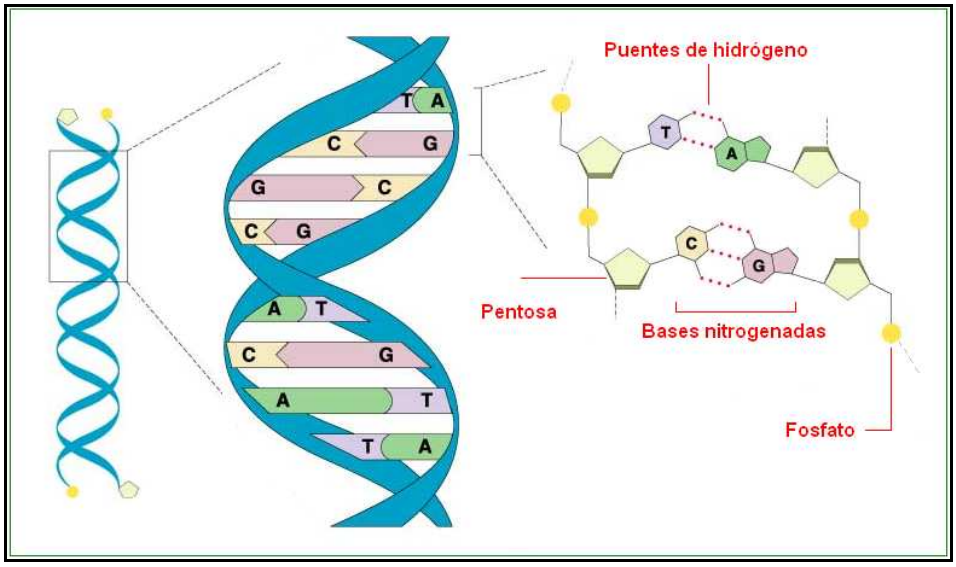


Figura 2.1. Estructura en doble hélice de una molécula de DNA.

Otra molécula importante es el RNA, que al igual que el DNA, está formado por nucleótidos. En el caso del RNA, en lugar de desoxiribosa, el azúcar de su estructura es ribosa, y en lugar del nucleótido T, tiene una molécula similar, Uracilo (U). Estas diferencias hacen que el RNA no forme habitualmente una doble hélice sino cadenas sencillas, que pueden formar estructuras espaciales complejas por los vínculos de complementariedad entre partes de la misma hebra. Existen muchos tipos de RNA que tienen diferentes funciones en la célula. En este trabajo, nos interesa fundamentalmente el mRNA (RNA mensajero) que es el intermediario entre el DNA y las proteínas en el proceso de expresión génica.

Como ya se ha indicado, tanto el DNA como el RNA son polímeros de nucleótidos llamados polinucleótidos. En biología molecular, muchas veces es necesario construir copias de un determinado gen, es decir, una región codificante concreta del genoma que correspondería a un mRNA concreto. La copia de esa región codificante concreta como DNA se llama cDNA (que corresponde al término "DNA codificante") y es lo que habitualmente se maneja en los laboratorios para clonar o manipular genes *in vitro*.

2.1.2. Genes y síntesis de proteínas

Un gen puede definirse como una región concreta del genoma –es decir, de DNA– que codifica la información de una proteína en base a la secuencia específica de nucleótidos que contiene. El caso más frecuente, y el que nos interesa en este trabajo, es el de los llamados "genes codificantes para proteínas" (a menudo denominados pcDNAs "*protein-coding DNAs*").

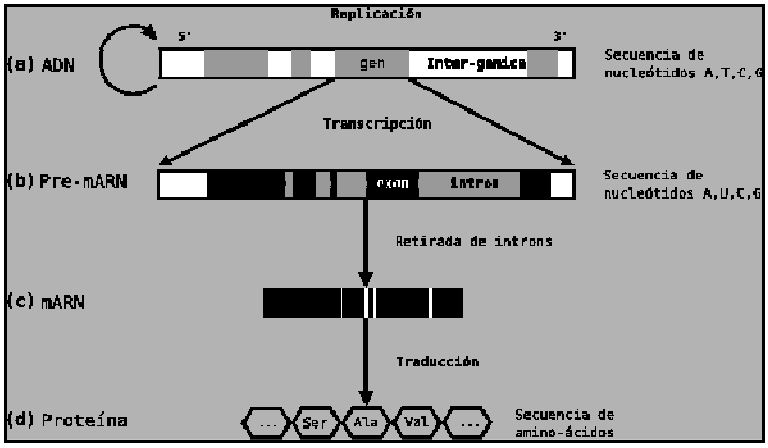


Figura 2.2. Dogma central de la biología molecular

La relación entre el DNA y la secuencia de aminoácidos de una proteína se establece a través de una ley o dogma central en biología molecular (figura 2.2). En virtud de esta ley, existe un sistema fundamental de mantenimiento y de flujo de la información genética en los organismos vivos. Por un lado, la información genética contenida en el DNA se mantiene y se preserva mediante su capacidad de replicación, y así se puede transmitir a las siguientes generaciones. Por otro lado, esta información se expresa dando lugar a proteínas que realizan trabajos y funciones específicas en la célula y en el organismo.

El proceso de replicación es el mecanismo que permite al DNA sintetizar dos copias idénticas que son transmitidas a las células hijas durante la división celular, permitiendo por tanto la transmisión de toda la información genética a las células descendientes. En la replicación es clave la característica de la complementariedad entre las bases, ya que permite que las dos cadenas complementarias del DNA original se separen, y cada una sirva de molde para la

síntesis de una nueva cadena, de forma que cada nueva doble hélice contiene una de las cadenas del DNA original.

Para realizar el paso de genes a proteínas es necesario un proceso de "decodificación" que a nivel biomolecular incluye tres etapas principales: (1) **transcripción**, (2) ***splicing*** y (3) **traducción**.

1. **Transcripción**: Es el proceso en el cual una región codificante del DNA se activa y es copiada a mRNA en forma de pre-mRNA, también denominado mRNA inmaduro. Este RNA es complementario a la secuencia de DNA que lo codifica.
2. ***Splicing***: Es el proceso en el que se eliminan las regiones del pre-mRNA que no son codificantes –que se denominan intrones– dejando las regiones codificantes para proteínas, denominadas exones. Los exones, se unen dando lugar al mRNA maduro. El número y tamaño de intrones y exones difiere considerablemente entre genes y entre especies. Se sabe que muchos genes tienen diferentes variantes en el *splicing*, de forma que la misma región de DNA codificante puede dar lugar a diferentes mRNA, lo que se conoce como *splicing* alternativo. Esto es muy frecuente en organismos metazoos superiores, como es el caso de los seres humanos.
3. **Traducción**: Es el proceso de síntesis de proteínas que se ensamblan mediante la unión de aminoácidos siguiendo la secuencia dictada por el mRNA. Cada aminoácido es determinado por tres nucleótidos adyacentes en el DNA, es decir por un triplete conocido como codón que según su secuencia concreta codifica para un aminoácido específico. La relación entre las secuencias de tres nucleótidos y los aminoácidos es lo que se conoce como "código genético", que es el que permite la traducción desde la secuencia del DNA a la secuencia de la proteína. Existen 64 codones y sólo 20 aminoácidos, por lo que este código es redundante, es decir que el mismo aminoácido puede ser codificado por varios codones.

Concluida la traducción se consigue la expresión de los genes codificantes y el producto final son las proteínas (a veces denominadas "productos génicos"), cuya secuencia se corresponde con la secuencia codificada por el mRNA.

Estos procesos constituyen un dogma central de la biología molecular, en el cual se fundamenta uno de los supuestos básicos en el análisis de datos de expresión génica: teniendo en cuenta que los genes se expresan transcribiendo y traduciendo su información en el mRNA responsable de la síntesis de la proteína, a partir de la cantidad de mRNA es posible saber qué genes y con qué intensidad se están expresando. Hoy se sabe que esto no es tan simple, y que el axioma "un gen, una proteína" no es cierto, puesto que un mismo gen a veces puede dar lugar a múltiples productos, por ejemplo, por el proceso descrito de *splicing* alternativo. Además, muchas proteínas, una vez sintetizadas, son modificadas por la unión de otras moléculas, pudiendo llegar a modificar radicalmente su actividad biológica. Son los cambios post-traducción que hacen que sea posible llegar a proteínas con una actividad biológica muy diversa partiendo de un mismo gen. Sin embargo, está claro que a

nivel genómico global la mayoría de los cambios en el estado de una célula se pueden relacionar con cambios en los niveles de mRNA, es decir cambios en los niveles de expresión de los genes, por lo que es muy interesante poder medir estos niveles de forma sistemática.

Para profundizar en el concepto de expresión génica y su significado biológico se pueden consultar textos básicos de bioquímica y biología molecular como **[Bolsover et al, 1997]** o **[Garrett y Grisham, 2002]**.

2.1.3. Hibridación de ácidos nucléicos

La hibridación de ácidos nucleicos (DNA o RNA) es el proceso por el cual dos cadenas de ácidos nucleicos con una secuencia específica de bases complementarias se unen. Los nucleótidos se unen con sus complementarios de manera ordenada y alineada, por lo que dos cadenas perfectamente complementarias se unen rápidamente y uno o pocos cambios en la secuencia, generan incompatibilidades entre las dos cadenas y dificultan o impiden la unión.

La hibridación se ha utilizado para identificar genes con secuencias específicas de DNA desde hace más de 30 años **[Alwine et al, 1977]**. Fragmentos cortos de DNA o RNA con secuencias conocidas, llamados oligonucleótidos (u oligos), también hibridan y por ello se pueden utilizar como "sondas" para pescar secuencias complementarias. Estas sondas se pueden marcar con sustancias radiactivas o fluorescentes –es decir, con marcadores químicos ligados– con el fin de hacer posible su posterior detección. Durante la hibridación se produce la unión entre la secuencia de DNA o RNA de interés –molécula diana– y la secuencia oligo que hace de sonda. Este tipo de marcajes se usan para visualizar la unión y, dado que cada molécula tiene ligado estequiométricamente un marcador, se puede cuantificar la cantidad de oligos que hay en una determinada muestra y también la cantidad de oligos que han hibridado en una determinada reacción de hibridación.

Los microarrays para medir la expresión son biochips o nano-dispositivos que están basados en el principio descrito de hibridación entre moléculas oligonucleótidos complementarios. Son dispositivos construidos con las nuevas nano-tecnologías que permiten incluir millones de moléculas en superficies microscópicas. Esta capacidad de gran escala les ha llevado a reemplazar técnicas de hibridación tradicionales en biología que podían detectar sólo unos pocos genes cada vez. A continuación se describe la tecnología de los microarrays para medir expresión génica.

2.2. Microarrays de expresión génica

2.2.1. La tecnología

Los microarrays de expresión son unos dispositivos desarrollados en el ámbito biotecnológico hace aproximadamente una década, que permiten la identificación y cuantificación del mRNA extraído a partir de células y muestras biológicas. Como ya se ha mencionado anteriormente, el número de moléculas de mRNA, procedentes de la transcripción de un gen conocido, puede considerarse como una aproximación del "nivel de expresión" de ese gen. Aunque existe gran número de situaciones biológicas en las que esta afirmación no es del todo correcta, por ejemplo siempre que el mRNA es alterado antes de dar lugar a proteínas; como idea general puede considerarse válida, y constituye la hipótesis central del análisis de datos de microarrays.

Existen varios tipos de microarrays para medir la expresión génica. Para una descripción más extensa puede consultarse **[Southern, 2001]** o **[Hardiman, 2002]**. Los dos tipos de microarrays más utilizados y tecnológicamente más distintos son: los arrays cDNA (o también llamados de dos canales) y los arrays de oligonucleótidos (también llamados de un canal). Las superficies empleadas para fijar el DNA o los oligonucleótidos son muy variables: vidrio, plástico e incluso de silicio. En ambos casos, el proceso de hibridación, representado en la figura 2.3, es la clave de su funcionamiento, pero se diferencian en el tipo de molécula utilizada como sonda y en la forma en la que miden la señal de hibridación.

- Microarrays de *spotted* cDNA o dos canales: en los que las sondas son moléculas de DNA complementario de cadena sencilla que se sintetizan *in vitro* y se colocan sobre la superficie. Cada sonda representa normalmente un gen. Estos arrays se describen en **[Schena et al, 1995]** y **[DeRisi et al, 1997]**.
- Microarrays de oligonucleótidos o de canal único: en los que las sondas son moléculas de oligonucleótidos que se sintetizan directamente sobre la superficie. Los oligonucleótidos son polímeros cortos de DNA de cadena sencilla de 25, 30 o 60 nucleótidos. En estos arrays cada gen está representado por varios oligonucleótidos de secuencia distinta. A este tipo de arrays nos referimos a continuación con más detalle ya que son los que se utilizarán en este trabajo.

Un microarray es una superficie sólida inerte sobre la que se inmovilizan, de forma ordenada, miles de secuencias conocidas de oligonucleótidos llamados sondas ó "*probes*". Cada sonda está repetida miles de veces en un espacio micro o nanoscópico y tiene una secuencia específica que se conoce por corresponder a un fragmento de secuencia de un gen. Cada gen suele estar representado en el microarray por un conjunto de sondas de secuencias distintas que constituyen el llamado "*probe-set*", distribuidas aleatoriamente en la superficie del microarray. De este modo, cada secuencia se asocia con un único gen y cada gen está representado por un conjunto de secuencias. En un microarray, que tiene capacidad para cientos de miles de sondas, se pueden incluir todos los genes conocidos de un genoma. En este trabajo utilizaremos microarrays de oligonucleótidos manufacturados por la

compañía *Affymetrix*, llamados GeneChips, que incluyen oligos de 25 nucleótidos y en los que cada gen está representado por un conjunto de 11 ó 16 sondas **[Affymetrix, 1999]**.

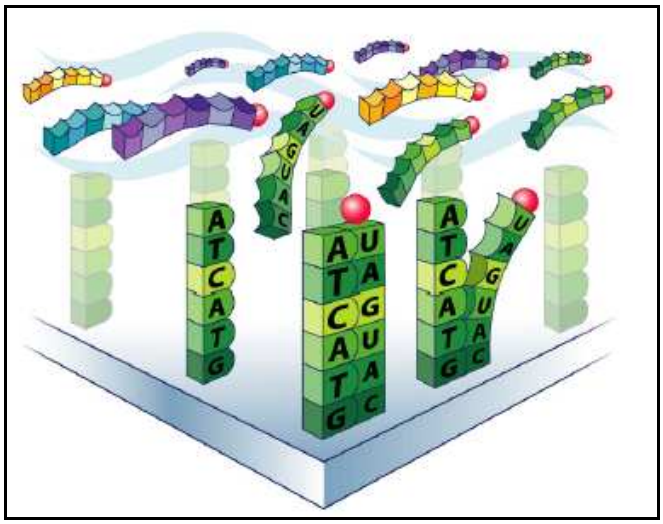


Figura 2.3. Proceso de hibridación en un microarray de DNA

Existen también otros tipos de microarrays de oligos cuyo objetivo no es medir la cantidad de expresión de los genes. Como ejemplo se describen dos:

- Microarrays de polimorfismos de un solo nucleótido o SNP (*Single Nucleotide Polymorphism, SNP Arrays*). Su objetivo es detectar SNPs en una población. Un SNP es una variación en la secuencia de DNA que afecta a una sola base, y es la variación más frecuente en el genoma. En el genoma humano, por ejemplo, existen alrededor de 10 millones de SNPs identificados. Los polimorfismos constituyen la base de las diferencias genéticas entre individuos. Mapear y obtener las frecuencias de SNPs puede ser útil para identificar genes responsables de enfermedades, predecir los efectos de determinadas condiciones ambientales e incluso respuestas a determinados fármacos.
- Microarrays de hibridación genómica (*CGH* Arrays) que se utilizan para detectar la presencia de ganancias (amplificación) o pérdidas (deleción) de segmentos del genoma. Se han usado mucho en el campo de la oncología para la detección de amplificaciones y deleciones genómicas asociadas a tumores humanos y la definición en detalle de regiones comúnmente amplificadas.

Como se ha indicado, este trabajo se centra en los microarrays de expresión. En lo que sigue, y con objeto de simplificar, el término microarray hará siempre referencia a este tipo de microarray.

2.2.2. El proceso de medición

El primer paso en un experimento con microarrays es la extracción del RNA de las células. Cada molécula de RNA extraída es troceada en oligos y marcada uniéndole una sustancia

fluorescente. Una vez preparada la muestra se coloca sobre el microarray, que se introduce en una cámara de hibridación durante algunas horas. Durante este tiempo, los fragmentos de RNA marcados se unen por hibridación a aquellas sondas del microarray que son complementarias. Después de eso, el microarray se lava para eliminar los fragmentos que no han hibridado. En el sistema de *Affymetrix* se hibrida sólo una muestra biológica por chip.

Finalizado el proceso de hibridación, cada sonda del microarray tendrá pegada o no, su oligo complementario marcado. Como existen miles de copias de cada sonda en la micro-región (micro-celda) del array donde está ubicada, la cantidad de hibridación debe ser proporcional al nivel de expresión del correspondiente mRNA; es decir del gen representado por esa sonda. Para determinar la cantidad de muestra hibridada, se utiliza un escáner láser de alta sensibilidad que sirve para excitar y detectar las moléculas marcadas con fluorescencia. Cada señal fluorescente de una micro-celda será proporcional a la cantidad de moléculas marcadas que hayan hibridado. Esta fluorescencia se captura como una imagen y se cuantifica. Al final se obtiene para toda la superficie del array una gran imagen, de alta resolución, formada por una malla de puntos brillantes, cada uno de los cuales se corresponde con una sonda. En la figura 2.4 se ilustra un ejemplo del tipo de imágenes que se obtienen para una muestra hibridada concreta.

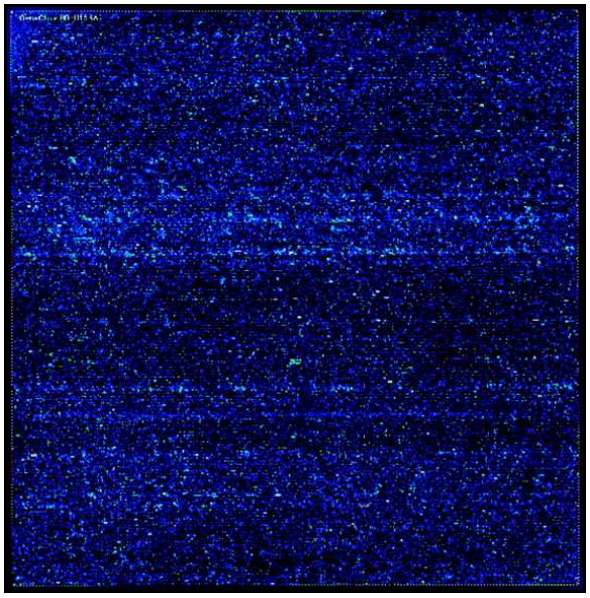


Figura 2.4. Imagen de un microarray de canal único. Las intensidades de luz revelan los niveles de expresión de las sondas que miden los correspondientes genes.

2.2.3. Medidas de expresión

Los microarrays cuantifican la expresión génica utilizando la intensidad de fluorescencia capturada a través del escaneo de una imagen. Cada *probe* se representa como un punto o *spot* en esa imagen. La imagen se divide en píxeles y cada *spot* puede estar representado

por varios píxeles. La intensidad de cada píxel se transformará en un número utilizando herramientas de análisis de imagen. Básicamente este proceso va a consistir en dos tareas, **(i)** segmentación de la imagen para identificar las áreas en las que existe expresión, y **(ii)** cálculo de la intensidad de los píxeles que forman parte del mismo *spot*. Se han utilizado muchos métodos, tanto para la segmentación de las imágenes, como para calcular las intensidades de píxel. Algunas revisiones sobre técnicas de análisis de imagen relacionadas con microarrays son, por ejemplo **[Brown et al, 2001]** y **[Jain et al, 2002]**.

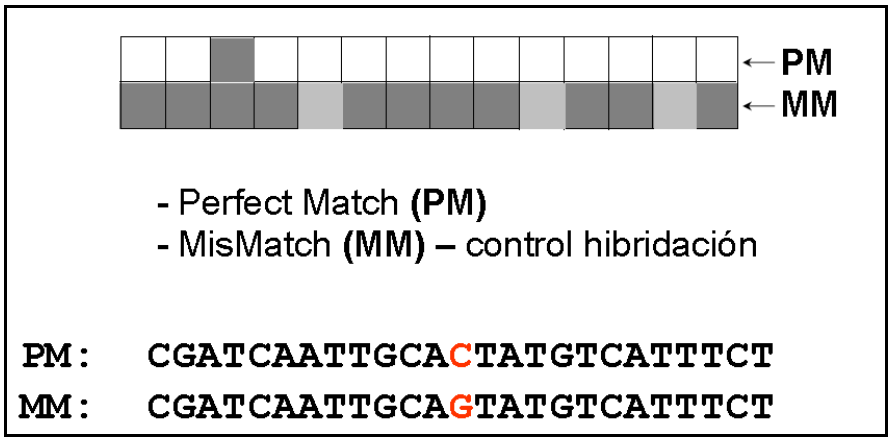


Figura 2.5. *Probe-set* de un GeneChip de *Affymetrix*.

Las medidas obtenidas del análisis de la imagen deben procesarse para proporcionar una estimación de una variable: la expresión de los genes. Ya se ha comentado que en los GeneChips de *Affymetrix* cada gen está representado por un conjunto de sondas (*probe-set*), cada una de las cuales se corresponde con un oligonucleótido. Además, en los microarrays de *Affymetrix* cada *probe* está organizada en parejas, *probe-pair*, compuestas por un *Perfect Match* (PM), que coincide exactamente con una secuencia del gen estudiado, y un *MisMatch* (MM), idéntico al PM excepto en el nucleótido central, que se reemplaza por el complementario (figura 2.5). La idea que subyace en este diseño, es que cualquier cosa que hibride con el MM no representa a la expresión real, es decir, puede ser considerado como *background*. *Affymetrix* **[Affymetrix, 2002]** propuso combinar estas dos medidas (PM y MM) en una única medida de expresión corregida por *background*.

Una vez que se tiene la señal de expresión cruda de todas las sondas de un microarray, es muy habitual utilizar como estimador final el logaritmo en base 2 (log2) de las medidas de expresión **[Stekel, 2003]**. La primera razón tiene que ver con que la distribución de las intensidades de señal es asimétrica por la derecha, ya que la mayoría de los genes se expresan a niveles bajos o muy bajos, cercanos a cero (figura 2.6). La transformación logarítmica corrige la asimetría de este tipo de distribuciones. Por otro lado, tomar logaritmos va a facilitar la interpretación. En un experimento con microarrays, para medir cambios en un gen, es muy frecuente utilizar el *fold-change* (FC), definido como un cociente entre el nivel de expresión en una muestra objetivo, respecto al valor de expresión en una

muestra tomada como referencia. Trabajar con cocientes tiene desventajas en cuanto a la comparación de genes sobre- e infra-expresados. Un valor de FC en el rango [1,∞) indica sobre-expresión del gen en la muestra objetivo respecto de la de referencia, y un valor de FC en el rango [0,1] indica infra-expresión del gen en la muestra objetivo respecto de la de referencia. En la escala logarítmica, el FC se convierte en una diferencia, y un determinado valor y su recíproco serán simétricos. Valores positivos indicarán sobre-expresión, negativos, infra-expresión y un gen expresado a nivel constante tendrá un logFC de 0. La base de la transformación es normalmente 2 y así, cambios de 1 unidad en la escala log2 doblan su valor en la escala original.

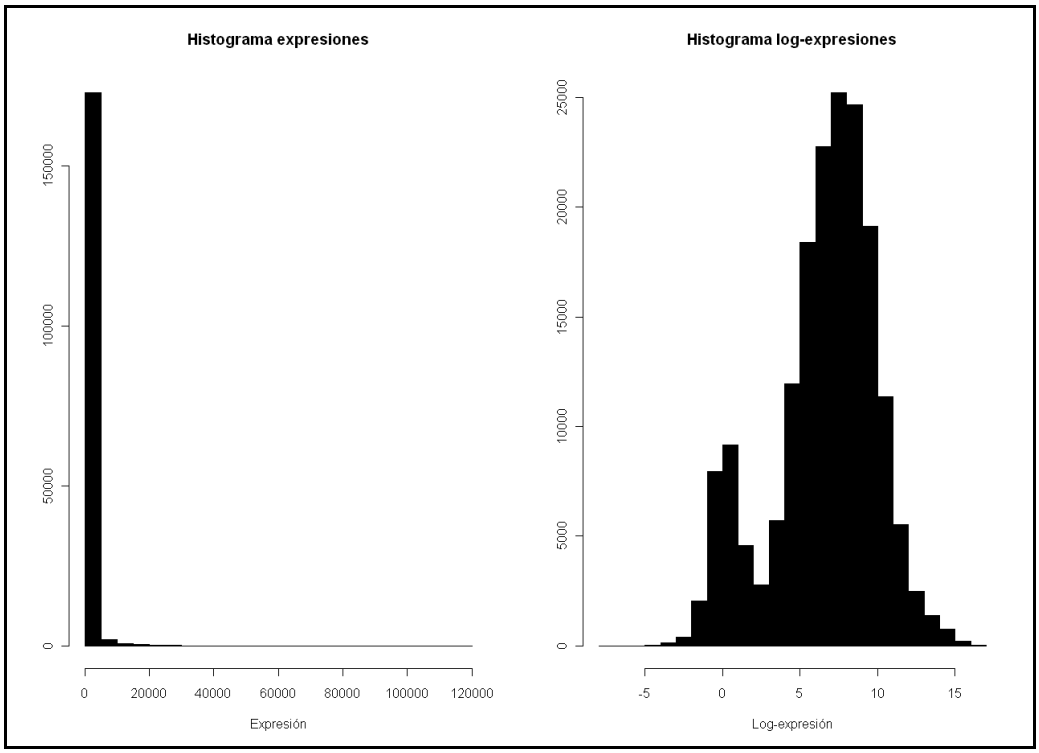


Figura 2.6. Ejemplo de la distribución de intensidades en los GeneChip de *Affymetrix*

Una explicación alternativa para el uso de la transformación logarítmica puede basarse en la ley de Weber-Frechner **[Weber, 1834] [Fechner, 1860]** o la ley de Steven **[Stevens, 1957]**, que establecen una relación logarítmica entre la intensidad de un estímulo y la percepción humana de dicho estímulo.

2.2.4. Aplicaciones de los microarrays

Los microarrays nacen de la necesidad de desarrollar tecnologías de alto rendimiento que permitan analizar, a la vez, el estado de todos los genes en genomas conocidos gracias a los grandes proyectos de secuenciación desarrollados en los últimos años **[Wood et al, 2002]**; **[Venter et al, 2001]**; **[International Human Genome Consortium, 2001]**; **[Adams et al, 2000]**; **[Blattner et al, 1997]**. En este marco, las principales aplicaciones se centran en el análisis de

la expresión génica mediante técnicas de hibridación y en el diagnóstico genético molecular mediante técnicas de secuenciación. Algunos de los tipos de estudios genómicos más frecuentes en todo este área son:

- Monitorización y análisis comparativo de la expresión genética: Una de las aplicaciones más comunes es la monitorización de los niveles de expresión de genes comparando dos o más condiciones. Este tipo de estudio proporciona información acerca de la función de determinados genes en un determinado estado. La hipótesis global que subyace es que los genes que comparten un patrón de expresión común deben compartir funcionalidad.
- Selección de nuevos fármacos identificando dianas terapéuticas más efectivas: La identificación del gen responsable de un determinado mecanismo biológico implicará la aparición de una nueva diana terapéutica, sobre la que potencialmente se podrá actuar para regular dicho mecanismo.
- Farmacogenómica: Estudio del impacto que tienen determinadas variaciones genéticas en la eficiencia y toxicidad de los fármacos. En este tipo de estudios, se correlacionan el perfil genético de los individuos con la respuesta de cada uno de ellos a un fármaco determinado.
- Diagnóstico molecular en patologías: Identificación de genes marcadores específicos para determinadas enfermedades, junto a la identificación molecular de microorganismos patógenos (virus, bacterias, hongos, etc) y de mecanismos implicados en su resistencia a ciertos agentes antimicrobianos.

2.2.5. Limitaciones de los microarrays

2.2.5.1. Limitaciones metodológicas de los microarrays de expresión

Como ya se mencionó anteriormente, el dogma central de la biología molecular, y su axioma "un gen, una proteína", es el fundamento del análisis de datos de expresión génica. Según esta visión, la relación uno a uno entre la secuencia de DNA y la de mRNA hace que sea posible, midiendo la cantidad de mRNA transcrito de un determinado gen, aproximar el nivel de expresión de dicho gen. Desafortunadamente, el proceso de expresión génica es más complejo y la visión moderna de este dogma central entiende dicha complejidad y debe estar abierto a que un gen pueda codificar más de una secuencia de mRNA.

Como se ha dicho ya, en el proceso de *splicing* no siempre se seleccionan los mismos exones, de forma que distintos *splicings* alternativos sobre un mismo gen pueden producir distintas moléculas de mRNA, que pueden codificar distintas proteínas. En **[Clancy, 2008]** se muestra el ejemplo de un gen expresado en el cerebro de la mosca de la fruta (*Drosophila melanogaster*) que puede ser particionado de 40000 formas diferentes.

Otras fuentes de variación se deben a los factores de transcripción (TF). Un factor de transcripción es una proteína que participa en la regulación de la transcripción del DNA mediante la unión específica a distintos sitios (*binding sites*) de los genes. Distintos factores

de transcripción o la combinación de varios de ellos, pueden producir la transcripción de un determinado locus génico de formas diversas. Cada factor de transcripción y/o sitio de unión al que se acople, pueden producir la transcripción de una cadena diferente de mRNA.

También es variable la tasa de traducción. La traducción de mRNA a proteína puede que no se produzca siempre con la misma frecuencia ni velocidad. De esta forma, la relación entre la cantidad de mRNA y la cantidad de proteína puede ser distinta en determinados momentos.

Por otra parte, se sabe que pueden producirse modificaciones en la estructura de la proteína después de la traducción. La proteína puede ser troceada en otras de menor tamaño, o algunas de sus regiones de aminoácidos pueden ser eliminadas o modificadas por la unión, ya sea temporal o permanente, de otros grupos. Estos cambios post-traducción pueden tener unos efectos drásticos en las propiedades funcionales y estructurales de la proteína, y por tanto en el comportamiento celular.

Con todo esto, hay que tener en cuenta que la cantidad de mRNA transcrito no tiene porque corresponder exactamente con el nivel de expresión del gen asociado, ni con la cantidad de proteína que codifica dicho gen. Por lo tanto, la tecnología genómica de microarrays de expresión, aunque es válida en términos globales, para casos concretos puede ser bastante imprecisa. Hoy en día, y gracias al crecimiento del campo de la proteómica, se están desarrollando nuevas tecnologías de gran escala que miden directamente las proteínas en sus distintos estados estructurales y funcionales. Estas tecnologías proteómicas son muy buen complemento a las tecnologías genómicas, y ambas deben ser utilizadas para lograr estudios profundos y válidos en sistemas biológicos concretos.

2.2.5.2. Limitaciones técnicas de los microarrays de expresión

Existen numerosas fuentes de error debidas a imprecisiones e imperfecciones en el instrumental técnico utilizado para la hibridación y escaneado de los microarrays de expresión. Algunos ejemplos serían, heterogeneidad en la cantidad de material biológico impreso en cada *spot* del microarray o en la cantidad o estiquimetría de reactivo fluorescente que se utiliza para el marcaje de las muestras, errores en la medición de luz por parte del escáner, etc.

Estos errores son intrínsecos a los aparatos o dispositivos de medida y por ello son principalmente un problema que deben resolver los fabricantes y desarrolladores de dichos dispositivos. La minimización de estos errores es necesaria para que los análisis posteriores, y las conclusiones derivadas de los mismos, sean fiables. Por lo tanto, tras la obtención de los datos crudos fiables, es necesario el llamado "pre-procesado" de los datos de microarrays, que incluye todo lo referente a la corrección de *background*, estimación de la señal-ruido, normalización de los datos intra-muestra y entre-muestras, etc. Todo esto es previo al análisis de datos propiamente dicho, pero es fundamental para poder obtener unos buenos resultados.

A pesar de todas estas limitaciones, la tecnología de microarrays de expresión, resulta relativamente barata y permite medir en paralelo niveles de expresión de miles de genes en cientos de muestras. Por tanto, son aceptados por la comunidad como una herramienta útil con la que extraer información acerca de los procesos celulares, aunque se sabe que los resultados obtenidos deben ser interpretados con cuidado y sobre todo, validados con información procedente de otras fuentes de datos.

2.3. Datos de microarrays

La tecnología de los microarrays de expresión proporciona unos resultados que no son directamente analizables. En esta sección se describen los pasos preparatorios para el análisis de los datos: **(i)** diseño experimental, **(ii)** control de calidad y **(iii)** pre-procesado.

En la figura 2.7 se muestra un esquema de todo el proceso del análisis de datos de microarrays de forma integrada.

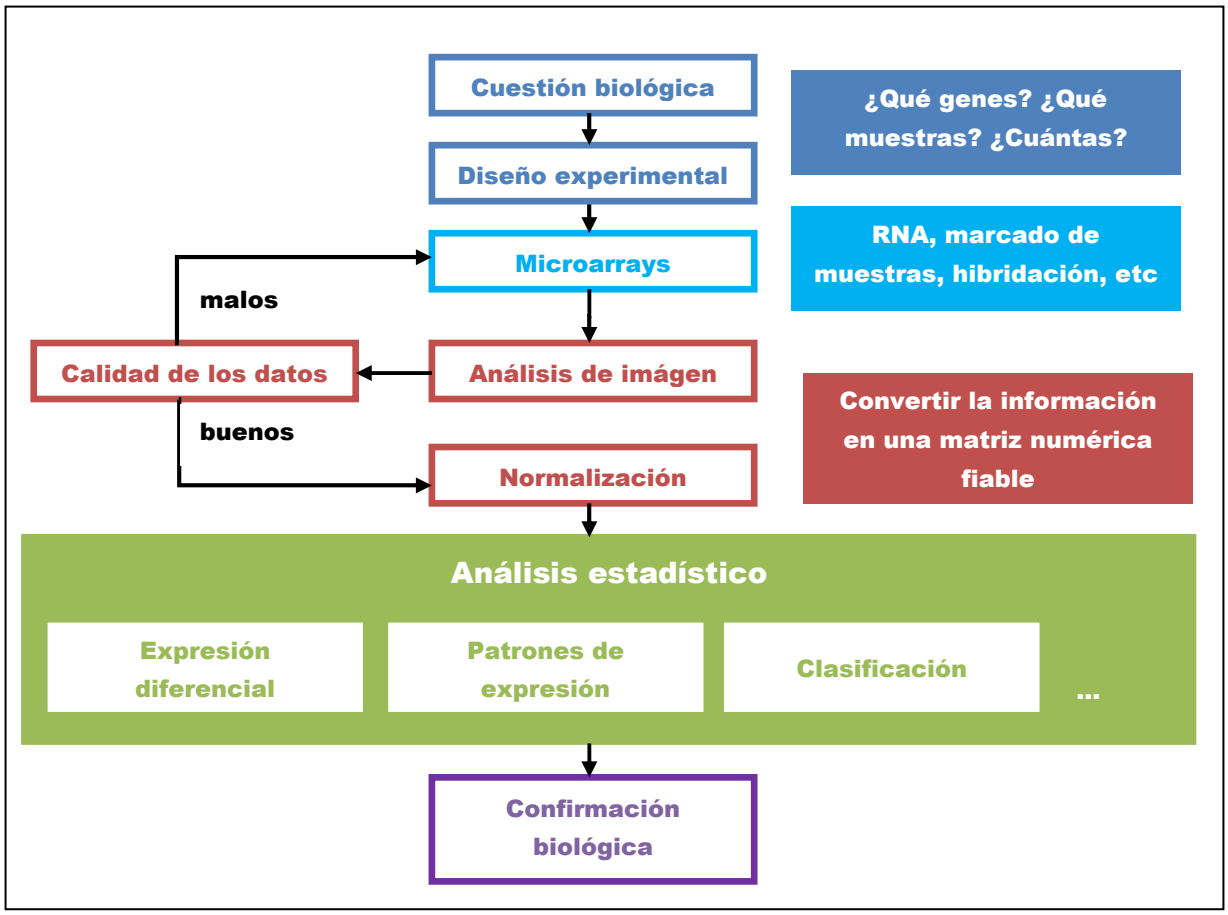


Figura 2.7. Diagrama de flujo de un proceso típico de análisis de datos de microarrays

2.3.1. Diseño experimental

2.3.1.1. Fuentes de variabilidad

Los datos genómicos son muy variables. La figura 2.8, adaptada de **[Geschwind y Gregg, 2002]**, ilustra alguna de las fuentes de variabilidad de este tipo de experimentos.

Como es habitual en muchas situaciones experimentales, se puede distinguir entre variabilidad sistemática y aleatoria. La primera de ellas se atribuye a los procedimientos técnicos, mientras que la aleatoria se atribuye tanto a razones técnicas como biológicas. Ejemplos de variabilidad sistemática podrían ser la extracción de RNA, marcaje o foto-detección. La variabilidad aleatoria puede estar relacionada como muchos más factores, como la calidad del DNA o las características biológicas de las muestras.

Las correcciones para la variabilidad sistemática se basan en su estimación en los procesos de calibración y normalización. En el caso de la variabilidad aleatoria, se deben utilizar diseños experimentales apropiados que permitan su control y herramientas estadísticas que permitan extraer conclusiones válidas.

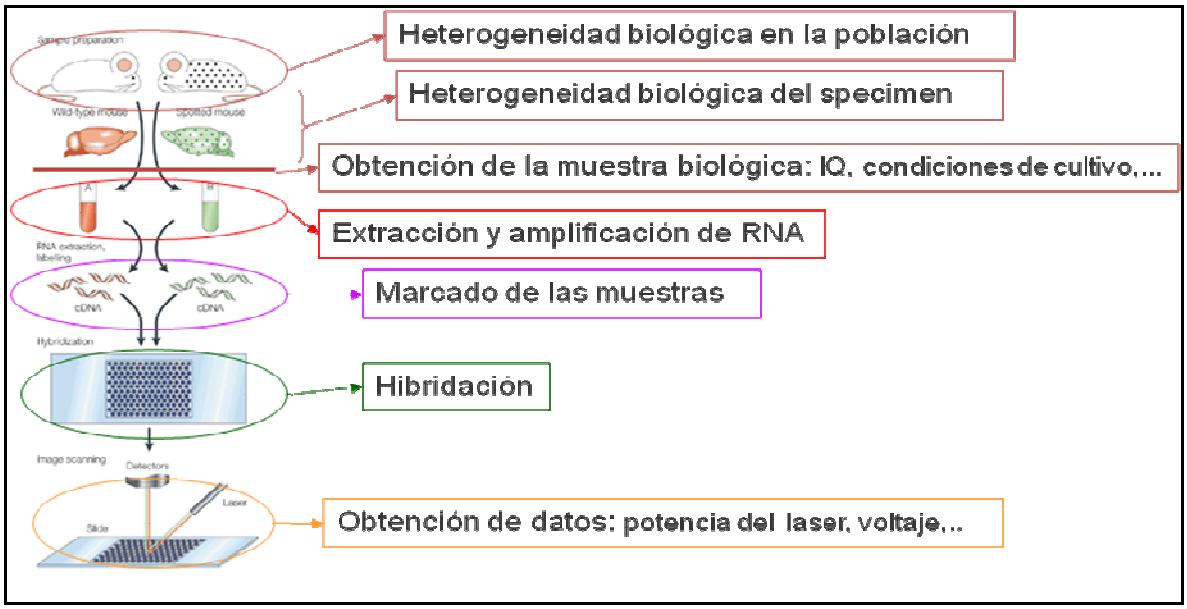


Figura 2.8. Fuentes de variabilidad en datos de microarrays

2.3.1.2. Replicación

La existencia de controles internos es muy útil para evaluar las numerosas fuentes de error en este tipo de experimentos, pero no son suficientes, y desde hace ya unos años, la comunidad acepta que es necesario replicar los experimentos **[Lee et al, 2000]**. Existen dos tipos de réplicas en un experimento de microarrays,

- Réplicas técnicas: cuando se hacen varias réplicas instrumetales del mismo material biológico (se suele denominar también alícuotas; son simplemente una muestra dividida

en varias con las que se realiza el mismo análisis o proceso de estudio tantas veces como réplicas haya).

- Réplicas biológicas: cuando se hacen mediciones de múltiples muestras biológicas diferentes tratadas del mismo modo o sometidas al mismo proceso (por ejemplo, una serie de ratones hermanos tratados igual o una serie de pacientes con la misma enfermedad, etc).

La replicación técnica proporciona estimaciones del nivel de error de la técnica, mientras que la biológica trata de estimar la variabilidad poblacional.

2.3.1.3. Potencia y tamaño muestral

Inicialmente, en los experimentos de microarrays se utilizaban pocas réplicas biológicas, principalmente por el alto coste del procesamiento de cada microarray. En pocos años, el coste de los chips ha descendido considerablemente, por lo que los tamaños muestrales han aumentado y han aparecido multitud de métodos específicos para el análisis de la potencia y el cálculo del tamaño muestral **[Pounds y Cheng, 2005]**; **[Jung et al, 2005]**; **[Tibshirani, 2006]**; **[Lin et al, 2010]**. A pesar de ello, no existe un claro candidato para su uso en situaciones prácticas, probablemente debido a las complejas estructuras de correlación que existen entre los genes, la mayoría de las veces desconocidas y difíciles de estimar. Sin embargo, sí que existe consenso sobre la necesidad de utilizar métodos específicamente diseñados para estudiar la potencia, y por supuesto, sobre la afirmación de que más cantidad de réplicas van a proporcionar mayor potencia **[Allison et al, 2006]**.

2.3.2. Pre-procesado

Un experimento de microarrays proporciona un conjunto de imágenes que deben ser transformadas a valores numéricos. Como en cualquier análisis estadístico, y en especial en el análisis de imágenes, se debe comprobar la calidad de los datos. Los datos de microarrays tienen una dificultad añadida, y es que las matrices de datos son enormes, haciendo que sea imposible detectar muchos problemas de manera visual. Esto ha provocado que se hayan desarrollado procedimientos específicos para el control de calidad.

2.3.2.1. Control de calidad

El objetivo del control de calidad es determinar si el proceso de medición ha sido lo suficientemente bueno para que los datos puedan considerarse fiables. Mayoritariamente, las herramientas que existen son gráficas, y aunque no hay métodos estándar, existe un grupo de trabajo conocido como MAQC (*MicroArray Quality Control*) **[Shi et al, 2006]**, en el que están involucrados los principales proveedores de plataformas de microarrays actuales, junto con algunos institutos nacionales de salud, institutos de tecnología y laboratorios universitarios de distintas partes del mundo. Su objetivo es precisamente desarrollar estándares en este ámbito de los análisis de microarrays.

2.3.2.2. Corrección de background, normalización y sumarización

Una vez que se ha realizado un control de calidad adecuado de los microarrays y se ha logrado la obtención de la señal numérica cruda, es necesario hacer lo que se denomina normalmente el "pre-procesamiento" de los datos. A partir de ahora, en esta memoria se hablará principalmente de términos y conceptos que se aplican a los microarrays de oligonucleótidos de *Affymetrix*, ya que todo nuestro trabajo posterior se ha centrado en este tipo de plataforma.

Habitualmente el "pre-procesado" consiste en tres pasos,
1. Corrección del *background*, cuyo objetivo será eliminar la señal producida por una hibridación no específica, es decir aquella que no se debe a la hibridación entre la muestra y las sondas (*probes*) de medida.
2. Normalización de los datos para corregir sesgos sistemáticos debidos a variaciones introducidas durante la preparación de muestras, durante la preparación de los arrays y/o durante su procesamiento.
3. Cálculo de la señal por gen o entidad génica, que consiste en calcular un único valor sumarizado o resumen a partir de las diferentes medidas de señal obtenidas de todas las sondas correspondientes a un mismo gen o entidad biológica medida. También se llama sumarización.

Existen muchos métodos específicamente desarrollados para este tipo de análisis, conocido habitualmente como análisis a bajo nivel (*low level microarray analysis*). Muchos de estos métodos combinan los tres pasos. Algunos ejemplos de los más utilizados son: MAS5 **[Hubbell et al, 2002]**, RMA (*Robust Multichip Average*) **[Irizarry et al, 2003]** o DFW (*Distribution Free Weighted*) **[Chen et al, 2007]**.

Como ya se ha mencionado anteriormente, en los GeneChip de *Affymetrix*, cada *probe* en un *probe-set* está organizada en un *probe-pair* compuesto por un PM (*Perfect Match*), que coincide exactamente con una parte del gen estudiado, y un MM (*Mismatch*), cuyo nucleótido central se sustituye por su complementario. Los MM están diseñados para dar una medida de la hibridación no específica asociada a su correspondiente PM. Sin embargo, en la práctica esto no funciona, y se sabe que aproximadamente el 30% de los MM tienen valores mayores que sus correspondientes PM **[Naef et al, 2001]**. *Affymetrix* introduce el concepto de *Ideal Mismatch* (IM) **[Affymetrix, 2002]**, para resolver el problema de valores de expresión negativos, garantizando, por diseño, que el valor del IM es menor que el del correspondiente MM. Sin embargo, su uso no es muy extendido, y muchos de los métodos de pre-procesado más populares resuelven el problema, simplemente, ignorando las probes MM.

En este trabajo se ha utilizado RMA **[Irizarry et al, 2003]** para normalizar todos los conjuntos de datos reales analizados. Este método utiliza únicamente los PM y asume un modelo lineal y aditivo para corregir el *background*. Para la normalización, utiliza un procedimiento conocido como normalización por cuantiles **[Bolstad et al, 2003]**, cuyo objetivo es hacer que la

distribución empírica de las intensidades de un gen para cada array sea la misma. En el paso de sumarización utiliza el algoritmo *median polish* **[Tukey, 1977]**.

2.4. Análisis estadísticos de datos de expresión génica

Los datos de microarrays, así como otros datos genómicos, son algo diferentes al tipo de datos habitual utilizado en análisis estadísticos. En consecuencia, en muchos casos, es necesario adaptar técnicas existentes o desarrollar herramientas nuevas que permitan abordar problemas clásicos de la estadística. En esta sección se hace un breve repaso de los problemas en los que más se ha trabajado en el área de la expresión génica junto con las soluciones más populares.

El punto de partida, es la matriz de expresión, que en caso de los microarrays se obtiene tras el pre-procesamiento de los datos. En las filas de esta matriz estarán representados los genes como variables medidas (entre miles y decenas de miles de genes según la plataforma) y en las columnas las muestras, observaciones o individuos (que normalmente son decenas o centenares). La estructura y tamaño de esta matriz es un poco diferente a las que se utilizan habitualmente en estadística, donde las filas representan variables y las columnas individuos, y normalmente, se tienen muchos más individuos testados que variables medidas.

2.4.1. Expresión diferencial

El problema de la expresión diferencial se define como la selección de genes cuyo nivel de expresión es significativamente diferente entre las diferentes condiciones experimentales o biológicas estudiadas. De cada uno de estos genes seleccionados se dirá que está diferencialmente expresado.

Se han desarrollado muchos modelos y métodos para este tipo de problema y existen muchas revisiones en la literatura, como por ejemplo **[Pan, 2002]**; **[Hatfield et al, 2003]** o **[Cui et al, 2005]**. Lo más habitual es encontrar métodos que analizan cada gen por separado, bien utilizando herramientas paramétricas como limma (*Linear Models for Microarray Data*) **[Smyth et al, 2005]** o alternativas no paramétricas como SAM (*Significance Analysis of Microarrays*) **[Tusher et al, 2001]**.

2.4.2. El problema de las comparaciones múltiples

El análisis de microarrays gen a gen requiere realizar un test por cada uno de los miles de genes analizados, y esto introduce el problema del cálculo de una tasa de error global a

todas las decisiones tomadas, el problema de las comparaciones múltiples. Si se realizan miles de contrastes, la probabilidad de que se obtengan falsos positivos aumenta y esto debe ser corregido. Una opción es controlar la FWER (*Family-Wise Error Rate*), definida como la probabilidad de tener uno o más falsos positivos en el conjunto de contrastes realizados. Sin embargo, criterios basado en el control de FWER pueden ser muy restrictivos y muchos biólogos prefieren controlar la FDR (*False Discovery Rate*) **[Benjamini y Hochberg, 1995]**, definida como la proporción de falsos positivos entre todos los genes inicialmente identificados como diferencialmente expresados. En **[Dudoit et al, 2003]** se hace una revisión muy extensa sobre el problema de las comparaciones múltiples aplicado al análisis con datos de microarrays.

2.4.3. Descubrimiento de grupos o clústers

El análisis clúster es el método más popular utilizado para, a partir de la matriz de expresión, identificar y agrupar genes expresados a nivel parecido, genes co-expresados. La idea subyacente es que genes con funciones biológicas relacionadas probablemente vayan a expresarse simultáneamente.

En el contexto de la expresión génica, las técnicas clásicas de *clustering* pueden aplicarse tanto a genes, como a condiciones, como a ambas. En este tipo de datos ocurre que muchos patrones de actividad de grupos de genes sólo se presentan bajo un determinado conjunto de condiciones experimentales, mientras que bajo otras condiciones los mismos genes pueden comportarse de forma diferente. Las técnicas de *biclustering* **[Van Mechelen et al, 2004]** pueden realizar el agrupamiento en las dos dimensiones de forma simultánea.

Algunos algoritmos de *clustering* que se han aplicado a este tipo de datos son, por ejemplo, el *clustering* jerárquico, aplicado por primera vez en **[Eisen et al, 1998]**, el algoritmo de las k-medias **[Tavazoie et al, 1999]** y métodos como SOM (*Self Organizing Maps*) **[Tamayo et al, 1999]**.

En cuanto al *biclustering*, desde su primera aplicación a datos de expresión génica **[Cheng y Church, 2000]**, han sido muy numerosos los algoritmos desarrollados para responder a este problema. Algunos intentos de clasificar toda esta metodología puede consultarse en revisiones como **[Madeira y Oliveira, 2004]**.

2.4.4. Predicción de clases

El objetivo de la predicción de clases es ajustar modelos que permitan clasificar a un nuevo individuo según ciertas características que intervienen en dicho modelo. Una revisión de los métodos empleados para este objetivo puede consultarse en **[Díaz-Uriarte, 2005]**.

El método más extendido es el DLDA (*Diagonal Linear Discriminant Analysis*) **[Dudoit et al, 2002a]**, variante del análisis discriminante clásico. También son muy populares los métodos de los K vecinos más próximos (KNN, K *Nearest Neighbor*), y otras herramientas dentro del campo de aprendizaje automatizado como el SVM (*Support Vector Machines*) o el *Random Forest*. Además se han utilizado métodos de agregación **[Breiman, 1996]**, en los que la combinación de varios predictores consigue mejorar los clasificadores globales. El *Bagging* **[Breiman, 1996]** y el *Boosting* **[Freund y Schapire, 1996]** son dos métodos de esta clase, que han sido aplicados con éxito a datos de microarrays. En **[Dudoit et al, 2002a]** y en **[Won Lee et al, 2005]** se hace una evaluación y comparación de más de 20 métodos utilizados en este campo.

2.5. Fuentes de información biológica y usos

Como resultado de un análisis de datos de expresión génica es frecuente encontrarse largas listas de genes que han sido seleccionados utilizando algún criterio. A partir de estas listas es necesario dar una interpretación biológica.

La comunidad científica que centra sus investigaciones en la genómica es muy grande. Gracias a revistas, bases de datos y repositorios, los resultados de esas investigaciones están disponibles casi en tiempo real. Los estudios genómicos no sólo utilizan este conocimiento biológico como una fuente externa de validación de resultados, sino que en muchas ocasiones, los experimentos se diseñan teniendo en cuenta esta información. A pesar de eso, no hay que olvidar, que el conocimiento biológico no es completo. En consecuencia, cuando se utilizan validaciones biológicas y no aparecen resultados llamativos, podría deberse a que la información disponible no es suficiente.

A continuación se resume brevemente alguna de las fuentes y bases de datos de información y anotación biológica.

2.5.1. Información sobre genes

En relación a los genes existen bases de datos en las que es posible consultar información sobre sus características básicas: nombre, descripción, organismo localización, secuencia, funciones biológicas asociadas, artículos científicos relacionados, etc. La dos bases de datos de genes más importante son Entrez Gene **[Magglott et al, 2005]** del NCBI (*National Center for Biotechnology Information*) y Ensembl **[Hubbard et al, 2002]**, del Instituto Sanger. Aunque este tipo de información es bastante estable, pueden existir cambios. En la figura 2.8 se muestra la evolución de las entradas de Ensembl desde 1985.

Para relacionar los genes con diferentes conceptos, procesos y funciones biológicas aparecen las anotaciones. Una de las anotaciones actualmente más aceptada es la Ontología de Genes (GO, *Gene Ontology*) **[Ashburner et al, 2000]**. GO está basado en el desarrollo de un vocabulario estructurado de términos biológicos, categorías GO, que son previamente definidos y tienen asignado un código o identificador único. Los términos están asociados por una clave de jerarquía ontológica, en la cual unos están relacionados con otros a modo de padres e hijos. Esta jerarquía es acíclica y un mismo término puede estar asociado a varias categorías más genéricas en esta jerarquía (es decir, puede tener varios términos "padre"). Los términos biológicos de GO son aplicables a los genes de todas las especies, con el propósito de anotar los genes de forma consistente y comparable en las diferentes bases de datos.

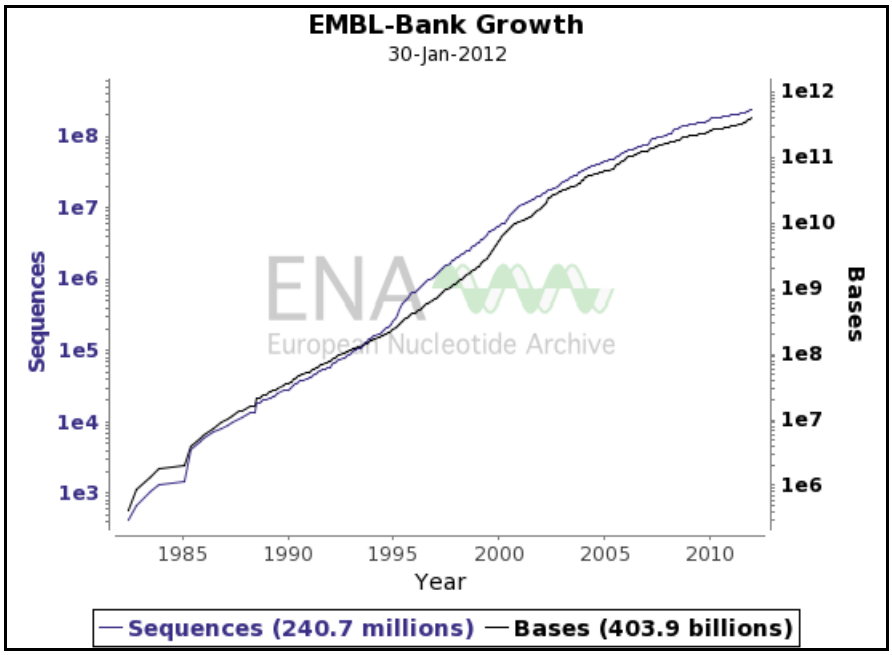


Figura 2.8. Crecimiento del número de entradas en la bases de datos Ensembl entre 1985 y enero de 2012 (fuente: http://www.ebi.ac.uk/ena/about/statistics#embl_growth).

Actualmente, GO está divida en tres sub-ontologías: **(i)** procesos biológicos (GO-BP, *Gene Ontology: Biological Process*), que contiene categorías relacionadas con los objetivos biológicos en los que el gen participa; **(ii)** funciones moleculares (GO-MF, *Gene Ontology: Molecular Function*), que define las actividades bioquímicas de cada gen, y **(iii)** componentes celulares (GO-CC, *Gene Ontology: Cellular Component*), relativas a los lugares de las células donde los genes están activos.

También existen bases de datos con información directamente relacionada con los genes, como son las bases de datos de proteínas, y las bases de datos que describen vías de señalización y vías metabólicas (*biological pathways*). UniProt **[The UniProt Consortium, 2012]** es la base de datos más importante que relaciona gen con proteínas. KEGG *(Kyoto*

Encyclopedia of Genes and Genomes) **[Kanehisa et al, 2006]**, es la mejor fuente para relacionar un gen con rutas metabólicas o de señalización celular. Existen otras bases de datos biológicas interesantes que complementan las citadas, como por ejemplo GAD (*Genetic Association Database*) **[Zhang et al, 2010]**, que relaciona genes con estudios de asociación genética de enfermedades complejas en humanos.

Todas estas fuentes son sólo una muestra muy pequeña de la gran cantidad de recursos, repositorios y bases de datos biológicas disponibles. Debido a esta diversidad, de cara a estudios bioinformáticos y bioestadísticos, se genera un problema con la gran heterogeneidad de "identificadores" de genes, lo que dificulta el acceso e integración de la información. Como ejemplo, la tabla 2.1 contiene una muestra de los diferentes nombres que recibe el gen HBA1. En este trabajo se ha identificado cada gen utilizando el código proporcionado en Ensembl junto con el nombre oficial del gen. Puesto que los datos de partida están en el formato de *Affymetrix*, se ha utilizado el re-mapeo proporcionado por GATExplorer (*Genomic and Transcriptomic Explorer*) **[Risueño et al, 2010]**, que relaciona cada sonda (*probe*) de un microarray de *Affymetrix* con un gen, logrando un re-mapeo de cada conjunto de sondas específico a la entidad biológica correspondiente.

Tabla 2.1. Heterogeneidad en los identificadores del gen HBA1

Fuente	Identificador
Nombre	HBA1 - Hemoglobin, alpha 1
NCBI	3039
Ensembl	ENSG00000206172
KEGG	hsa:3039
Affymetrix	204018_x_at, 209458_x_at, 211699_x_at, 211745_x_at, 214414_x_at, 217414_x_at
GAD	124933

2.5.2. Información sobre los experimentos

La gran diversidad de formatos de microarrays y tipos de experimentos hace muy complicado establecer un *gold* estándar en cuanto a sistemas de bases de datos. Lo que sí existe, es un acuerdo acerca de la mínima información sobre un experimento de microarrays que necesita ser almacenada para interpretar los resultados sin ambigüedades y reproducir el experimento. Es lo que se conoce como MIAME *(Minimum Information About a Microarray Experiment*) **[Brazma et al, 2001b]**.

En cuanto a la disponibilidad de los datos, desde la aparición de los microarrays, la comunidad biológica está de acuerdo en que los datos de experimentos publicados deben estar disponibles públicamente. Esto ha creado la necesidad de que existan repositorios de microarrays públicos, donde cualquier usuario puede almacenar sus datos. Al mismo tiempo esto hace que exista una gran cantidad de datos disponibles para re-analizarlos por cualquiera que desee hacerlo. Uno de los repositorios más utilizados es GEO *(Gene Expression Omnibus*) **[Edgar et al, 2002]**.

2.5.3. Validación de listados de genes a través del análisis de la información biológica contenida

Hay algunos métodos y modelos para procesar listas de genes intentando relacionarlos con anotaciones funcionales **[Khatri y Drâghici, 2005]**. Habitualmente, se lleva a cabo un análisis de enriquecimiento funcional (FEA, *Functional Enrichment Analysis*), que se puede realizar por varios métodos. En este tipo de análisis se busca un enriquecimiento estadístico de los genes, tratando de establecer si una categoría dada, que representa por ejemplo, un proceso biológico de GO o una vía metabólica de KEGG, aparece de manera más frecuente en la lista de genes seleccionada que en la población de la que procede (genoma, array o conjunto de genes). Habitualmente, se ofrece un p-valor para cada categoría biológica basado en la distribución hipergeométrica, aunque existe cierta controversia acerca de lo apropiado de este tipo de contrastes **[Hubbard, 2006]**; **[Anderson et al, 2000]**.

Actualmente existen varias herramientas *on-line* que permiten hacer este tipo de análisis. Una de las más utilizadas es DAVID (*Database for Annotation, Visualization and Integrated Discovery*) **[Huang da et al, 2009]**.

2.6. El proyecto Bioconductor

El crecimiento del uso de técnicas genómicas de gran escala, como los microarrays, en numerosos estudios biológicos y biomédicos, ha provocado la aparición de proyectos que pretenden desarrollar nuevos métodos para modelizar y analizar los datos. Para implementar estos métodos, así como para almacenar, acceder y organizar la gran cantidad de datos disponibles, han sido necesarias nuevas herramientas bioinformáticas, computacionales y estadísticas.

En el caso de los datos de expresión génica y datos obtenidos con microarrays, para realizar un análisis avanzado se necesitan dos elementos fundamentales: **(i)** disponer de repositorios o bases de datos públicas abiertas donde se almacenen dichos datos (como los que se han citado en la sección 2.5), y **(ii)** disponer de software y entornos de desarrollo avanzado donde se puedan implementar o usar distintas herramientas y algoritmos computacionales y estadísticos de análisis de datos.

Existen decenas de programas o plataformas disponibles, ya sea vía *web* o que se pueden instalar localmente (*stand-alone programs*), para trabajar con datos de microarrays. Ver http://www.nslij-genetics.org/microarray/soft.html para una extensa clasificación. Estas herramientas, aunque fáciles de utilizar, no suelen ser muy útiles para enfrentarse a análisis complejos, bien porque no tienen implementados muchos métodos, o simplemente porque no es posible automatizar tareas o desarrollar nuevos métodos basados en código propio.

Frente a estos programas cerrados existe la posibilidad de analizar datos biológicos complejos utilizando software estadístico y lenguajes de programación avanzado, como Matlab o R, con librerías específicamente diseñadas para el análisis de datos de expresión genética. Aunque existen algunas extensiones para Matlab, como por ejemplo MatArray **[Venet, 2003]**, es R la herramienta utilizada mayoritariamente por ser de código abierto (*open-source*) y además gratuita.

El proyecto Bioconductor (http://www.bioconductor.org) comenzó en 2001 como un proyecto de desarrollo de software abierto para el análisis e interpretación de datos genómicos, basado en R. Actualmente, prácticamente la mayoría de los métodos disponibles en análisis de microarrays tiene su propio paquete en este entorno. Todos los métodos, análisis y resultados mostrados en este trabajo, se han desarrollado en este marco.

Capítulo 3

Análisis de expresión en múltiples muestras y múltiples clases: búsqueda del núcleo

El análisis de expresión génica en una multiplicidad de condiciones es un área en creciente desarrollo como demuestran, entre otros, los trabajos de **[Spencer et al, 2011]**; **[Bamps y Hope, 2008]**; **[Fowlkes et al, 2008]**; **[Tomancak et al, 2007]**; **[Su et al, 2004]**. Los datos a analizar se corresponden con niveles de expresión de multitud de genes tomados en muchas condiciones biológicas distintas. La hipótesis de partida en este capítulo, corresponde a, en estos casos, para un elevado porcentaje de genes existirá un grupo mayoritario de observaciones, no necesariamente las mismas, que describen un patrón general y, que podrá existir o no, otro conjunto más reducido que mostrará un comportamiento diferenciado, correspondiendo a observaciones que están sobre-expresadas o infra-expresadas respecto del nivel típico de expresión de cada gen. Esta situación aparece, por ejemplo, en los genes relacionados con el funcionamiento celular y los ligados a muchas enfermedades en tejidos humanos **[Liang et al, 2006]**.

Bajo la mencionada hipótesis, tendrá interés, para cada gen, caracterizar el patrón de funcionamiento *normal.* Esto permitiría detectar y cuantificar los comportamientos atípicos, posiblemente ligados a patologías, tejidos biológicos, momentos temporales, etc. La existencia de estos comportamientos amplía el concepto de dato atípico, inicialmente atribuido a errores de medición o a artefactos en el pre-procesado de los datos, para incluir también a los de origen biológico: grupos de muestras con respuesta genética heterogénea. Este tipo de patrón es común, por ejemplo, en los datos de expresión ligados al cáncer, donde los oncogenes pueden estar activados en algunos, pero no necesariamente en todos los individuos enfermos. En términos estadísticos, nuestro interés estará centrado en estimar la distribución de las observaciones que definen el patrón de comportamiento mayoritario. Para ello son necesarios estimadores robustos que no se vean influenciados por la asumida contaminación contenida en las observaciones.

Posiblemente el estimador robusto más simple de localización, en el caso unidimensional, es la media recortada, que elimina del análisis una proporción prefijada de las observaciones

con valores más altos y más bajos. Este estimador generalizaría la mediana, paradigma de la estimación robusta, que aparecería en el límite, cuando el porcentaje de observaciones a eliminar por cada lado se aproxima al 50%. Basada también en una idea de contaminación simétrica, aparece como estimador robusto de localización la media biponderada de Tukey **[Mosteller y Tukey, 1977]**, que ha sido muy utilizada para resumir datos de expresión génica y es parte esencial de uno de los procedimientos de normalización más utilizados, el *Affymetrix Global Scaling* **[Affymetrix, 2002]**. En el análisis de datos de microarrays, ha sido habitual reemplazar la media y la desviación típica por medidas robustas de localización y escala como la mediana, la mencionada media recortada o el rango intercuartílico. Recientemente han sido propuestos, como aproximación robusta a la estimación conjunta de localización y escala en la búsqueda de expresión diferencial, estimadores recortados que eliminan observaciones utilizando las definiciones de atipicidad desarrolladas para el diagrama de cajas (*boxplot rules*) **[Gleiss et al, 2011]**.

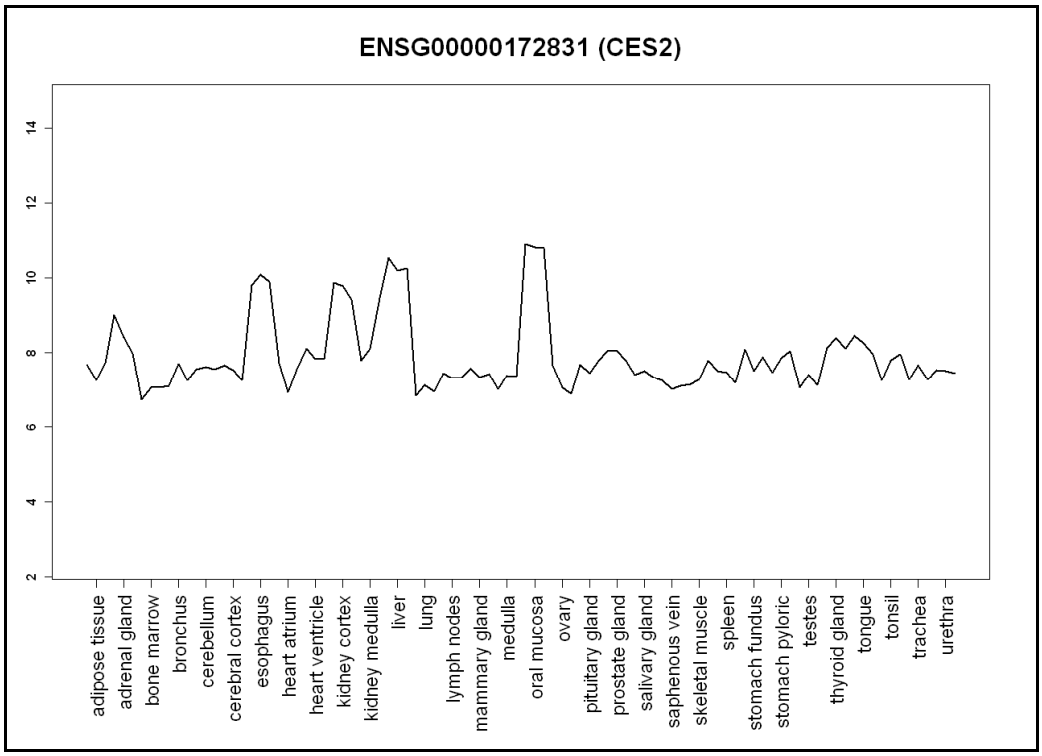


Figura 3.1. Ejemplo del perfil de expresión de un gen del *dataset* de Tejidos Humanos.

Todas estas propuestas mencionadas, están basadas en asumir un comportamiento simétrico para la contaminación, bien porque eliminan el mismo porcentaje de observaciones por arriba y por abajo, o porque consideran como observaciones contaminantes las que distan más de una cantidad del percentil 25 o del 75 o de la mediana. Bajo este postulado, incluso medias que eliminan un 50% de datos, en una situación que incluya una contaminación unilateral próxima al 50%, podrían estar basadas hasta en casi un 25% de atípicos. En los datos de expresión génica, que aparecen en microarrays, es relativamente frecuente observar contaminación solo en una dirección, más frecuentemente correspondiente a sobre-

expresión. La figura 3.1 muestra un ejemplo de esta situación, correspondiente a la expresión observada en un gen del *dataset* de Tejidos Humanos descrito en el apéndice B (B.1).

Este tipo de problema llevó a **[Gleiss et al, 2011]** a proponer una metodología que trata de encontrar el recorte unilateral óptimo, entre todos los posibles, en el sentido del mejor p-valor observado en un contraste de dos poblaciones. Como no se dispone de algoritmo para la búsqueda de ese óptimo, la aplicación de esta metodología es solo posible para conjuntos de datos pequeños, que todavía son muy frecuentes, como el que ellos utilizaron de solo 10 muestras. En ellas tratan de encontrar el recorte unilateral óptimo entre los correspondientes a eliminar 0, 1 o 2 arrays.

Nuestra propuesta para estimar el núcleo de expresión estará basada en el uso de recortes imparciales **[Gordaliza, 1991]**. Utilizando estimadores basados en esta idea, serán los propios datos los que nos guíen en la elección de la mejor forma de recortar. Dentro de esta familia de estimadores, la media recortada imparcial de nivel α aparece como solución del problema de búsqueda conjunta de la mejor estimación para la localización de las observaciones y del mejor conjunto conteniendo una proporción $1-\alpha$ de ellas (del mejor candidato a conjunto para las observaciones *buenas*). A diferencia de los anteriores estimadores mencionados, una media recortada imparcial con un nivel de recorte próximo al 50% podría resistir una contaminación unilateral cercana a ese tamaño. Este estimador es la concreción univariante del estimador MCD (Mínimo Determinante de la matriz de Covarianzas) **[Rousseeuw, 1985]** para localización y dispersión multivariante y tienen probada utilidad para localizar una región central con alta densidad de observaciones no contaminadas. Un problema no resuelto con la media recortada univariante, y en general con el MCD, es el de la estimación insesgada (en sentido asintótico; en adelante, cuando nos refiramos a la insesgadez siempre será como una propiedad asintótica correspondiente a la definición estadística de consistencia) de la variabilidad del núcleo no contaminado. Si se utilizara para ello la desviación típica de los datos no recortados, cuando haya menos contaminación que el nivel de recorte utilizado se infra-estimaría la variabilidad. Para evitar esto, es muy habitual utilizar la desviación típica anteriormente señalada, multiplicada por una constante elegida para conseguir, bajo el modelo normal de no contaminación, eliminar este sesgo. Pero el uso de este estimador modificado por la constante, producirá sobre-estimación de la variabilidad en el caso de que la muestra contenga contaminación.

Debido a la necesidad de estimadores insesgados para la variabilidad, independientemente del grado de contaminación existente en la muestra, propondremos para estimar la distribución del núcleo de expresión el estimador smart **[Cuesta-Albertos et al, 2008]**, que también pertenece a la familia de estimadores basados en recortes imparciales. Este estimador es insesgado para la media y la desviación típica del núcleo de la distribución, asumiendo distribución normal para las observaciones genuinas, siempre que la contaminación, independientemente de su tamaño, sea exterior a la zona central de esta

distribución. El estimador se construye en dos pasos, el primero corresponde a la obtención de la zona más informativa de la distribución, la relacionada con la parte central del núcleo, y en el segundo paso se estima la media y la desviación típica del núcleo utilizando únicamente las observaciones en la zona obtenida en el primer paso y censurando o truncando las observaciones exteriores a ésta zona, según el propio estimador decida como más conveniente. Como sub-producto, este estimador permite obtener el porcentaje de observaciones contaminantes en la muestra, o lo que es equivalente, el porcentaje de observaciones genuinas. Aunque este estimador tiene probadas propiedades estadísticas y de robustez, no ha sido nunca utilizado en el análisis de datos de expresión génica.

Para la obtención de las estimaciones hemos adaptado el algoritmo disponible para la versión multivariante **[Cuesta-Albertos et al, 2008]** a la situación univariante que se presenta en la estimación del comportamiento del núcleo de un gen. Hemos escogido, para esta aproximación que corresponde a una situación univariante, la versión del estimador smart que utiliza como estimador (de primer paso), para conseguir la zona de observaciones *buenas*, la media recortada imparcial. El análisis del algoritmo para esta situación, en la que su primer paso estaría basado en la concreción univariante del MCD, nos ha permitido probar la equivalencia de los estimadores de localización y de los parámetros de forma de la matriz de covarianzas del estimador smart basado en el MCD (smart-MCD) con los del propio MCD. Esto facilita la estimación para esta versión del smart en el caso univariante, pero también simplifica en gran medida la estimación del smart-MCD multivariante, ya que reduce el segundo paso de la estimación, a la búsqueda de un único parámetro, el correspondiente al tamaño de la variabilidad (el determinante de la matriz de covarianzas). De esta forma, la estimación corresponde a la obtención de la media recortada imparcial en el primer paso y a una sustitución de la estimación iterativa del segundo paso, por la búsqueda del máximo de una función en un intervalo. Esta consecución supone un gran avance en la dirección de hacer factible la obtención de estimaciones del smart-MCD en tiempo razonable, y permite su utilización en el análisis de las matrices de datos de expresión génica donde, en cada conjunto de datos, es necesario aplicarlo masivamente, tanto por la elevada cantidad de genes como por el hecho de que la obtención de p-valores está basada en re-muestreo.

En este capítulo se pretende obtener el *core,* el núcleo de expresión de cada gen, a través de la determinación de los parámetros de localización y variabilidad de su distribución. Adicionalmente dispondremos de una estimación del porcentaje de contaminación para cada gen, que, en este caso, corresponderá al porcentaje de condiciones analizadas que difieren del nivel de expresión común del gen, y dispondremos de estimaciones de *calidad* para otros parámetros muy utilizados en este ámbito, como el coeficiente de variación del núcleo del gen. La estimación del núcleo, permitirá traducir en puntuaciones estandarizadas por gen, los valores de expresión observados en cada muestra. Estos valores estandarizados permitirán medir el nivel de atipicidad de cada individuo, o convenientemente resumidos, podrán servir también, para caracterizar a grupos de individuos en relación a su comportamiento típico.

3.1. Búsqueda del núcleo central

3.1.1. Estimador smart

El estimador smart **[Cuesta-Albertos et al, 2008]** es un estimador máximo verosímil para los parámetros de un modelo normal multivariante contaminado. La contaminación asumida por este modelo puede tener distribución libre pero el soporte de ésta, tiene que ser exterior a la zona central, a la zona más informativa de la distribución. El estimador se obtiene en dos pasos, en un primer paso se estima la localización de la zona más informativa de la distribución normal y en un segundo paso, mediante máxima verosimilitud aplicada a las observaciones en dicha zona, se estiman los parámetros de la distribución.

Como estimadores para el primer paso se pueden utilizar el, ya mencionado, MCD o el MVE (Elipsoide de Mínimo Volumen), ambos con un nivel de recorte, α, elegido para ser superior a la proporción esperada de contaminación. Estos estimadores están basados en recorte imparcial y buscan la zona más informativa (con las observaciones más concentradas), conteniendo una proporción $1-\alpha$ de los datos de una distribución multivariante. El MCD, busca la zona de menor variabilidad en el sentido de que el determinante de la matriz de covarianzas sea mínimo, y el MVE, en el sentido del volumen del elipsoide más pequeño que contenga las observaciones. En los dos casos, las correspondientes zonas obtenidas son elipsoides que contienen la proporción de observaciones $1-\alpha$ prefijada. Los dos estimadores son robustos, en el sentido de que soportarían, cuando el recorte aplicado es próximo a α, hasta una proporción cercana a α de observaciones contaminantes colocadas en cualquier posición, sin que las estimaciones respondan arbitrariamente. Por tanto, situando el nivel de recorte en el 50% pueden soportar un porcentaje de contaminación próximo al 50%. Este funcionamiento está garantizado en su aplicación a matrices de datos típicas de análisis multivariante (con $n >> p$, siendo n y p el número de individuos y de dimensiones respectivamente). El MCD es más eficiente (menor varianza), su distribución converge a una normal al aumentar el tamaño muestral, pero cuando se incrementa el número de dimensiones, el sesgo máximo, el máximo daño que un porcentaje de observaciones contaminantes puede hacerle, crece muy rápido. En el sentido de los algoritmos disponibles para ambos, el correspondiente al MCD es muy superior al del MVE. Por el contrario, el MVE tiene una baja eficiencia pero es muy interesante por su bajo sesgo máximo en un alto número de dimensiones. Todo ello hace recomendable la utilización del MVE en aplicaciones con un elevado número de dimensiones, pero cuando el número de dimensiones es muy bajo, el mejor estimador es el MCD. Los parámetros de localización y forma de estos estimadores determinan una zona central conteniendo la proporción de observaciones prefijada, que será la base del segundo paso de la estimación smart.

Este segundo paso del estimador smart corresponde a la estimación máximo verosímil de los parámetros de la distribución normal basada únicamente en las observaciones contenidas en

la zona obtenida por el primer paso de estimación. Bajo la suposición de que todas las observaciones en la zona elegida en el primer paso son legítimas, es decir, provienen del modelo normal, para realizar la estimación en este segundo paso, se debería escoger entre el estimador del modelo normal censurado, que adicionalmente a las observaciones en la zona, utiliza también un porcentaje de observaciones observado en el exterior de la zona; y el estimador truncado, que solo utiliza las observaciones en la zona y olvida el número de observaciones exterior. Si no hay contaminación, el mejor estimador es el censurado porque la información adicional que utiliza, el porcentaje de observaciones exteriores, es correcta. Si hay contaminación exterior, el mejor es el truncado, porque no utiliza esta información adicional que, bajo este supuesto, es errónea. El estimador smart lleva incorporada una estimación del porcentaje de contaminación en el exterior de la zona utilizada, y según esta estimación sea mayor o menor que 0 escoge truncar o censurar. El hecho de que este estimador escoja automáticamente entre censurar o truncar, gestionando de la manera más conveniente la utilización del porcentaje de observaciones exteriores a la zona, justifica su nombre como estimador inteligente o smart.

Dada una muestra de observaciones $y_1,...,y_n$ y un conjunto que contiene una proporción $1-\alpha$ de las observaciones, A_α, la función objetivo del estimador smart es la log-verosimilitud de un modelo normal, contaminado por una distribución arbitraria cuyo soporte es exterior al conjunto A_α, que solo obtiene información completa de las observaciones en ese conjunto A_α, y cuyo peso representa una proporción π en la mezcla. Esta función objetivo viene dada por la expresión,

$$\sum_{i=1}^{n}\left(I_{A_\alpha}(y_i)\log\left((1-\pi)\varphi_{\mu,\sigma}(y_i)\right)+(1-\pi)\log p_{\mu,\sigma}(A_\alpha)+\pi\right) \quad con \quad \pi>0 \qquad (3.1)$$

donde $\varphi_{\mu,\sigma}()$ y $p_{\mu,\sigma}()$ son, respectivamente, las funciones de densidad y de probabilidad asignada a un conjunto para una distribución normal con parámetros μ y σ, y, π es la proporción de observaciones contaminantes.

Este estimador, como se ha mencionado, coincide, cuando la estimación de la proporción de contaminación π es igual a 0, con el estimador para el modelo normal que censura las observaciones fuera de A_α y, cuando es mayor que 0, con el estimador truncado que elimina completamente la información fuera de A_α.

El estimador smart tiene probadas propiedades de robustez, heredadas del estimador elegido para la estimación del conjunto A utilizado. Como ya se ha señalado, el estimador smart es insesgado para la media, pero también, como característica inusual en un estimador robusto, es también insesgado para la variabilidad, independientemente del grado de contaminación existente en la muestra. Para obtener estimaciones se necesitan algoritmos para el primer

paso (MCD o MVE) y para el segundo paso. Para el MCD y el MVE existen algoritmos disponibles en R. Para el segundo paso, se puede utilizar el algoritmo descrito en **[Cuesta-Albertos et al, 2008]**. Este algoritmo, por la necesidad de computar probabilidades normales multivariantes utilizando el método de Monte Carlo, no es tan eficiente como el disponible para el MCD. En su aplicación a situaciones univariantes, al disponer de funciones que, de forma explícita, nos permiten calcular una probabilidad normal en intervalos, este problema no aparece.

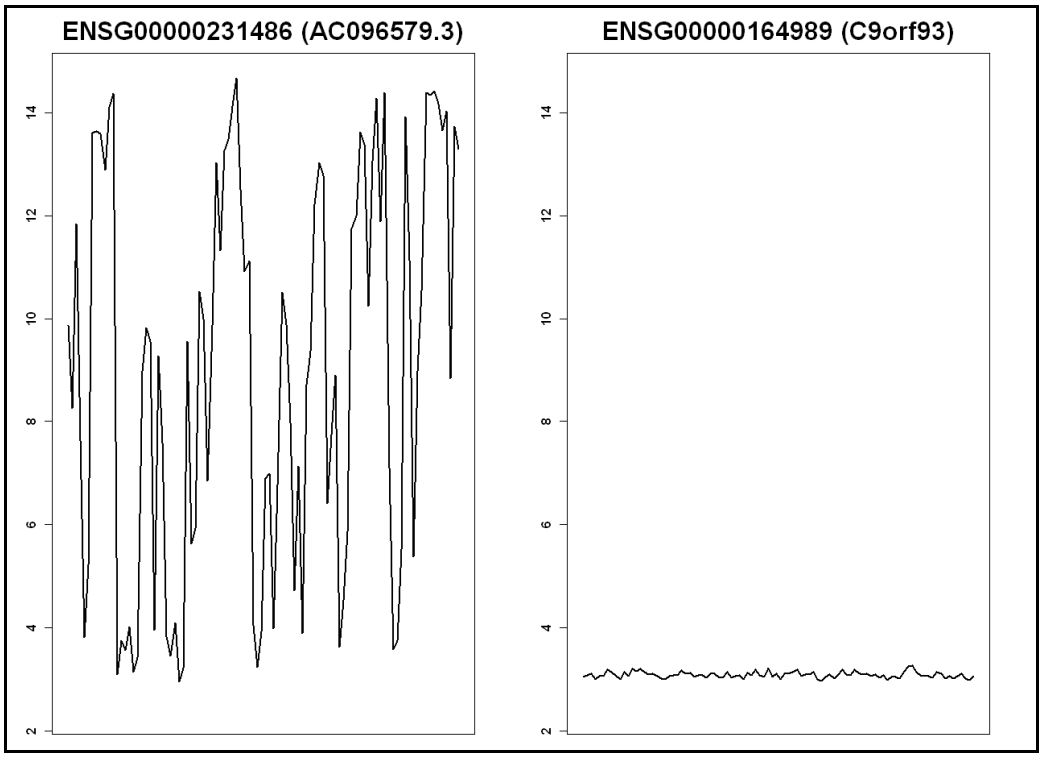


Figura 3.2. Ejemplo del perfil de expresión de dos genes pertenecientes al *dataset* de Tejidos Humanos.

3.1.2. Estimador smart para el núcleo de un gen

Encontrar el comportamiento típico de un gen en cuanto a su nivel de expresión, es de gran utilidad. En una matriz de expresión conviven genes con distribuciones de expresión muy diferentes. A modo de ejemplo, en la figura 3.2 se representan los patrones de expresión de dos genes pertenecientes al mismo conjunto, el *dataset* de Tejidos Humanos. Uno de ellos tiene niveles de expresión en toda la escala de posibles valores frente al otro gen, con niveles de expresión muy concentrados alrededor de su valor central. Para poder situar convenientemente el valor de expresión observado en un individuo respecto de la distribución correspondiente a un gen, es necesario estimar el comportamiento típico de ese gen.

Basar esta distribución en la media y la desviación típica muestral puede tener efectos indeseables, como el mostrado en la figura 3.3, donde claramente se observan varias observaciones con niveles de expresión muy por encima del nivel de expresión mayoritario, y que, sin embargo, se encuentran a menos de tres desviaciones típicas de la media. Este patrón de expresión no es un caso aislado. Por el contrario, es muy frecuente en conjuntos de datos donde el número de condiciones biológicas diferentes analizadas es grande.

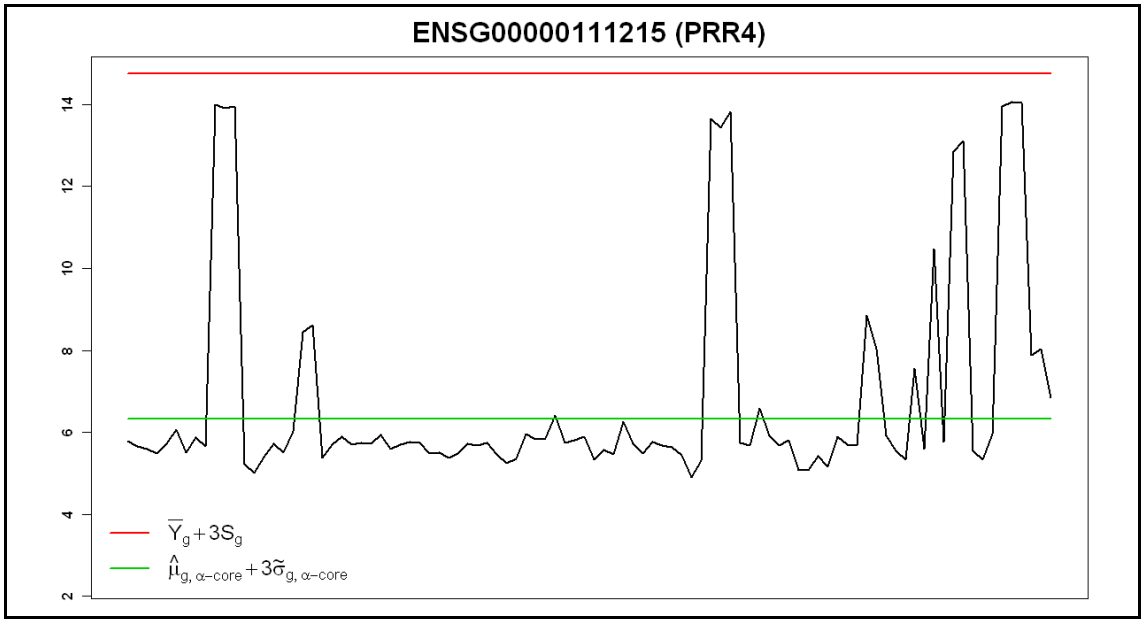


Figura 3.3. Perfil de expresión de un gen con varias muestras sobre-expresadas entre los 32 tejidos humanos sanos del *dataset* de Tejidos Humanos.

La aplicación del estimador smart para identificar el núcleo de comportamiento de un gen, a partir de un conjunto de observaciones o muestras obtenidas en diferentes condiciones, requiere de la versión univariante del estimador. Para esta aplicación, como estimador del primer paso escogemos el MCD, ya que es un estimador más eficiente que el MVE, dispone de algoritmo más eficiente y su peor sesgo máximo en altas dimensiones no va a tener efecto en esta aplicación a una situación univariante. A la versión del estimador smart que utiliza el MCD como zona inicial la denominamos smart-MCD.

La obtención del smart-MCD se beneficia de la simplificación ofrecida por el teorema 1, enunciado más adelante: la estimación de la media smart coincide con la media recortada imparcial, que es la concreción del MCD a una situación univariante. Por ello, la estimación del smart-MCD, una vez obtenido el MCD, corresponderá a la obtención de la desviación típica, ya que **[Cuesta-Albertos et al, 2008]** probaron que el otro parámetro π, se puede obtener como función de σ, y que su óptimo se alcanza en $\pi = \dfrac{1-\alpha-p_{\mu\sigma}(A)}{p_{\mu\sigma}(A)}$ (con la notación de (3.1)), si esta cantidad es positiva, o en $\pi = 0$, en caso contrario.

Para la estimación del MCD con recorte α, que en el caso univariante corresponde a la media recortada imparcial con el mismo nivel de recorte, utilizamos el algoritmo FastMCD **[Rousseeuw y Van Driessen, 1999]**, implementado en la librería robustbase **[Rousseeuw et al, 2011]** disponible en R. La zona que determina las observaciones incluidas en la estimación del segundo paso corresponde a la proporción $1-\alpha$ de observaciones más próximas a la media recortada imparcial. Para la estimación de la desviación típica smart, buscamos el máximo de la función en (3.1), utilizando una rejilla de valores para σ, el intervalo delimitado por 0 y la desviación típica muestral. La estimación smart-MCD de la desviación típica será el valor de σ para el que se alcance el máximo de la función objetivo, en la rejilla antes mencionada, utilizando cualquiera de las dos versiones mencionadas para π.

Uno de los principales inconvenientes del método propuesto, es la elección del nivel de recorte. Un nivel de recorte equivocado, por debajo del nivel de contaminación, producirá sesgo en el estimador mientras que un nivel de recorte inferior al nivel de contaminación aumentará la variabilidad de los estimadores. Entre los dos errores, el que tiene como efecto el aumento del sesgo es el menos deseado. El otro efecto, el de la reducción de la eficiencia del estimador, se puede suplir aumentando el número de observaciones. Por ello, siempre se debería elegir el nivel de recorte superior al nivel de contaminación esperado.

En esta memoria, cuando hagamos mención de la utilización del estimador smart en aplicaciones a matrices de expresión génica, siempre nos estaremos refiriendo al smart-MCD, aunque no lo mencionemos.

Teorema 1. La estimación de la media y la matriz de forma correspondiente estimadores smart-MCD coincide con la media y la matriz de forma del MCD.
Demostración. La demostración está incluida en el Apéndice A.

3.2. Funcionamiento del método

3.2.1. Datos simulados

Generamos $n(=100)$ observaciones de una mixtura de normales, $Y \to \omega N(\delta,1)+(1-\omega)N(0,1)$, con $\omega \in \{0, 0.1, 0.3, 0.45\}$, la proporción de contaminación correspondiente a sobre-expresión, y $\delta \in \{0,2,4,6\}$, la cantidad sobre-expresada. Para tener en cuenta la variabilidad del método, se simulan $n_{sim}(=150)$ conjuntos para cada combinación.

Localización

Media muestral, $\overline{Y}$

Mediana, Me

Media biponderada de Tukey [Mosteller y Tukey, 1977], definida como,

$$Tb = \frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i}$$

con $w_i = \begin{cases} \left(1-u_i^2\right)^2 & \text{si } |u_i| \le 1 \\ 0 & \text{si } |u_i| > 1 \end{cases}$, i=1,...,n , $u_i = \frac{y_i - Me}{c \cdot mad}$, $mad = Mediana\left(\left\{|y_i - Me|;\ i = 1,...,n\right\}\right)$ y c es una constante establecida en función de la resistencia a outliers que se desee obtener. Se ha establecido $c = 5$, el valor utilizado en el procedimiento de normalización de microarrays MAS5 **[Hubbell et al, 2002].**

Media $\alpha/2$ ***-recortada no-imparcial***, $\overline{Y}_{\alpha/2}$, definida como la media de las $\left[(1-\alpha)\cdot n\right]$ observaciones centrales,

$$\overline{Y}_{\alpha/2} = \frac{\sum_{i=\left[\alpha/2 \cdot n\right]+1}^{n-\left[\alpha/2 \cdot n\right]} y_{(i)}}{\left[(1-\alpha)\cdot n\right]}$$

con $\left[a\right]$ el entero más próximo a a

Dispersión

Varianza muestral, S_Y^2

Varianza basada en Tukey**,** ponderada utilizando los pesos estimados en el cálculo de la media biponderada de Tukey,

$$\hat{\sigma}_{Tb}^2 = \frac{\sum_{i=1}^{n} w_i}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n} w_i^2} \sum_{i=1}^{n} w_i \left(y_i - Tb\right)^2$$

Varianza basada en el MAD, definida como,

$$\hat{\sigma}_{mad}^2 = \left(MAD\right)^2$$

donde MAD se calcula como la mediana de las desviaciones respecto de la mediana, en valor absoluto, y multiplicada por la constante 1.4826, para conseguir que sea un estimador insesgado para la desviación típica en el modelo normal.

Figura 3.4. Estimadores de localización y dispersión utilizados en las comparaciones con el estimador smart.

En la tabla 3.1 comparamos el estimador $\hat{\mu}_{\alpha-smart}$ que resulta del α-recorte imparcial de $Y = \left\{y_1, y_2, ..., y_n\right\}$ variando α en el conjunto $\alpha \in \left\{0.1, 0.3, 0.45\right\}$, con los estimadores de localización descritos en la figura 3.4. Consideramos, la media muestral, la mediana, la

media bi-ponderada de Tukey y la media $\alpha/2$-recortada no imparcial, con $\alpha \in \{0.1, 0.3, 0.45\}$.

Tabla 3.1. Datos simulados, estimación de μ_{core}

ω	δ	$\bar{Y}$	Me	Tb	$\bar{Y}_{\alpha/2}$			$\hat{\mu}_{\alpha-smart}$		
					$\alpha=0.1$	$\alpha=0.3$	$\alpha=0.45$	$\alpha=0.1$	$\alpha=0.3$	$\alpha=0.45$
0	**0**	0.01 ± 0.10	0.00 ± 0.13	0.00 ± 0.12	0.00 ± 0.11	0.00 ± 0.11	0.00 ± 0.12	0.00 ± 0.14	-0.02 ± 0.17	-0.02 ± 0.22
0.1		0.19 ± 0.11	0.12 ± 0.14	0.11 ± 0.12	0.16 ± 0.11	0.14 ± 0.11	0.13 ± 0.11	0.09 ± 0.13	0.08 ± 0.17	0.08 ± 0.25
0.3	**2**	0.62 ± 0.10	0.50 ± 0.14	0.53 ± 0.12	0.59 ± 0.10	0.56 ± 0.11	0.54 ± 0.11	0.50 ± 0.14	0.39 ± 0.22	0.36 ± 0.31
0.45		0.91 ± 0.10	0.87 ± 0.14	0.89 ± 0.11	0.90 ± 0.10	0.89 ± 0.11	0.89 ± 0.11	0.87 ± 0.15	0.81 ± 0.26	0.82 ± 0.37
0.1		0.40 ± 0.09	0.13 ± 0.12	0.05 ± 0.11	0.29 ± 0.10	0.17 ± 0.10	0.15 ± 0.10	0.02 ± 0.11	0.02 ± 0.16	0.06 ± 0.21
0.3	**4**	1.2 ± 0.1	0.57 ± 0.16	0.65 ± 0.18	1.13 ± 0.10	0.97 ± 0.11	0.83 ± 0.12	0.78 ± 0.11	0.05 ± 0.14	0.06 ± 0.21
0.45		1.8 ± 0.09	1.28 ± 0.22	1.64 ± 0.12	1.78 ± 0.1	1.74 ± 0.1	1.7 ± 0.11	1.61 ± 0.19	0.79 ± 0.33	0.10 ± 0.25
0.1		0.60 ± 0.09	0.13 ± 0.12	0.03 ± 0.10	0.40 ± 0.09	0.17 ± 0.10	0.15 ± 0.10	0.00 ± 0.09	0.01 ± 0.17	0.07 ± 0.21
0.3	**6**	1.8 ± 0.1	0.56 ± 0.15	0.31 ± 0.21	1.69 ± 0.10	1.4 ± 0.1	1.12 ± 0.11	1.22 ± 0.10	0.00 ± 0.11	0.02 ± 0.20
0.45		2.71 ± 0.11	1.34 ± 0.22	2.27 ± 0.19	2.69 ± 0.11	2.61 ± 0.12	2.54 ± 0.12	2.32 ± 0.18	1.08 ± 0.12	0.01 ± 0.13

Obviamente se observa una influencia del nivel de recorte. Cuando nos equivocamos en la aplicación del recorte, porque el recorte imparcial aplicado es inferior al nivel de contaminación, el estimador smart sobre-estima el verdadero parámetro, aunque el error cometido, será del mismo orden que el de otros estimadores robustos, como la mediana o la media bi-ponderada de Tukey. Cuando el recorte es mayor que la contaminación (señalado en verde en la tabla 3.1), la media imparcial α-recortada funciona. Las diferencias entre estimadores son más llamativas cuando la localización de la sobre-expresión, δ, es mayor. El recorte imparcial es muy superior cuando el nivel de sobre-expresión es alto.

Tabla 3.2. Datos simulados, estimación de σ^2_{core}.

ω	δ	S_y^2	$\hat{\sigma}^2_{T_b}$	$\hat{\sigma}^2_{mad}$	$\hat{\sigma}^2_{\alpha-smart}$		
					$\alpha=0.1$	$\alpha=0.3$	$\alpha=0.45$
0	**0**	0.99 ± 0.15	0.68 ± 0.13	0.98 ± 0.23	0.9 ± 0.16	0.78 ± 0.25	0.54 ± 0.29
0.1		1.35 ± 0.18	0.84 ± 0.17	1.2 ± 0.29	1.14 ± 0.22	0.94 ± 0.32	0.64 ± 0.35
0.3	**2**	1.86 ± 0.24	1.34 ± 0.22	1.95 ± 0.41	1.77 ± 0.26	1.49 ± 0.42	1.14 ± 0.54
0.45		2.01 ± 0.26	1.57 ± 0.23	2.39 ± 0.46	1.96 ± 0.25	1.84 ± 0.34	1.63 ± 0.57
0.1		2.44± 0.27	0.81 ± 0.19	1.3 ± 0.33	1.1 ± 0.21	0.97 ± 0.37	0.67 ± 0.42
0.3	**4**	4.39± 0.35	2.57 ± 0.5	3.31 ± 0.75	4.18 ± 0.41	1.19 ± 0.36	1.17 ± 0.76
0.45		5.01 ± 0.4	4.54 ± 0.39	8.12 ± 1.08	4.96 ± 0.4	4.49 ± 0.75	1.37 ± 0.59
0.1		4.28 ±0.41	0.76 ± 0.13	1.31 ± 0.29	1.06 ± 0.17	0.97 ± 0.32	0.74 ± 0.39
0.3	**6**	8.58± 0.57	2.05 ± 1.01	3.42 ± 0.94	8.41 ± 0.62	1.01 ± 0.2	1.17 ± 0.76
0.45		10.01 ± 0.63	9.06 ± 0.69	14.24 ± 2.81	9.91 ± 0.62	8.66 ± 1.3	0.9 ± 0.22

Considerando los mismos escenarios de simulación, en la tabla 3.2 se resumen cuatro estimaciones para el parámetro σ^2_{core}: S^2_Y, la varianza muestral; $\hat{\sigma}^2_{T_b}$, la varianza ponderada por los pesos estimados en el cálculo de la media biponderada de Tukey; $\hat{\sigma}^2_{mad}$ el cuadrado de la mediana de las desviaciones abolutas (MAD) multiplicada por la constate 1.4826 que aproxima a la desviación estándar en distribuciones normales; y $\hat{\sigma}^2_{\alpha-smart}$, el estimador smart variando el nivel de recorte en el conjunto $\alpha \in \{0.1, 0.3, 0.45\}$.

Una mala elección del nivel de recorte imparcial para el estimador $\hat{\sigma}^2_{\alpha-smart}$ va a tener el mismo efecto que el observado en las estimaciones de localización, ya que, necesariamente, intervendrán en la estimación observaciones contaminadas. Cuando el nivel de recorte es superior a la contaminación, el smart es el mejor estimador. El funcionamiento de los estimadores $\hat{\sigma}^2_{T_b}$ y $\hat{\sigma}^2_{mad}$ empeora al aumentar la cantidad de sobre-expresión. Un estimador considerado robusto, como el basado en el MAD, con niveles altos de sobre-expresión funciona peor que la varianza muestral. En la figura 3.5 se muestra la distribución de los cuatro estimadores para el caso en el que (a) $\alpha = 0.3$, $\omega = 0.3$ y $\delta = 4$, y (b) $\alpha = 0.45$, $\omega = 0.45$ y $\delta = 6$.

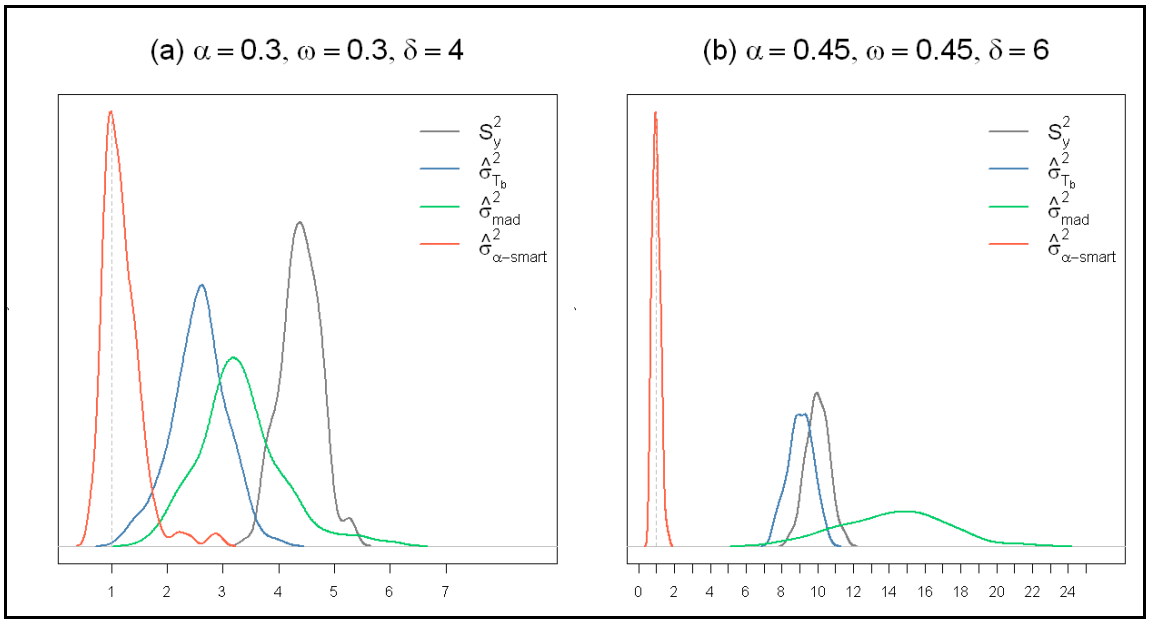


Figura 3.5. Distribución de los estimadores del parámetro σ^2_{core} con (a) $\alpha = 0.3$, $\omega = 0.3$ y $\delta = 4$ y (b) $\alpha = 0.45$, $\omega = 0.45$ y $\delta = 6$.

Otro aspecto que interesa evaluar es la identificación de las observaciones que van a pertenecer al núcleo de expresión mayoritario. En la tabla 3.3 se comparan, en términos del porcentaje de aciertos en la detección de observaciones sobre-expresadas, tres reglas de

clasificación: **(i)** basada en la media y desviación típica muestral, regla 2σ, que considera atípica a toda observación alejada de la media más de 2 desviaciones típicas; **(ii)** la regla basada en el *boxplot* considerando atípicos los valores fuera del intervalo percentil 75 (25) más (menos) f rangos intercuartílicos, con $f \in \{0.5, 1, 1.5\}$; y **(iii)** basada en el recorte imparcial.

Tabla 3.3. Datos simulados, porcentaje de acierto de observaciones sobre-expresadas.

ω	δ	Regla 2σ	Regla boxplot			Recorte imparcial		
			$f = 0.5$	$f = 1$	$f = 1.5$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.45$
0.1		29.8 ± 11.95	65.33 ± 16.04	37.07 ± 18.15	15.40 ± 14.7	41.07 ± 18.32	52.2 ± 22.49	59.4 ± 24.23
0.3	**2**	8.96 ± 3.75	31.82 ± 9.39	8.18 ± 5.51	1.33 ± 2.28	12.87 ± 6.76	24.53 ± 16.58	34.11 ± 23.39
0.45		3.99 ± 2.62	17.16 ± 6.31	2.56 ± 2.76	0.16 ± 0.64	4.96 ± 3.64	8.12 ± 7.43	17.9 ± 17.48
0.1		70.07 ± 12.56	98.87 ± 3.58	93.13 ± 8.91	76.73± 18.77	96.67 ± 6.09	97.6 ± 5.64	97.67 ± 5.95
0.3	**4**	7.87 ± 4.07	25.47± 11.35	1.42 ± 2.42	0.02 ± 0.27	20.16 ± 8.2	95.2 ± 6.18	92.27 ± 12.86
0.45		0.98 ± 1.47	3.78 ± 3.67	0.04 ± 0.31	0.00 ± 0.00	2.09 ± 2.35	18.56 ± 13.77	93.08 ± 11.94
0.1		91.67 ± 7.89	100±0.00	99.87 ± 1.15	99.67 ± 1.80	100 ± 0	99.93 ± 0.82	100 ± 0
0.3	**6**	4.38 ± 3.21	5.02 ± 6.18	0.00 ± 0.00	0.00 ± 0.00	14.64 ± 6.04	100 ± 0	99.93 ± 0.61
0.45		0.06 ± 0.36	0.16 ± 0.58	0.00 ± 0.00	0.00 ± 0.00	0.4 ± 0.89	19.04 ± 13.44	99.99 ± 0.18

La media y la desviación típica quedan influenciadas por la contaminación asimétrica que hemos generado, por lo que la regla 2σ va a funcionar muy mal. La regla basada en el *boxplot*, para porcentajes de contaminación bajos, $\omega = 0.1$, y niveles de recorte altos, que se corresponden con valores de f pequeños, va a funcionar de forma parecida a los recortes imparciales. La razón es que, en estos casos, los percentiles 25 y 75 no se van a ver muy afectados por los valores extremos. Sin embargo, con porcentajes de contaminación superiores al 25%, la detección de puntos extremos empeora sensiblemente. Además existe un efecto producido por la cantidad de sobre-expresión, contrario al que uno desearía, cuanto más alejadas se encuentran las muestras expresadas, peor son identificadas.

Tabla 3.4. Datos simulados, porcentaje de fallo de observaciones no-expresadas.

ω	δ	Regla 2σ	Regla boxplot			Recorte imparcial		
			$f = 0.5$	$f = 1$	$f = 1.5$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.45$
0.1		2.13 ± 1.36	14.26 ± 4.07	3.11 ± 2.33	0.48 ± 0.81	3.34 ± 1.86	6.93 ± 5.54	13.05 ± 9.87
0.3	**2**	1.72 ± 1.45	10.53 ± 4.36	1.44 ± 1.88	0.11 ± 0.42	1.72 ± 1.65	3.34 ± 3.76	7.68 ± 8.49
0.45		2.52 ± 1.75	12.75 ± 5.19	1.62 ± 1.81	0.13 ± 0.52	2.91 ± 2.6	3.68 ± 4.2	9.99 ± 11.52
0.1		0.3 ± 0.53	12.76 ± 4.35	2.61 ± 2.08	0.36 ± 0.63	3.42 ± 1.62	6.58 ± 5.11	12.94 ± 9.58
0.3	**4**	0.04 ± 0.23	2.23 ± 2.09	0.02 ± 0.16	0 ± 0	0.02 ± 0.16	3.26 ± 2.09	5.9 ± 5.81
0.45		0.27 ± 0.71	2.01 ± 2.21	0.01 ± 0.15	0 ± 0	0.25 ± 0.63	0.06 ± 0.33	3.12 ± 2.57
0.1		0.01 ± 0.09	13.38 ± 4.33	2.64 ± 2.14	0.37 ± 0.74	3.86 ± 1.66	6.99 ± 5.31	12.67 ± 10.1
0.3	**6**	0 ± 0	0.16 ± 0.51	0 ± 0	0 ± 0	0 ± 0	3.92 ± 1.65	6.35 ± 6.22
0.45		0 ± 0	0.18 ± 0.59	0 ± 0	0 ± 0	0 ± 0	0 ± 0	4.01 ± 1.91

Para completar la evaluación en la clasificación, en la tabla 3.4 se muestra el porcentaje observaciones no expresadas identificadas como *outlier* en cada uno de los escenarios. Obviamente, este porcentaje aumenta al aumentar el nivel de recorte. En los casos más extremos, la regla del *boxplot* con $f = 0.5$ y la regla basada en el recorte imparcial con $\alpha = 0.45$, las tasas de fallo son similares.

3.2.2. Aplicación al *dataset* de Tejidos Humanos

La matriz de expresión cuenta con 20172 genes procedente de 96 muestras de 32 tejidos humanos sanos, con 3 réplicas en cada uno de ellos. La descripción completa de este conjunto de datos puede consultarse en el apéndice B, sección B.1.

Para determinar el *core* de cada uno de los genes analizados, la primera decisión que tenemos que tomar está relacionada con la elección del nivel de recorte. Como hipótesis biológica asumimos que, como máximo, para la mayoría de los genes habrá 9 tipos de tejido de los 32 diferencialmente expresados respecto de su núcleo, por lo que el nivel de recorte se fija en el 30%. Aplicando el estimador smart a la matriz de datos de expresión obtenemos la estimación de la distribución de los estimadores de localización y dispersión que mostramos en la tabla 3.5.

Tabla 3.5. Distribución de los estimadores $\hat{\mu}_{0.3-smart}$ y $\hat{\sigma}_{0.3-smart}$ en los 20172 genes del *dataset* de Tejidos Humanos.

	Percentiles									
	0%	**10%**	**25%**	**50%**	**70%**	**75%**	**90%**	**95%**	**99%**	**100%**
$\hat{\mu}_{0.3-smart}$	2.67	3.52	4.27	5.43	6.34	6.58	7.55	8.22	9.92	13.84
$\hat{\sigma}_{0.3-smart}$	0.04	0.13	0.17	0.25	0.34	0.38	0.56	0.75	1.24	4.03

En este caso hemos utilizado el mismo nivel de recorte para todos los genes. En aquellos genes para los que el nivel de recorte empleado haya sido superior al nivel de contaminación, al nivel de sobre/infra-expresión, el estimador habrá funcionado como se espera. En el resto, aquellos en los que la contaminación es superior al 30% de nivel de recorte empleado, la estimación está basada no solo en los datos del núcleo, también en datos contaminados. Asumimos que estos últimos representan una proporción pequeña. La elección del nivel de recorte tiene que ser un balance entre la pérdida que podemos sufrir debida a sesgo en aquellos genes con mayor nivel de contaminación que de recorte y la pérdida debida a precisión que podemos sufrir en aquellos genes con mayor nivel de recorte que de contaminación. Aunque la primera de las pérdidas, la debida al sesgo, es mucho más indeseable, no podemos tampoco permitirnos aplicar un recorte del 50% para asegurarnos de que el último de los genes quede bien estimado. Por ello, es muy importante disponer de información biológica que nos ayude en la elección del nivel de recorte además de disponer,

asimismo, de alguna herramienta exploratoria que nos permita hacernos una idea, basada en los datos, de cuantos genes pueden quedar mal estimados con el nivel de recorte elegido.

Para aproximarnos al porcentaje de genes que han podido sufrir porcentajes de contaminación superiores al elegido, vamos a estudiar la distribución del estadístico cociente de varianzas recortadas ($\hat{\sigma}^2_{\alpha-core}$) a dos niveles distintos, para una distribución normal y diferentes hipótesis relacionadas con el porcentaje de contaminación.

La expresión del estadístico para el gen g y niveles de recorte α_1 y α_2, es,

$$V_{g,\alpha_1,\alpha_2} = \frac{\hat{\sigma}^2_{g,\alpha_1-core}}{\hat{\sigma}^2_{g,\alpha_2-core}} \tag{3.2}$$

Tabla 3.6. Distribución de $V_{g,0.5,0.3}$, $V^{\omega}_{g,0.5,0.3}$ y porcentaje de genes con nivel de recorte insuficiente en el *dataset* de Tejidos Humanos.

	Percentiles										% genes con $\omega_g > 0.3$
	0%	1%	5%	10%	25%	30%	50%	75%	90%	100%	
$V_{g,0.5,0.3}$	0.006	0.250	0.316	0.342	0.382	0.391	0.422	0.461	0.498	0.851	-
$V^{0}_{g,0.5,0.3}$	0.267	0.321	0.350	0.370	0.405	0.412	0.437	0.469	0.495	0.562	12.00
$V^{0.05}_{g,0.5,0.3}$	0.287	0.312	0.345	0.368	0.397	0.407	0.431	0.467	0.503	0.637	10.61
$V^{0.1}_{g,0.5,0.3}$	0.275	0.315	0.339	0.360	0.395	0.402	0.429	0.472	0.513	0.645	8.98
$V^{0.15}_{g,0.5,0.3}$	0.229	0.290	0.333	0.351	0.380	0.388	0.417	0.461	0.503	0.658	7.73
$V^{0.2}_{g,0.5,0.3}$	0.227	0.281	0.315	0.332	0.366	0.376	0.402	0.435	0.478	0.595	4.92
$V^{0.25}_{g,0.5,0.3}$	0.198	0.256	0.294	0.314	0.344	0.351	0.379	0.415	0.448	0.602	2.87
$V^{0.3}_{g,0.5,0.3}$	0.158	0.200	0.234	0.256	0.287	0.294	0.319	0.355	0.388	0.515	0.69

Aplicando el método de Monte Carlo hemos obtenido los percentiles 5 de las distribuciones de $V^{\omega}_{g,0.5,0.3}$ para diferentes hipótesis de contaminación variando en $\omega \in [0,0.3]$. Luego hemos comparado la distribución de $V_{g,0.5,0.3}$ en la matriz de expresión con la que estamos trabajando, con los umbrales anteriores y hemos obtenido el porcentaje de genes que tienen un valor de $V_{g,0.5,0.3}$ por debajo de estos percentiles 5. Estos valores pueden ser utilizados como aproximación del porcentaje de genes que necesitarían un nivel de recorte mayor al 30%. En la tabla 3.6, se muestran las distribuciones de $V_{g,0.5,0.3}$, $V^{\omega}_{g,0.5,0.3}$ y una estimación de este porcentaje de genes para cada hipótesis de porcentaje de contaminación, ω, en $\omega = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

A partir de la información de esta tabla, podríamos ofrecer una única estimación del porcentaje de genes con nivel de contaminación por encima del 30%, utilizando una media ponderada de los porcentajes obtenidos para cada nivel de contaminación. Para obtener los pesos correspondientes a cada uno de los porcentajes en la ponderación, utilizamos la distribución de la estimación smart de la proporción de observaciones contaminadas. La figura 3.6 muestra esta distribución.

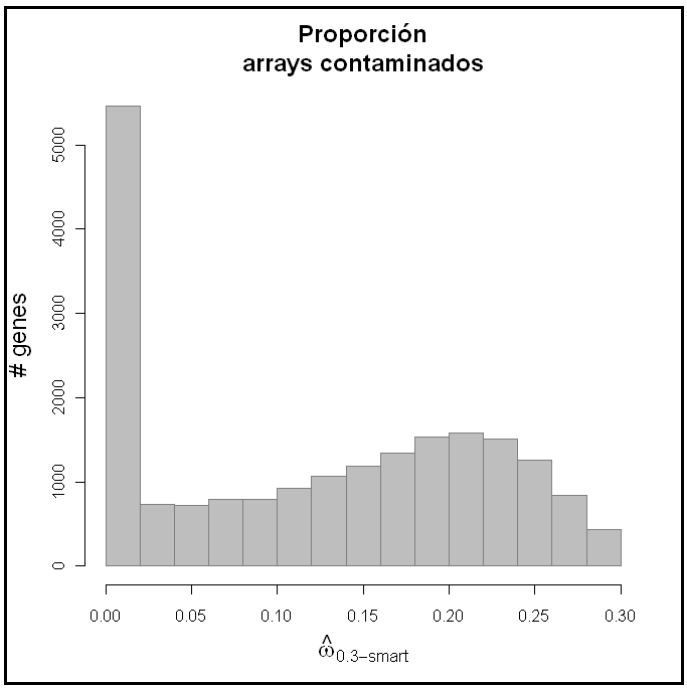


Figura 3.6. Distribución del porcentaje de arrays contaminados en el *dataset* de Tejidos Humanos

Aplicando este criterio obtenemos que el porcentaje estimado de genes que necesitan un nivel de recorte superior al 30% es el 7%, aproximadamente 1425 genes de los 20172 estudiados.

3.2.2.1. Identificación de genes *housekeeping*

Los genes *housekeeping* (HK) **[Watson et al, 1987]**, se definen como aquellos genes que se expresan a un nivel constante, necesario para mantener las funciones celulares **[Butte et al, 2001]**. Esta familia de genes es muy utilizada como conjunto de referencia para normalizar niveles de expresión entre diferentes muestras, y por su importancia, muchos trabajos han centrado su objetivo en identificar genes HK. Algunos ejemplos recientes corresponden a, **[Chang et al, 2011]** o **[Dong et al, 2011]**.

Conjuntos de datos como el *dataset* de Tejidos Humanos, en el que se pretende representar los tejidos de un individuo sano, puede ser un buen punto de partida para identificar este

tipo de genes. En este caso, esperamos que un gen HK tenga un alto nivel de expresión en todos los tejidos simultáneamente y que, además, no presente expresión diferencial frente al núcleo mayoritario en ninguna de las condiciones consideradas. El consenso actual sobre este tipo de gen, admite la posibilidad de que alguna condición pueda tener expresión diferencial frente al resto **[Greer et al, 2010]**. En la figura 3.7 se representa como ejemplo, el perfil del gen ENSG00000160710 (ADAR) que codifica a la enzima *Double-stranded RNA-specific adenosine deaminase* implicada en el proceso del *splicing*, y que en este conjunto aparece infra-expresado en músculo esquelético y en lengua.

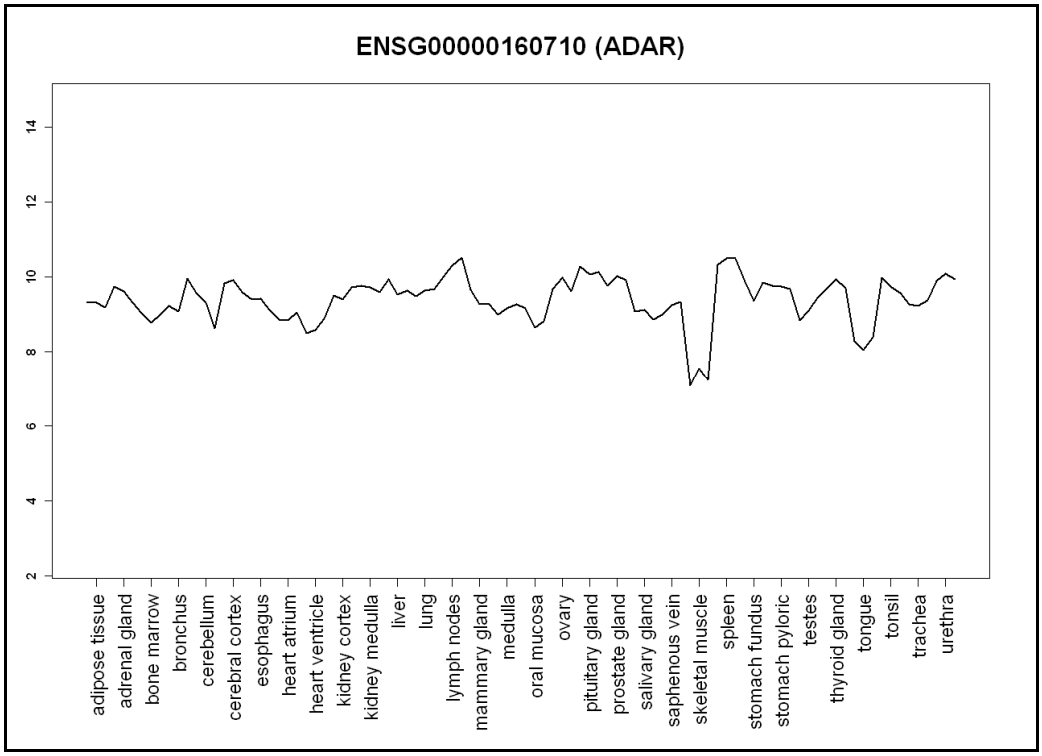


Figura 3.7. Perfil de expresión del gen HK ENSG00000160710 (ADAR).

A continuación ofrecemos unos criterios parametrizables para definir este tipo de genes a partir de la estimación del comportamiento típico de cada gen y de la respuesta relativa a este comportamiento ofrecida por cada uno de los tejidos. Para ser aplicados, un investigador debería elegir valores para los correspondientes parámetros y automáticamente obtendría la correspondiente lista de genes HK.

Los criterios que proponemos son,

(i) Nivel medio de expresión $\hat{\mu}_{g,\alpha-smart}$ alto, obtenido con un nivel de recorte no muy grande. Se espera que los genes HK estarán sobre-expresados en la mayoría de los tejidos, lo que se traduce en esperar un valor alto para $\hat{\mu}_{g,\alpha-smart}$ con α pequeño.

(ii) Valores no muy elevados del cociente de varianzas estimadas, V_{g,α_1,α_2} para dos niveles de recorte con $\alpha_1 > \alpha_2$. Se espera que la expresión de los genes HK sea muy homogénea en todos los grupos estudiados. Por tanto, no deberían observarse cambios grandes en cocientes del tipo indicado.

(iii) No muchas condiciones deberían mostrar sobre-expresión diferencial frente al núcleo del gen y la magnitud de esta expresión diferencial debería ser moderada. Para recoger esta idea utilizaremos,

$$D^{\alpha}_{g,\max} = \max_{1\le j\le K}\left(\overline{Y}_{gj\cdot} - \hat{\mu}_{g,\alpha-smart}\right) \tag{3.3}$$

para un nivel de recorte α no muy grande. Los genes HK no deberían aparecer con valores grandes en este máximo de desviaciones respecto del núcleo central.

Tabla 3.7. Número de genes que cumplen cada uno de los criterios de gen HK según la cota fijada en cada uno de ellos

$\hat{\mu}_{g,0.15-smart} \ge c$	**c**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
	# g	20172	19854	16394	12105	7673	3510	1249	428	186	87	34	8
	% g	100	98.42	81.27	60.01	38.04	17.40	6.19	2.12	0.92	0.43	0.17	0.04
$V_{g,0.5,0.15} \ge c$	**c**	**0.00**	**0.05**	**0.09**	**0.14**	**0.18**	**0.23**	**0.27**	**0.32**	**0.36**	**0.41**	**0.45**	**0.50**
	# g	20172	20060	19789	19099	16795	10937	4409	981	148	25	2	0
	% g	100	99.44	98.10	94.68	83.26	54.22	21.86	4.86	0.73	0.12	0.01	0
$D^{0.15}_{g,\max} \le c$	**c**	**5**	**3**	**2**	**1.78**	**1.56**	**1.33**	**1.11**	**0.89**	**0.67**	**0.44**	**0.22**	**0**
	# g	19419	17500	15068	14229	13238	11998	10353	8214	5285	1976	168	0
	% g	96.27	86.75	74.70	70.54	65.63	59.48	51.32	40.72	26.20	9.80	0.83	0.00

Para concretar con parámetros los criterios ofrecidos empezaremos eligiendo los niveles de recorte. Proponemos niveles de recorte pequeños en los criterios (i) y (iii), $\alpha = 0.15$, y en (ii) $\alpha_1 = 0.5$ y $\alpha_2 = 0.15$.

La tabla 3.7 ofrece información relativa al número de genes que cumplen cada uno de estos criterios por separado, según diferentes elecciones para los parámetros correspondientes.

La elección de parámetros que proponemos en esta memoria para definir genes HK corresponde a escoger como umbral 6 para el criterio (i), lo que equivale a $\hat{\mu}_{g,0.15-smart} \ge 6$; 0.3 para el criterio (ii), equivalente a $V_{g,0.5,0.15} \ge 0.3$; y 2 para el criterio (iii) correspondiente a $D^{0.15}_{g,\max} \le 2$. La tabla 3.8 resume esta parametrización. Con ella, aparecen como genes HK 662 genes, que representan el 3.28% del total de genes analizados.

Tabla 3.8. Resumen de los criterios utilizados para determinar la lista HK6 de genes HK

Característica	Score	Parámetros	Criterio
Alto nivel de expresión	$\hat{\mu}_{g,\alpha-smart}$	$\alpha = 0.15$	$\hat{\mu}_{g,0.15-smart} \geq 6$
Expresión diferencial mayoritaria homogénea	V_{g,α_1,α_2} (3.2)	$\alpha_1 = 0.5$ y $\alpha_2 = 0.15$	$V_{g,0.5,0.15} \geq 0.3$
Bajo nivel de sobre-expresión minoritaria	$D^{\alpha}_{g,\max}$ (3.3)	$\alpha = 0.15$	$D^{0.15}_{g,\max} \leq 2$

Existen en la literatura distintas listas de genes HK obtenidas aplicando diferente métodología a distintos conjuntos de datos. El consenso entre ellas tiende a ser bajo. En la tabla 3.9 se comparan cinco listas de genes HK obtenidas a partir de muestras de tejidos procesadas en microarrays de *Affymetrix*. Los porcentajes de concordancia entre ellas van desde el 12% al 78%. En esta tabla aparece también, la concordancia entre nuestra lista y las cinco restantes, ésta se sitúa en valores inferiores al 23%.

Tabla 3.9. Comparación de listas de genes HK obtenidas a partir de análisis de datos de microarrays de Affymetrix en los que se analizan distintos tipos de tejidos

Estudio	# array	# tejido	# gen HK	% genes únicos	% genes en otras listas					
					HK1	HK2	HK3	HK4	HK5	HK6
HK1 **[Hsiao et al, 2001]**	59	19	350	10.29	100	43.71	48.29	78.00	73.14	8.00
HK2 **[Eisenberg y Levanon, 2003]**	46	32	542	25.28	28.23	100	35.61	61.99	57.56	6.09
HK3 **[Tu et al, 2006]**	142	79	1334	28.34	12.67	14.47	100	61.09	47.90	6.22
HK4 **[Zhu et al, 2008]**	18	18	2170	27.37	12.58	15.48	37.56	100	55.12	6.13
HK5 **[Chang et al, 2011]**	1431	43	1990	29.95	12.86	15.68	32.11	55.12	100	7.64
HK6 (este trabajo)	96	32	662	63.75	4.23	4.98	12.54	20.09	22.96	100

En la figura 3.8 representamos la distribución de los índices en los que se basan nuestros criterios en grupos creados a partir de las listas HK disponibles en la literatura. El primer criterio (figura 3.8 (a)) es prácticamente común a todas las listas. La distribución de $V_{g,0.5,0.15}$ (figura 3.8 (b)) de los genes clasificados como HK por los trabajos HK1-HK5 es muy similar a la de los genes no clasificados como HK por ninguna de las listas, y significativamente menor que 0.3, la cota fijada en este trabajo. En cuanto al criterio $D^{0.15}_{g,\max}$ (figura 3.8 (c)), todos los genes clasificados como HK por alguna de las listas, muestran valores ligeramente más pequeños. La diferencia de la clasificación HK6 respecto al resto, está en los valores máximos, la cota propuesta en este trabajo, 2, está próxima al percentil 75 de la distribución de $D^{0.15}_{g,\max}$ en los genes clasificados como HK por alguna de las otras listas evaluadas.

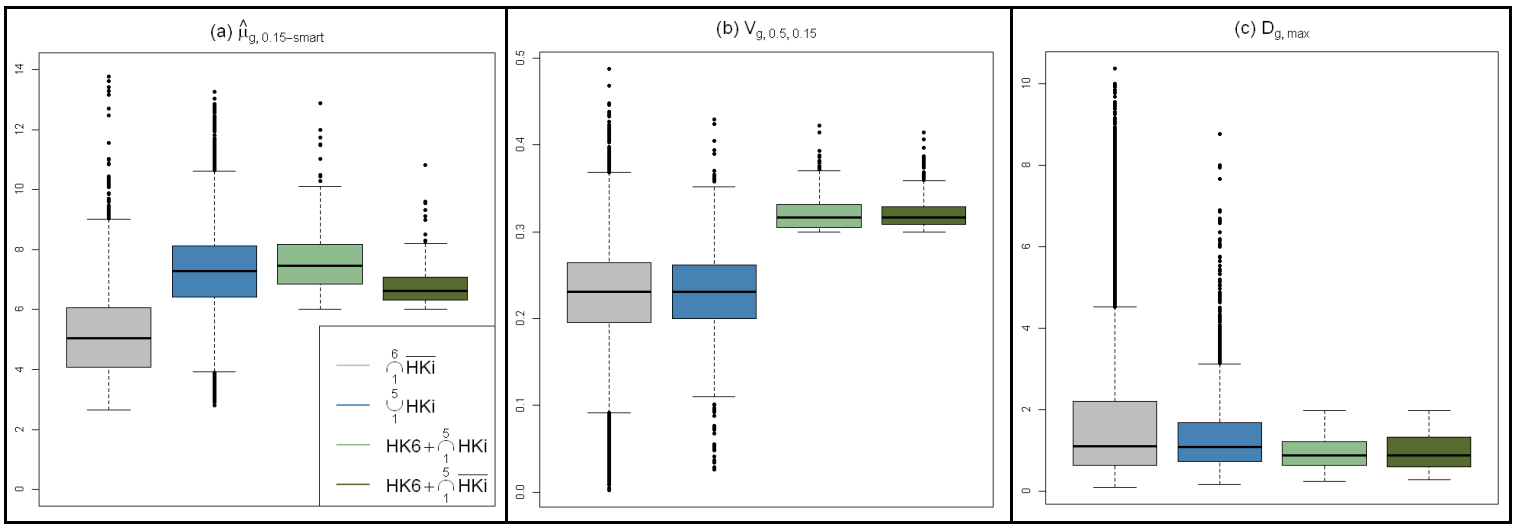


Figura 3.8. Distribución de (a) $\hat{\mu}_{g,0.15-smart}$, (b) $V_{g,0.5,0.15}$ y (c) $D_{g,\max}$, según la clasificación de gen HK en cada uno de los listados analizados.

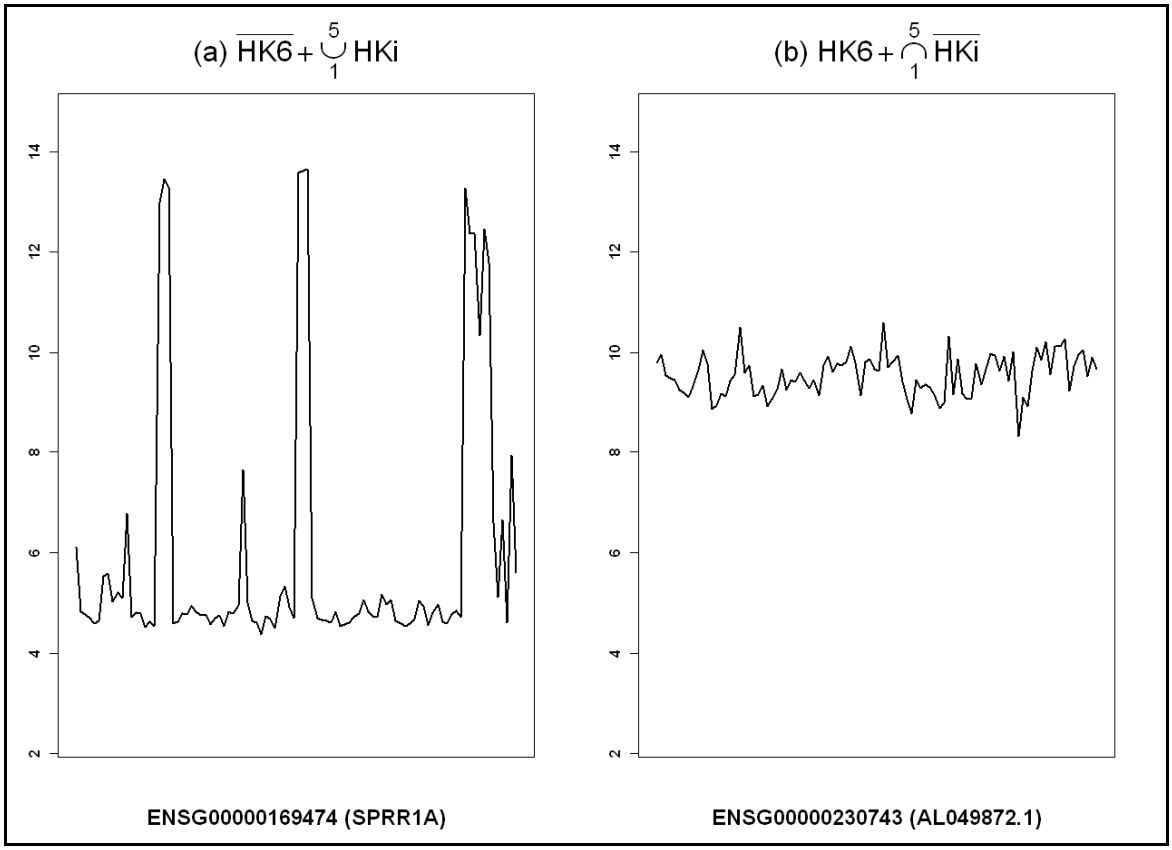


Figura 3.9. (a) Gen ENSG00000169474 (SPRR1A) no clasificado como HK por HK6 e incluido en alguna de las listas HK1-HK5. (b) Gen ENSG00000230743 (AL049872.1) clasificado como HK en HK6 y no incluido en ninguna de las listas HK1-HK5.

En la figura 3.9 se muestran algunos ejemplos extremos de genes clasificados como HK por HK1-HK5 y no por HK6 (figura 3.9 (a)), y genes HK según HK6 y no según el resto (figura 3.9 (b)). El gen ENSG00000169474 (SPRR1A), claramente no tiene un comportamiento HK. De hecho, aparece sobre-expresado en esófago, mucosa oral, lengua y amígdala, lo que es coherente con el trabajo de **[Robinson et al, 1994]**, en el que se demostró expresión diferencial en tejido oral humano. Por otra parte, el gen ENSG99999230743 (AL049872.1) tiene un comportamiento característico de HK: nivel de expresión alto para todos los tipos de tejidos y en ninguno de ellos aparece diferencialmente expresado.

Para comprobar la coherencia biológica del listado obtenido, en la tabla 3.10 se muestran las categorías más relevantes tras un análisis de enriquecimiento funcional en términos de GO-BP realizado con la herramienta DAVID. Todas ellas tienen que ver con funciones propias de HK.

Tabla 3.10 Categorías más significativas en el análisis de enriquecimiento funcional en la lista HK6. En estas categorías están anotados 625 genes, aproximadamente el 95% de los genes evaluados.

		# genes	%	p-valor FDR
GO:0008152~ metabolic process		**315**	**50.4**	
	GO:0016071~mRNA metabolic process	30	9.52	8.86E-06
	GO:0044265~cellular macromolecule catabolic process	44	13.97	9.08E-06
	GO:0009057~macromolecule catabolic process	45	14.29	2.73E-05
	GO:0006396~RNA processing	36	11.43	4.22E-05
	GO:0006793~phosphorus metabolic process	51	16.19	5.10E-05
	GO:0006796~phosphate metabolic process	51	16.19	5.10E-05
	GO:0016310~phosphorylation	43	13.65	4.23E-04
	GO:0006350~transcription	82	26.03	6.77E-04
	GO:0006412~translation	25	7.94	8.26E-04
	GO:0006397~mRNA processing	24	7.62	0.0017521
GO:0009987 ~cellular process		**398**	**63.7**	
	GO:0016071~mRNA metabolic process	30	7.54	0.0023477
	GO:0046907~intracellular transport	43	10.80	0.0025326
	GO:0006886~intracellular protein transport	30	7.54	0.0028881
	GO:0034613~cellular protein localization	31	7.79	0.006656
	GO:0070727~cellular macromolecule localization	31	7.79	0.0076178

Tabla 3.10 (Continuación)

		# genes	%	p-valor FDR
GO:0016043~ cellular component organization		**102**	**16.3**	
	GO:0043933~macromolecular complex subunit organization	32	31.37	5.33E-13
	GO:0065003~macromolecular complex assembly	31	30.39	7.22E-13
	GO:0051276~chromosome organization	25	24.51	1.51E-10
	GO:0070271~protein complex biogenesis	24	23.53	3.03E-09
	GO:0006461~protein complex assembly	24	23.53	3.03E-09
	GO:0006325~chromatin organization	20	19.61	6.22E-08
	GO:0016044~membrane organization	19	18.63	5.93E-07
	GO:0016568~chromatin modification	16	15.69	2.73E-06
	GO:0007005~mitochondrion organization	12	11.76	9.77E-06
	GO:0046907~intracellular transport	20	19.61	5.26E-04
	GO:0034621~cellular macromolecular complex subunit organization	15	14.71	6.38E-04
	GO:0034622~cellular macromolecular complex assembly	14	13.73	0.0010764
	GO:0016570~histone modification	9	8.82	0.0056237
	GO:0016569~covalent chromatin modification	9	8.82	0.0071576
	GO:0006333~chromatin assembly or disassembly	9	8.82	0.0075921
	GO:0016192~vesicle-mediated transport	17	16.67	0.0084748
GO:0044085~ cellular component biogenesis		**47**	**7.5**	
	GO:0065003~macromolecular complex assembly	31	65.96	2.06E-25
	GO:0043933~macromolecular complex subunit organization	31	65.96	1.45E-24
	GO:0070271~protein complex biogenesis	24	51.06	4.73E-18
	GO:0006461~protein complex assembly	24	51.06	4.73E-18
	GO:0034622~cellular macromolecular complex assembly	14	29.79	4.01E-08
	GO:0034621~cellular macromolecular complex subunit organization	14	29.79	1.69E-07
	GO:0042254~ribosome biogenesis	9	19.15	1.03E-05
	GO:0022613~ribonucleoprotein complex biogenesis	10	21.28	1.16E-05
	GO:0016072~rRNA metabolic process	8	17.02	4.68E-05
	GO:0006364~rRNA processing	7	14.89	9.52E-04
	GO:0034660~ncRNA metabolic process	9	19.15	0.0013937
	GO:0051259~protein oligomerization	8	17.02	0.0027292
GO:0016265 ~death		**36**	**5.8**	
	GO:0008219~cell death	36	100.00	1.67E-42
	GO:0016265~death	36	100.00	2.15E-42
	GO:0012501~programmed cell death	31	86.11	8.65E-33
	GO:0006915~apoptosis	30	83.33	6.20E-31
	GO:0042981~regulation of apoptosis	20	55.56	1.02E-11
	GO:0043067~regulation of programmed cell death	20	55.56	1.24E-11
	GO:0010941~regulation of cell death	20	55.56	1.32E-11
	GO:0043065~positive regulation of apoptosis	14	38.89	3.32E-08
	GO:0043068~positive regulation of programmed cell death	14	38.89	3.62E-08
	GO:0010942~positive regulation of cell death	14	38.89	3.84E-08
	GO:0006917~induction of apoptosis	12	33.33	4.14E-07
	GO:0012502~induction of programmed cell death	12	33.33	4.28E-07

Capítulo 4
Expresión diferencial en múltiples clases independientes

Uno de los objetivos del análisis de datos de expresión génica es identificar genes diferencialmente expresados en distintas condiciones experimentales o biológicas. La aproximación más sencilla para la obtención de una lista de genes diferencialmente expresados (genes DE) es considerar como significativos todos los genes con valor de FC superior a un nivel dado. Muchos autores consideran una diferencia significativa si es de al menos dos o tres-*fold* **[Draghici 2002]**; **[Jiang et al, 2001]**; **[DeRisi et al, 1997]**; **[Schena et al, 1996]**. Este criterio ligado a la significación clínica, es muy simple e intuitivo, e incluso es aplicable a conjuntos de datos en los que no hay réplicas biológicas, muy comunes en los primeros experimentos con microarrays. Sin embargo, tiene importantes desventajas como la arbitrariedad en la elección del umbral, o que no tiene en cuenta la posiblemente diferente variabilidad en los genes ni la variabilidad debida al muestreo.

En aquellos experimentos en los que hay réplicas biológicas, es posible estimar la variación asociada a la estimación del FC en cada gen, y aplicar, en el caso de dos poblaciones, el estadístico t-Student en cada uno de ellos. El problema que se plantea, es que el número de réplicas en cada clase suele ser bajo, y la falta de eficiencia en la estimación del error estándar puede producir inestabilidad en el estadístico t **[Cui y Churchill, 2003]**. Para solucionar este problema **[LaPointe et al, 2012]**, **[Zhu et al, 2010]**, **[Tanaka et al, 2000]** y **[Arfin et al, 2000]** proponen utilizar una estimación de la varianza común a todos los genes o modificar el estadístico t-Student añadiendo información de la variabilidad de todos los genes simultáneamente, como en el *Empirical Bayes Moderated t-test* **[Smyth, 2004]**, y el *regularized t-test* **[Long et al, 2001]**; **[Baldi y Long, 2001]**, también conocido como CyberT. Esta última opción es la empleada en el, muy utilizado, estadístico SAM **[Tusher et al, 2001]**, que añade una constante positiva al denominador del estadístico t, estimada a partir de todos los genes. Todas estas propuestas se pueden generalizar para ser aplicadas en el análisis de la varianza en situaciones con más de dos poblaciones.

Otro problema que aparece en las matrices de expresión génica ligado a la aplicación de este tipo de contrastes, está relacionado con la multiplicidad de los mismos, su repetida aplicación produce que muchos de ellos den resultados positivos por azar. En este caso, controlar la tasa de error global es demasiado restrictivo. En la mayoría de los casos, el interés de estos

análisis es exploratorio, interesa identificar muchos genes diferencialmente expresados, incluso a costa de que incluyamos en la lista algún falso positivo. Ello lleva a la utilización de la *False Discovery Rate* (FDR) **[Benjamini y Hochberg, 1995]** definida como la proporción esperada de falsos positivos en todos los contrastes realizados. Para estimar la FDR se utilizan métodos basados en permutaciones, como por ejemplo, el *empirical Bayes method* **[Efron et al, 2001]**, MMM (*Mixture Model Method*) **[Pan, 2003]**, y, sobre todo, el SAM **[Tusher et al, 2001]**. Estos procedimientos consideran todos los genes simultáneamente, estimando la distribución utilizando la distribución empírica de todos los valores obtenidos de las permutaciones. La utilización de todos los genes para construir la distribución nula se apoya en la suposición de que los estadísticos, bajo la hipótesis nula, están idénticamente distribuidos. Sin embargo, los estadísticos de los genes DE no tienen la misma distribución que los de los genes EE, y permutar los niveles de expresión de los genes DE incrementa la varianza de la distribución nula **[Xie et al, 2005]**. Algunos autores tratan de eliminar este sesgo modificando el procedimiento de estimación de esta distribución utilizando pesos para intentar que los genes DE tengan menos influencia **[Guo y Pan, 2004]**. En el extremo, autores como **[Xie et al, 2005]**; **[Zhang, 2007]** o **[Jiao y Zhang, 2008]** proponen excluir completamente de la estimación de la FDR los genes DE. En este sentido, **[Jiao y Zhang, 2008]** demuestran por simulación que en el caso de eliminar todos aquellos genes declarados como DE, la FDR pasará a infra-estimarse.

En el análisis de expresión diferencial, es muy común la estrategia de trabajar en dos fases. En una primera etapa, los datos se utilizan para identificar y eliminar aquellos genes no informativos, y en una segunda, se contrasta la hipótesis de interés en aquellos genes que pasaron el filtro inicial. La razón es que, en estos experimentos, por lo general, existe un porcentaje importante de genes que, o bien no están expresados, o están por debajo del límite detectable por la técnica. Además hay otro conjunto importante de genes, que aunque están expresados y es posible detectarlos, no están DE. Eliminar del análisis los genes no informativos va a reducir la penalización por las comparaciones múltiples y así, aumentar la potencia **[van Iterson et al, 2010]**.

Existen distintos métodos de filtrado en la literatura. Los más utilizados son métodos descriptivos, basados en el nivel medio de expresión y dispersión entre arrays, que no van a tener en cuenta la información relativa a la clase de pertenencia. Son los llamados filtros no específicos y algunos autores apuntan que son preferibles porque no interfieren en análisis posteriores **[Bourgon et al, 2010]**; **[Hackstadt y Hess, 2009]**. Sin embargo, hay bastante controversia en este tema y otros autores prefieren hacer un filtrado que permita seleccionar los genes que se espera tengan más probabilidades de tener algún tipo de expresión diferencial **[McCarthy y Smith, 2009]**; **[Zhang y Cao, 2009]**. Una forma natural de hacer esta selección inicial de genes es utilizar contrastes del tipo ANOVA, en los que se comprueba una hipótesis global sobre la existencia de expresión diferencial en alguna de las condiciones estudiadas **[Chen et al, 2005]**; **[Pavlidis, 2003]**. Cuando se trabaja con un conjunto de genes pre-filtrados se va a potenciar el sesgo que se produce en la estimación de la FDR debido a

la mezcla de genes DE y genes EE, produciéndose un exceso de falsos positivos. En cualquier caso, y a pesar de este problema, se asume que eliminar de los análisis genes no informativos es una buena estrategia.

En este capítulo abordamos el problema de identificar expresión diferencial entre múltiples clases independientes. Proponemos la realización de contrastes, para cada gen, basados en las diferencias entre cada clase y el núcleo de expresión obtenido aplicando recortes imparciales. Asumimos por tanto, que en cada gen existe un núcleo de expresión del que, posiblemente se diferencien algunas clases. Nuestra intención es identificar los genes que contienen clases *DE* respecto de ese núcleo y, para cada uno de esos genes, las clases que muestran ese comportamiento diferencial.

El método trabaja en dos etapas. En la primera etapa, tratamos de eliminar el conjunto de genes no informativos, a partir de un contraste tipo ANOVA. En el estadístico interviene el porcentaje de clases que más diferencia ha mostrado respecto del núcleo, siendo este porcentaje un parámetro del método que hay que pre-fijar. De esta forma, intentamos mejorar la potencia frente a alternativas relacionadas con la expresión diferencial de un porcentaje de clases inferior al pre-fijado. En la segunda etapa, utilizado únicamente los genes que rechazaron el primer contraste, realizamos contrastes múltiples de cada clase frente al núcleo de expresión del gen. **[Alvarez-Esteban et al, 2012]** aplicaron con éxito esta idea, en una aproximación no paramétrica, para detectar diferencias debidas al origen de los estudiantes en un examen nacional. Con este tipo de aproximación reducimos el número de contrastes (a tantos como el número de clases) frente a realizar contrastes entre todos los pares de clases. Adicionalmente, esta aproximación, facilita la interpretación ya que nos proporciona una afirmación por gen, relacionada con si éste se diferencia o no del núcleo, no de cada uno de los demás.

Aplicamos la metodología SAM **[Tusher et al, 2001]** para la obtención de la listas de genes y clases *DE* y la estimación de la correspondiente FDR. Los estadísticos que proponemos para estos contrastes incorporan la constante típica de SAM, ya mencionada, en el denominador. En nuestro caso recomendamos la elección de dicha constante en relación con la detección de diferencias biológicamente relevantes. En la estimación de la FDR se va a aplicar un p-valor ajustado propuesto en **[Dudoit et al, 2002b]**. Además, y con el objetivo de corregir el sesgo en la estimación de esta tasa, se incluye una adaptación del método propuesto en **[Guo y Pan, 2004]**. Para conseguir que los genes DE tengan menos influencia en la estimación de la distribución nula, se propone utilizar la probabilidad de que un gen sea EE como probabilidad de inclusión en la construcción de una muestra de genes con reemplazamiento, cuyas permutaciones se utilizarán para estimar la FDR. Estas probabilidades se calculan a partir de la distribución normal con parámetros los que se obtienen de un recorte imparcial de los estadísticos.

La motivación biológica que hay detrás de esta propuesta, es la identificación de genes con patrones de expresión característicos de determinados tipos de tejidos, genes

predominantemente expresados en unos pocos tipos de tejidos u órganos a la vez, y como caso particular, los genes específicos de un tejido. Las matrices de datos que se manejan en estos casos recogen información de múltiples clases independientes, cada una de ellas ligada a un tipo de tejido. Así, se espera que los resultados sean biológicamente más completos cuantas más clases se analicen simultáneamente.

Dada la importancia de la metodología SAM en las herramientas que proponemos en este capítulo, en la siguiente sección hacemos una breve descripción de esta metodología. El método está implementado en el *package* de R samr **[Tibshirani et al, 2010]**, y todos los detalles técnicos pueden encontrarse en **[Chu et al, 2011]**.

4.1. La metodología SAM

La metodología SAM **[Tusher et al, 2001]** trata de resolver los problemas que se plantean en el análisis de expresión diferencial ofreciendo un estadístico, modificación del estadístico t-Student, junto con un procedimiento, basado en permutaciones, para estimar la FDR y a partir de ella, obtener un listado de genes DE.

El estadístico SAM trata de eliminar el problema de la inestabilidad del estadístico t, añadiendo una constante positiva, s_0, denominada *fudge factor*, común a todos los genes, en el denominador del estadístico. Esta constante, tiene como objetivo que la variación en el valor del estadístico, denotado por d_g^{SAM} con $1 \leq g \leq n_G$, sea similar para todos los niveles de intensidad. Así, a medida que se va aumentando el valor de s_0, va ir disminuyendo la importancia que el estadístico asigna a algunos genes por tener un error estándar bajo, consiguiendo que genes con efectos pequeños dejen de ser significativos.

Existen varias propuestas para calcular el valor de esta constante. Los propios autores proponen un procedimiento automático que asigna a s_0 el valor de la estimación de los errores estándar por gen, que minimiza el coeficiente de variación del estadístico d_g^{SAM}. Otros autores, como **[Choe et al, 2005]**, tratan de simplificar y, directamente recomiendan usar percentiles altos de la distribución de estas estimaciones, habitualmente el percentil 90. En otros trabajos, como en **[Broberg, 2003]**, se plantea estudiar la tasa de falsos positivos y la de falsos negativos para distintos valores de s_0, eligiendo el valor que minimice ambas tasas a la vez.

Uno de los aspectos más valorados de esta metodología, es su estrategia para decidir que genes son estadísticamente significativos. La idea general es considerar significativos los primeros genes, ordenados según el valor del estadístico, de forma que, entre ellos, el porcentaje de genes EE no superen un valor máximo fijado para la FDR. Esta tasa se estima

utilizando las permutaciones de todos los genes simultáneamente. A partir de B permutaciones, se obtiene el estadístico esperado para el gen que ocupa la posición g, $\overline{d}_{(g)}^{SAM}$, como el promedio de los B valores del estadístico que ocupan esa misma posición en cada una de las permutaciones. Utilizando un QQ-*plot*, en el que se representa el estadístico esperado frente al observado, y fijando un umbral Δ, para la diferencia entre estos dos estadísticos, se obtienen dos puntos de corte globales, κ_1 y κ_2, que dan lugar a una región de rechazo del tipo $d_g^{SAM} < \kappa_1$ y $d_g^{SAM} > \kappa_2$. Una característica interesante, es que estas regiones pueden ser asimétricas, lo que va a permitir que los contrastes sean más potentes en situaciones donde hay más genes sobre-expresados que infra-expresados o viceversa.

Dado un umbral Δ, la estimación de la FDR se basa en la propuesta de **[Storey, 2002]**,

$$\widehat{FDR} = \hat{\pi}_0 \frac{\widehat{FP}}{R} \tag{4.1}$$

donde $\widehat{FP}$ es una estimación de los falsos positivos, calculada como el promedio del número de estadísticos declarados significativos en cada una de las B permutaciones, R es el número de contrastes declarados como significativos y $\hat{\pi}_0$ es una estimación de la proporción de genes EE que se define como,

$$\hat{\pi}_0 = \min\left(\frac{\#\left\{ d_g^{SAM} \in (q_1, q_2) \right\}}{0.5 \cdot n_G}, 1 \right) \tag{4.2}$$

con q_1 y q_2 los percentiles 25 y 75 de los estadísticos obtenidos de las permutaciones. Para estimar π_0, se está asumiendo que la mitad de estos valores están en la región de rechazo.

El proceso de estimación de la FDR se repite para un conjunto de umbrales Δ, y se elige la solución que proporciona una estimación de FDR más próxima, por debajo, a un valor fijado. El listado de genes diferencialmente expresados, serán aquellos cuyo valor del estadístico esté contenido en la región de rechazo correspondiente al umbral elegido. Para permitir tomar decisiones simplemente fijando un nivel de significación, en **[Dudoit et al, 2002b]** se propone una forma de calcular los p-valores ajustados, adaptada a la estimación de la FDR que proporciona esta metodología.

4.2. Método propuesto

Los datos de partida serán una muestra $\{Y_{gks}\}$ de niveles de expresión donde g representa al gen, con $g = 1, \ldots, n_G$, k la condición o clase, con $k = 1, \ldots, K$ y s la réplica biológica, con

$s = 1, \ldots, n_k$. Se trata de evaluar la expresión diferencial en un determinado gen comparando las K clases, con K grande.

4.2.1. Primer paso: contraste global

4.2.1.1. Definición del estadístico

Sean $\mu_{g,core}$ y $\sigma_{g,core}$, los parámetros de localización y escala que determinan el núcleo de expresión del gen g, $\hat{\mu}_{g,\alpha-smart}$ y $\hat{\sigma}_{g,\alpha-smart}$ sus estimaciones, y μ_{gk} la media de la clase k en el gen g, estimada a partir de la media de las n_k réplicas biológicas del gen g y la clase k ($\overline{Y}_{gk\cdot}$). Para cada gen g y clase k, la diferencia entre la expresión media de la clase y la expresión típica del gen, en valor absoluto, será una primera aproximación de la expresión diferencial de esa clase en ese gen. Se denota a esta medida como $\left|FC_{gk,\alpha-core}\right|$.

Sea $\overline{Y}_{g(k)}$ la estimación del nivel de expresión en el gen g y en la clase que ocupa la posición k-ésima de la muestra de $\left|FC_{gk,\alpha-core}\right|$ ordenada de mayor a menor. Considerando la proporción pk de clases con mayor expresión diferencial, se contrasta la hipótesis,

$$H_{g0} : \mu_{g1} = \ldots = \mu_{gK} = \mu_{g,core} \tag{4.3}$$

utilizando el estadístico,

$$F_{g,core} = \frac{\lceil pk \cdot K \rceil \cdot \sum_{k=1}^{\lceil pk \cdot K \rceil} n_{(k)} \left(\overline{Y}_{g(k)\cdot} - \hat{\mu}_{g,\alpha-smart} \right)^2}{\sum_{k=1}^{\lceil pk \cdot K \rceil} \left(1 - \frac{n_{(k)}}{n_{core} + \sum_{j=1}^{\lceil pk \cdot K \rceil} n_{(j)}} \right) S^2_{g(k)} + \left(1 - \frac{n_{core}}{n_{core} + \sum_{j=1}^{\lceil pk \cdot K \rceil} n_{(j)}} \right) \hat{\sigma}^2_{g,\alpha-smart} + s_0^{(global)}} \tag{4.4}$$

donde n_k representa el número de réplicas en la clase k, n_{core} el número de observaciones del núcleo de expresión, común a todos los genes, S^2_{gk} la cuasivarianza muestral del gen g y la clase k, y $s_0^{(global)}$ es una constante, común a todos los genes, cuyo objetivo es minimizar el problema de la inestabilidad en la estimación de la varianza. $\lceil a \rceil$ denota el entero más pequeño mayor que a.

Se trata de un estadístico del mismo tipo que el utilizado en los modelos ANOVA, calculado como la tasa entre la variación de las $pk \cdot K$ clases más expresadas respecto del *core* y la variación dentro de cada una de las clases consideradas incluyendo el *core*. Para evitar

declarar como significativos genes con efectos pequeños, en términos de $FC_{gk,core}$, se añade una constante positiva en el denominador, $s_0^{(global)}$.

4.2.1.2. Parámetro $s_0^{(global)}$

Los mismos autores que diseñaron el estadístico SAM proponen un método automático para la estimación de la constate s_0 cuyos detalles técnicos pueden consultarse en **[Chu et al, 2011]**. El principal problema que tienen los métodos automáticos para establecer un parámetro, es que en la mayoría de los casos, el usuario, no es capaz de evaluar el efecto real que puede tener sobre sus datos. A esto se añade que existe una cierta desconexión, en cuanto a interpretación, entre el valor de s_0 y su verdadero objetivo: impedir que genes con diferencia de medias pequeña sean considerados significativos. Desde el punto de vista biológico, es mucho más interpretable y fácilmente elegible un umbral para la diferencia de medias, por debajo del cual, encontrar expresión diferencial carece de interés. Sea δ_0 este umbral, en términos del error estándar, uno podría traducir δ_0 en s_0 a partir de la relación,

$$\delta_0 = 2\frac{s_0}{\sqrt{n}} \tag{4.5}$$

donde n representa el tamaño muestral.

Fijado δ_0, $s_0^{(global)}$ se obtendrá añadiendo una constante positiva, generalmente pequeña, a la estimación de la varianza del *core* que impida la existencia de genes significativos por varianzas pequeñas,

$$F_{g,core} = \frac{\lceil pk\cdot K\rceil \cdot \sum_{k=1}^{\lceil pk\cdot K\rceil} n_{(k)}\left(\overline{Y}_{g(k).} - \hat{\mu}_{g,\alpha-smart}\right)^2}{\sum_{k=1}^{\lceil pk\cdot K\rceil}\left(1-\frac{n_{(k}}{n_{core}+\sum_{j=1}^{\lceil pk\cdot K\rceil} n_{(j)}}\right)S^2_{g(k)} + \left(1-\frac{n_{core}}{n_{core}+\sum_{j=1}^{\lceil pk\cdot K\rceil} n_{(j)}}\right)\left(\hat{\sigma}^2_{g,\alpha-smart}+s_0^2\right)} \tag{4.6}$$

con

$$s_0 = \frac{\delta_0}{2}\sqrt{\min_{1\le k\le\lceil pk\cdot K\rceil}\left(n_{(k)}\right)} \tag{4.7}$$

Sustituyendo s_0 por su expresión y, sin más que despejar en el segundo término del denominador de la expresión (4.6), se obtiene que,

$$s_0^{(global)} = \left(\frac{\delta_0}{2}\right)^2 \cdot \min_{1 \le k \le \lceil pk \cdot K \rceil} \left(n_{(k)}\right) \cdot \left(1 - \frac{n_{core}}{n_{core} + \sum_{j=1}^{\lceil pk \cdot K \rceil} n_{(j)}}\right) \tag{4.8}$$

4.2.2. Segundo paso: contraste por pares

El contraste global identifica genes que muestran diferencias significativas en los niveles de expresión de las clases más expresadas. Sin embargo, no identifica entre qué condiciones se encuentran estas diferencias. En principio, para solucionar este problema, uno debería plantearse un procedimiento en el que se consideraran todas las $\frac{K \cdot (K-1)}{2}$ posibles combinaciones de dos clases, lo cual puede ser muy costoso, especialmente cuando el número de clases es muy grande. En este caso, y puesto que se tiene una buena estimación de la expresión típica de cada gen, es posible reducir el problema a una única decisión por cada clase en cada gen, sin más que evaluar las diferencias de expresión entre cada clase y el *core*.

4.2.2.1. Definición del estadístico

Sean $\mu_{g,core}$ y $\sigma_{g,core}$, los parámetros de localización y escala que determinan el núcleo de expresión del gen g, $\hat{\mu}_{g,\alpha-smart}$ y $\hat{\sigma}_{g,\alpha-smart}$ sus estimaciones, y μ_{gk} la media de la clase k en el gen g, estimada utilizando $\overline{Y}_{gk\cdot}$. Para cada gen y clase, se contrasta la hipótesis,

$$H_{gk0} : \mu_{gk} = \mu_{g,core} \tag{4.9}$$

utilizando el estadístico,

$$t_{gk,core} = \frac{\overline{Y}_{gk\cdot} - \hat{\mu}_{g,\alpha-smart}}{S_{gk}^{(par)} + s_0^{(par)}} \tag{4.10}$$

con

$$S_{gk}^{(par)} = \sqrt{\frac{S_{gk}^2}{n_k} + \frac{\hat{\sigma}_{g,\alpha-smart}^2}{n_{core}}} \tag{4.11}$$

donde n_k representa el número de réplicas en la clase k, n_{core} el número de observaciones del core, S_{gk}^2 la cuasivarianza muestral del gen g y la clase k, y $s_0^{(par)}$ es una constante, común a todos los genes, cuyo objetivo es minimizar el problema de la inestabilidad en la estimación de la varianza. Una vez fijada por el usuario la diferencia de expresión biológicamente relevante, δ_0, y de forma totalmente análoga al caso del contraste global detallado en el apartado anterior,

$$s_0^{(par)} = \frac{\delta_0}{2}\sqrt{\frac{\min\left(n_k\right)}{n_{core}}} \tag{4.12}$$

4.2.3. Estimación de la FDR

Para estimar y controlar la FDR en el problema de las comparaciones múltiples, se utiliza el procedimiento propuesto en el método SAM **[Tusher et al, 2001]**, junto con el p-valor ajustado propuesto en **[Dudoit et al, 2002b]**. El procedimiento completo se resume en la figura 4.1, incluyendo una adaptación del método propuesto en **[Guo y Pan, 2004]**, basado en el recorte imparcial de los estadísticos, para conseguir que los genes DE tengan menos influencia en la estimación de la distribución nula. Este ajuste se detalla en la sección siguiente. El uso de esta rectificación, es especialmente recomendable en el caso del segundo contraste, puesto que se parte de un conjunto de genes pre-filtrados, lo que puede originar un sesgo importante en la estimación de la FDR.

4.2.3.1. Corrección de la estimación de la distribución nula del estadístico

Una idea simple para corregir el sesgo en la estimación de la FDR producido por la mezcla de genes EE y genes DE, es asignar pesos diferentes a los genes según su probabilidad de tener expresión diferencial y utilizar estos pesos en el procedimiento de estimación de la FDR **[Guo y Pan, 2004]**. Una manera de estimar esta probabilidad es utilizar un *core* de estadísticos que permitan determinar cuanto se aleja el valor del estadístico asociado a un determinado gen, respecto de ese comportamiento mayoritario. Para ello, es necesario asumir que ese comportamiento mayoritario existe. En el caso del primer contraste, es ampliamente aceptado que el número de genes con expresión diferencial, entre todos los genes que se estudian en un microarray, es pequeño **[Su et al, 2002]**. En el segundo contraste, se tiene un estadístico por par gen - clase. Si el número de clases es grande, es muy posible que la expresión diferencial, si existe, se dé predominantemente, en pocas clases simultáneamente **[Liang et al, 2006]**.

Para estimar los pesos el procedimiento es el siguiente,

1) Determinar el *core* del conjunto de estadísticos, $\left\{Z_g,\ 1 \le g \le n_G\right\}$ donde n_G es el número de genes en el caso del primer contraste, o el número de pares gen-clase en el segundo. Para ello y, dado un nivel de recorte, $\alpha_q \in \left[0,\ 1\right]$ se determina el conjunto de los $\left\lceil \left(1-\alpha_q\right)\cdot n_G \right\rceil$ estadísticos menos alejados respecto de su centro. Este *core* estará determinado por un parámetro de localización, $\overline{Z}_{\alpha_q - core}$, y un parámetro de escala, $\sigma^2_{\alpha_q - core}$, estimados a partir de las observaciones pertenecientes a dicho núcleo según la metodología descrita en el capítulo 3.

1) Calcular el estadístico test para cada uno de los genes. Se denota en general por Z_g, $g = 1, 2, ..., n_G$, haciendo referencia en este problema bien a $F_{g,core}$ o bien a $t_{gk,core}$.

2) Construir el correspondiente estadístico ordenado, $Z_{(1)} \le Z_{(2)} \le ... \le Z_{(n_G)}$

3) Calcular la probabilidad de que cada gen sea EE a partir de la distribución normal con parámetros estimados a partir del α_q -recorte imparcial de la muestra de estadísticos $\left\{Z_g\right\}_{1 \le g \le n_G}$.

4) Considerar B permutaciones de los valores observados en cada gen. Para cada permutación b, calcular el estadístico $\left\{Z_g^{*b}\right\}_{1 \le g \le n_G}$ y el correspondiente estadístico ordenado, $Z_{(1)}^{*b} \le Z_{(2)}^{*b} \le ... \le Z_{(n_G)}^{*b}$. En este punto existen dos opciones,

> **4.1)** Tener en cuenta los n_G genes, igual que en el método SAM.
>
> **4.2)** Corregir la estimación de la distribución nula. Para cada una de las permutaciones, calcular los estadísticos pesados $\left\{Z'^{*b}_g\right\}_{1 \le g \le n_G}$ como una muestra aleatoria con reemplazamiento de cada vector de estadísticos $\left\{Z_g^{*b}\right\}_{1 \le g \le n_G}$ utilizando como probabilidades de inclusión las obtenidas en el punto 3.

5) Para cada $g = 1, 2, ..., n_G$, estimar el estadístico esperado,

> **5.1)** teniendo en cuenta los n_G genes, exactamente igual que en la metodología SAM
>
> **5.2)** a partir de los estadísticos pesados: $\overline{Z}'_{(g)}$ definido en (4.15)

6) Determinar los genes diferencialmente expresados para un conjunto de umbrales Δ, en el rango de posibles valores de Z, o bien a partir de un número de genes DE fijado. Para cada umbral los pasos a seguir son los siguientes

> **6.1)** Encontrar el primer $g = g_1$ tal que $Z_{(g)} - \overline{Z}_{(g)} \le \Delta$ o $Z_{(g)} - \overline{Z}'_{(g)} \le \Delta$. Todos los contrastes g con $Z_{(g)} < Z_{(g_1)}$ serán considerados significativos negativos. El punto de corte inferior $\kappa_1(\Delta)$, se define como el mayor Z_g entre todos los contrastes significativos negativos.
>
> **6.2)** Encontrar el primer $g = g_2$ tal que $Z_{(g)} - \overline{Z}_{(g)} \ge \Delta$ o $Z_{(g)} - \overline{Z}'_{(g)} \ge \Delta$. Todos los contrastes g con $Z_{(g)} > Z_{(g_2)}$ serán considerados significativos positivos. El punto de corte superior $\kappa_2(\Delta)$, se define como el menor Z_g entre todos los contrastes significativos positivos.
>
> **6.3)** Si $\kappa_1(\Delta) > \kappa_2(\Delta)$ establecer $\kappa_1(\Delta) = \kappa_2(\Delta) = 0$

7) Estimar π_0 utilizando la expresión (4.2). En el caso del estadístico $F_{g,core}$, que sólo puede tomar valores positivos, q_1 y q_2 son los percentiles 0 y 50, respectivamente, de los estadísticos permutados.

8) Estimar FDR para cada umbral Δ según la expresión (4.1) o (4.16), o el p-valor ajustado propuesto en **[Dudoit et al, 2002b]**.

Figura 4.1. Procedimiento para resolver el problema de las comparaciones múltiples. Se señalan las modificaciones incluidas en el método SAM para corregir el sesgo en la estimación de la FDR.

2) La probabilidad de que un determinado gen g sea EE será

$$q_g = p\left(\left|Z_g\right| \ge \left|Z_g\right|\right) \tag{4.13}$$

con $\mathrm{Z}_g \to N\left(\bar{Z}_{\alpha_q - core}, {\sigma_{\alpha_q - core}}^2\right)$. En el caso del contraste por pares, se estima una única probabilidad por gen utilizando,

$$q_g = p\left(\left|\mathrm{Z}_g\right| \geq \min_{1\leq k\leq K}\left(\left|Z_{gk}\right|\right)\right) \tag{4.14}$$

Para incorporar estos pesos al método SAM de estimación de la FDR, se definen los estadísticos pesados $\left\{Z'^{*b}_g\right\}_{1\leq g\leq n_G}$ como los estadísticos $\left\{Z^{*b}_g\right\}_{1\leq g\leq n_G}$ procedentes de una muestra aleatoria con reemplazamiento de genes con probabilidades de inclusión $\left\{q_g \Big/ \sum_{g=1}^{n_G} q_g\right\}$. Considerando el correspondiente estadístico ordenado, $Z'^{*b}_{(1)} \leq Z'^{*b}_{(2)} \leq \ldots \leq Z'^{*b}_{(n_G)}$, para definir la lista de genes DE se compara cada estadístico observado, $Z_{(g)}$, con su correspondiente estadístico esperado calculado utilizando la expresión ,

$$\bar{Z}'_{(g)} = \frac{\sum_{b=1}^{B} Z'^{*b}_{(g)}}{B} \tag{4.15}$$

Para estimar la FDR se utilizará la expresión (4.1) sustituyendo $\widehat{FP}$ por,

$$\widetilde{FP} = \frac{\sum_{b=1}^{B} \#\left\{Z'^{*b}_g : Z'^{*b}_g \leq \kappa_1(\Delta) \text{ o } Z^{*b}_g \geq \kappa_2(\Delta)\right\}}{B} \tag{4.16}$$

donde $\kappa_1(\Delta)$ y $\kappa_2(\Delta)$ se calculan de la manera estándar, como el máximo (mínimo) estadístico observado entre todos los contrastes significativos negativos (positivos).

4.3. Evaluación del método propuesto

4.3.1. Datos simulados

Se simulan datos de expresión con $n_G(=1000)$ genes en $K(=20)$ clases con $n_k(=5)$ réplicas en cada una. El nivel de expresión del gen g se obtiene de una distribución normal con varianza σ_g^2 y media $\mu_{gk}\sigma_g$. Los G_1 primeros genes estarán sobre-expresados en 1, 2 o 3 clases. Para los G_0 genes EE, $\mu_{gk} = 0 \quad \forall k = 1,2,\ldots,K$. Para los G_1 genes DE, $\mu_{gk} \in \{2,4,6\}$, según los patrones de la tabla 4.1. El conjunto de n_G varianzas se genera como una muestra aleatoria de las varianzas observadas en el *dataset* de Tejidos Humanos

detallado en apéndice B (B.1) y etiquetado como $32.\mathrm{HUM}$. Las clases DE se asignan aleatoriamente en cada gen. El *core* se establece a partir de un recorte del 30%. Para tener en cuenta la variabilidad del estadístico, los resultados que se muestran son media de $nsim(=10)$ conjuntos simulados.

Tabla 4.1. Patrones de expresión de los G_1 genes con expresión diferencial

	C1	C2	C3
k1	2	0	0
k2	4	0	0
k3	6	0	0
k4	2	2	0
k5	2	4	0
k6	2	6	0
k7	4	4	0
k8	4	6	0
k9	6	6	0
k10	2	2	2
k11	2	2	4
k12	2	2	6
k13	2	4	4
k14	2	4	6
k15	2	6	6
k16	4	4	4
k17	4	4	6
k18	4	6	6
k19	6	6	6

Se generan 3 conjuntos de datos cuyas características se resumen en la tabla 4.2. Las principales diferencias entre ellos tienen que ver con el número de genes DE.

Tabla 4.2. Parámetros de las simulaciones

Parámetro	**Simulación A**	**Simulación B**	**Simulación C**
# genes, n_G	1000	1000	1000
# genes DE, G_1	0	175	475
expresión media, μ_{gk}	$0\ \forall g,k$	$\{2,4,6\}$	$\{2,4,6\}$
# genes por patrón tabla 4.1	k1- k19: 0	k1-k3: 5 k4-k19: 10	k1- k19: 25
varianza por gen, σ_g^2	m.a.s. 32.HUM	m.a.s. 32.HUM	m.a.s. 32.HUM
# réplicas por grupo, n_g	5	5	5
# permutaciones, B	100	100	100
# simulaciones, $nsim$	10	10	10

Se muestran los resultados para los dos contrastes propuestos en este trabajo, **(i)** el contraste global en el que se consideran el $pk(=15\%)$ de las clases más diferencialmente expresadas de cada gen, y **(ii)** en un conjunto de genes seleccionados con el contraste global, se contrasta la hipótesis de igualdad de medias entre cada par gen-clase y el núcleo de expresión. En todos los casos se fija $\delta_0 = 1$.

Tabla 4.3. Estimación π_0.

		Simulación A	Simulación B	Simulación C
Contraste global	π_0	1	0.825	0.475
	$\hat{\pi}_0$	0.9932 ± 0.0085	0.8306 ± 0.0249	0.5248 ± 0.0291
Contraste pares	# genes	1000	200	500
	α_q	0	0.3	0.3
	π_0	1	0.9172 ± 0.0034	0.9137 ± 0.0012
	$\hat{\pi}_0$	0.9948 ± 0.0053	0.9608 ± 0.0059	0.9593 ± 0.0054

Tabla 4.4. Simulación A. Estimaciones de FDR con el contraste global. Parámetros: $pk = 15\%$, $\delta_0 = 1$, $s_0^{(global)} = 0.193$.

# genes DE	$\widehat{FP}$ (# genes ± DT)	$\widehat{FDR}$ (% ± DT)
50	47.47 ± 7.04	94.94 ± 14.08
100	96.84 ± 10.04	96.84 ± 10.04
150	146.58 ± 9.15	97.72 ± 6.1
160	156.84 ± 9.44	98.02 ± 5.9
170	166.88 ± 10.54	98.17 ± 6.2
180	176.7 ± 10.21	98.17 ± 5.67
190	186.74 ± 12.16	98.28 ± 6.4
200	195.88 ± 10.08	97.94 ± 5.04
250	245.06 ± 11.23	98.03 ± 4.49
300	297.57 ± 12.5	99.19 ± 4.17
350	346.21 ± 15.34	98.92 ± 4.38
400	395.15 ± 15.22	98.79 ± 3.8
450	446.64 ± 15.86	99.25 ± 3.52
500	497.08 ± 16.92	99.42 ± 3.38

En la tabla 4.3 se muestra el promedio de las estimaciones de π_0: proporción de genes EE, junto con su desviación típica en cada conjunto de simulación y contraste. En las simulaciones con algún gen DE se sobre-estima ligeramente esta proporción. Además, esta estimación es peor cuanto mayor es el número de genes realmente expresados. Sin

embargo, en todos los casos, asumir que la mitad de los estadísticos calculados con los datos permutados están en la región de rechazo, funciona razonablemente bien.

Las tablas 4.4 - 4.9 muestran la estimación de la FDR para cada simulación y para cada contraste. Para determinar si un contraste es significativo se fija el número de genes DE o el de pares gen-grupo DE en el contraste por pares. Se prueban diferentes umbrales, etiquetados en las tablas como "# genes DE". En cada caso, se muestran las siguientes medidas resumen, **(i)** la potencia media del test (potencia), estimada como 1 menos el promedio de la tasa de falsos negativos en cada una de las $nsim$ simulaciones, **(ii)** el número de genes (pares) DE detectados como significativos (VP), **(iii)** el número de genes (pares) EE detectados como significativos (FP), **(iv)** la estimación de FP según el método propuesto ($\widehat{FP}$), **(v)** la FDR calculada como el promedio de falsos positivos en cada una de las $nsim$ simulaciones, y **(vi)** su estimación ($\widehat{FDR}$).

En el caso de la simulación A (tablas 4.4 y 4.5), donde no hay genes DE, la potencia es el 100% independiente del número de genes DE fijado, el número de verdaderos positivos es 0, el de falsos positivos el número de genes DE y la FDR, 100%. Así, las tablas sólo muestran las estimaciones (iv) y (vi). En todos los casos, tanto la estimación de los falsos positivos, como la estimación de la FDR se aproxima mucho a su verdadero valor.

Tabla 4.5. Simulación A. Estimaciones de FDR con el contraste para el par. Parámetros: $\delta_0 = 1$, $\alpha_q = 0$, $s_0^{(par)} = 0.1336$.

# genes DE	$\widehat{FP}$ (# genes ± DT)	$\widehat{FDR}$ (% ± DT)
300	292.61 ± 23.21	97.54 ± 7.74
325	318.51 ± 21.41	98 ± 6.59
350	341.83 ± 21.36	97.66 ± 6.1
375	368.91 ± 22.25	98.37 ± 5.93
400	394.3 ± 21.23	98.58 ± 5.31
425	418.7 ± 23.48	98.52 ± 5.53
450	444.38 ± 22.8	98.75 ± 5.07
475	470.41 ± 25.19	99.03 ± 5.3
500	491.68 ± 25.85	98.34 ± 5.17
550	540.2 ± 29.46	98.22 ± 5.36
600	587.56 ± 25.39	97.93 ± 4.23
700	683.97 ± 24.1	97.71 ± 3.44
800	782.07 ± 29.29	97.76 ± 3.66
900	879.97 ± 34.94	97.77 ± 3.88
1000	979.29 ± 37.35	97.93 ± 3.73

Tabla 4.6. Simulación B. Estimaciones de FDR con el contraste global. Parámetros: $pk = 15\%$, $\delta_0 = 1$, $s_0^{(global)} = 0.2206$.

# genes DE	*Potencia* (%)	*VP* (# genes ± DT)	$\widehat{FP}$ (# genes ± DT)	*FP* (# genes ± DT)	$\widehat{FDR}$ (% ± DT)	*FDR* (% ± DT)
50	29.86	50 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
100	59.72	100 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
150	88.97	149 ± 1.56	1.31 ± 1.03	1 ± 1.56	0.87 ± 0.69	0.67 ± 1.04
160	92.55	155 ± 2.62	5.53 ± 1.83	5 ± 2.62	3.46 ± 1.15	3.12 ± 1.64
170	94.99	159.1 ± 3.28	12.81 ± 3.01	10.9 ± 3.28	7.53 ± 1.77	6.41 ± 1.93
180	96.06	160.9 ± 3.35	19.18 ± 4.57	19.1 ± 3.35	10.66 ± 2.54	10.61 ± 1.86
190	96.84	162.2 ± 3.43	27.35 ± 3.21	27.8 ± 3.43	14.39 ± 1.69	14.63 ± 1.8
200	97.73	163.7 ± 3.71	35.63 ± 2.7	36.3 ± 3.71	17.82 ± 1.35	18.15 ± 1.86
250	98.68	165.3 ± 4.06	83.2 ± 7.35	84.7 ± 4.06	33.28 ± 2.94	33.88 ± 1.62
300	99.22	166.2 ± 3.29	134.37 ± 9.94	133.8 ± 3.29	44.79 ± 3.31	44.6 ± 1.1
350	99.34	166.4 ± 3.34	183.34 ± 15.12	183.6 ± 3.34	52.38 ± 4.32	52.46 ± 0.95
400	99.58	166.8 ± 3.19	232.12 ± 18.43	233.2 ± 3.19	58.03 ± 4.61	58.3 ± 0.8
450	99.82	167.2 ± 3.26	282.49 ± 17.05	282.8 ± 3.26	62.77 ± 3.79	62.84 ± 0.72
500	99.82	167.2 ± 3.26	332.66 ± 22.75	332.8 ± 3.26	66.53 ± 4.55	66.56 ± 0.65

Tabla 4.7. Simulación B. Estimaciones de FDR con el contraste para el par a partir de 200 genes seleccionados en el contraste global. Parámetros: $\delta_0 = 1$, $\alpha_q = 0.3$, $s_0^{(par)} = 0.1336$.

# genes DE	*Potencia* (%)	*VP* (# genes ± DT)	$\widehat{FP}$ (# genes ± DT)	*FP* (# genes ± DT)	$\widehat{FDR}$ (% ± DT)	*FDR* (% ± DT)
300	86.59	286.4 ± 5.72	14.29 ± 5.02	13.6 ± 5.72	4.76 ± 1.67	4.53 ± 1.91
325	90.18	298.4 ± 7.85	25.31 ± 7.16	26.6 ± 7.85	7.79 ± 2.2	8.18 ± 2.41
350	92.76	307 ± 9.63	40.88 ± 8.57	43 ± 9.63	11.68 ± 2.45	12.29 ± 2.75
375	94.8	313.8 ± 11.13	61.42 ± 12.29	61.2 ± 11.13	16.38 ± 3.28	16.32 ± 2.97
400	96.18	318.4 ± 12.4	81.91 ± 13.46	81.6 ± 12.4	20.48 ± 3.36	20.4 ± 3.1
425	97.22	321.9 ± 13.49	109.05 ± 16.84	103.1 ± 13.49	25.66 ± 3.96	24.26 ± 3.18
450	97.77	323.7 ± 13.2	138.16 ± 19.91	126.3 ± 13.2	30.7 ± 4.43	28.07 ± 2.93
475	98.46	326 ± 13.61	168.37 ± 24.89	149 ± 13.61	35.45 ± 5.24	31.37 ± 2.86
500	98.92	327.5 ± 13.04	200.53 ± 25.73	172.5 ± 13.04	40.11 ± 5.15	34.5 ± 2.61
550	99.46	329.3 ± 13.38	265 ± 22.79	220.7 ± 13.38	48.18 ± 4.14	40.13 ± 2.43
600	99.73	330.2 ± 13.8	328.82 ± 24.78	269.8 ± 13.8	54.8 ± 4.13	44.97 ± 2.3
700	99.84	330.6 ± 14.14	472.7 ± 27.07	369.4 ± 14.14	67.53 ± 3.87	52.77 ± 2.02
800	99.91	330.8 ± 13.88	595.36 ± 26.05	469.2 ± 13.88	74.42 ± 3.26	58.65 ± 1.73
900	99.97	331 ± 13.63	715.28 ± 31.73	569 ± 13.63	79.48 ± 3.53	63.22 ± 1.51
1000	100	331.1 ± 13.72	837.92 ± 35.93	668.9 ± 13.72	83.79 ± 3.59	66.89 ± 1.37

Tabla 4.8. Simulación C. Estimaciones de FDR con el contraste global. Parámetros: $pk = 15\%$, $\delta_0 = 1$, $s_0^{(global)} = 0.2206$.

# genes DE	*Potencia* (%)	*VP* (# genes ± DT)	$\widehat{FP}$ (# genes ± DT)	*FP* (# genes ± DT)	$\widehat{FDR}$ (% ± DT)	*FDR* (% ± DT)
300	66.55	300 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
400	88.49	398.9 ± 1.2	0.85 ± 0.49	1.1 ± 1.2	0.21 ± 0.12	0.28 ± 0.3
425	93.08	419.6 ± 2.22	4.65 ± 1.37	5.4 ± 2.22	1.09 ± 0.32	1.27 ± 0.52
450	95.92	432.4 ± 3.37	14.64 ± 3.59	17.6 ± 3.37	3.25 ± 0.8	3.91 ± 0.75
475	97.69	440.4 ± 4.09	29.04 ± 4.25	34.6 ± 4.09	6.11 ± 0.9	7.28 ± 0.86
500	98.6	444.5 ± 3.5	46.6 ± 6.14	55.5 ± 3.5	9.32 ± 1.23	11.1 ± 0.7
550	99.51	448.6 ± 3.5	87.49 ± 9.78	101.4 ± 3.5	15.91 ± 1.78	18.44 ± 0.64
600	99.71	449.5 ± 3.89	129.93 ± 11.34	150.5 ± 3.89	21.65 ± 1.89	25.08 ± 0.65
650	99.87	450.2 ± 4.26	178.03 ± 15.4	199.8 ± 4.26	27.39 ± 2.37	30.74 ± 0.66
700	99.93	450.5 ± 4.25	225.86 ± 25.18	249.5 ± 4.25	32.27 ± 3.6	35.64 ± 0.61
750	99.96	450.6 ± 4.14	276.93 ± 29.83	299.4 ± 4.14	36.92 ± 3.98	39.92 ± 0.55
800	99.98	450.7 ± 4.19	332.66 ± 29.93	349.3 ± 4.19	41.58 ± 3.74	43.66 ± 0.52
900	100	450.8 ± 4.18	431.86 ± 30.03	449.2 ± 4.18	47.98 ± 3.34	49.91 ± 0.46
1000	100	450.8 ± 4.18	524.42 ± 29.17	549.2 ± 4.18	52.44 ± 2.92	54.92 ± 0.42

Tabla 4.9. Simulación C. Estimaciones de FDR con el contraste para el par a partir de 500 genes seleccionados en el contraste global. Parámetros: $\delta_0 = 1$, $\alpha_q = 0.3$, $s_0^{(par)} = 0.1336$.

# genes DE	*Potencia* (%)	*VP* (# genes ± DT)	$\widehat{FP}$ (# genes ± DT)	*FP* (# genes ± DT)	$\widehat{FDR}$ (% ± DT)	*FDR* (% ± DT)
600	69.36	598.5 ± 0.71	1.73 ± 1.12	1.5 ± 0.71	0.29 ± 0.19	0.25 ± 0.12
700	80.11	691.3 ± 3.06	11.13 ± 5.65	8.7 ± 3.06	1.59 ± 0.81	1.24 ± 0.44
850	91.79	792.1 ± 10.63	72.67 ± 13.86	57.9 ± 10.63	8.55 ± 1.63	6.81 ± 1.25
900	93.84	809.8 ± 10.06	110.45 ± 15.01	90.2 ± 10.06	12.27 ± 1.67	10.02 ± 1.12
950	95.32	822.6 ± 9.06	156.57 ± 23.57	127.4 ± 9.06	16.48 ± 2.48	13.41 ± 0.95
1000	96.54	833.1 ± 9.33	205.05 ± 33.06	166.9 ± 9.33	20.51 ± 3.31	16.69 ± 0.93
1050	97.28	839.5 ± 10.28	261.35 ± 36.27	210.5 ± 10.28	24.89 ± 3.45	20.05 ± 0.98
1100	97.79	843.9 ± 10.94	322.28 ± 34.51	256.1 ± 10.94	29.3 ± 3.14	23.28 ± 0.99
1200	98.59	850.8 ± 11.44	452.85 ± 39.67	349.2 ± 11.44	37.74 ± 3.31	29.1 ± 0.95
1300	99.11	855.3 ± 11.75	582.81 ± 39.6	444.7 ± 11.75	44.83 ± 3.05	34.21 ± 0.9
1400	99.41	857.9 ± 11.75	713.88 ± 47.43	542.1 ± 11.75	50.99 ± 3.39	38.72 ± 0.84
1500	99.71	860.5 ± 11.27	839.61 ± 44.78	639.5 ± 11.27	55.97 ± 2.99	42.63 ± 0.75
1600	99.82	861.4 ± 11.48	968.15 ± 58.03	738.6 ± 11.48	60.51 ± 3.63	46.16 ± 0.72
1800	99.92	862.3 ± 11.16	1231.02 ± 56.25	937.7 ± 11.16	68.39 ± 3.12	52.09 ± 0.62
2000	99.97	862.7 ± 11.49	1490.05 ± 60.17	1137.3 ± 11.49	74.5 ± 3.01	56.86 ± 0.57

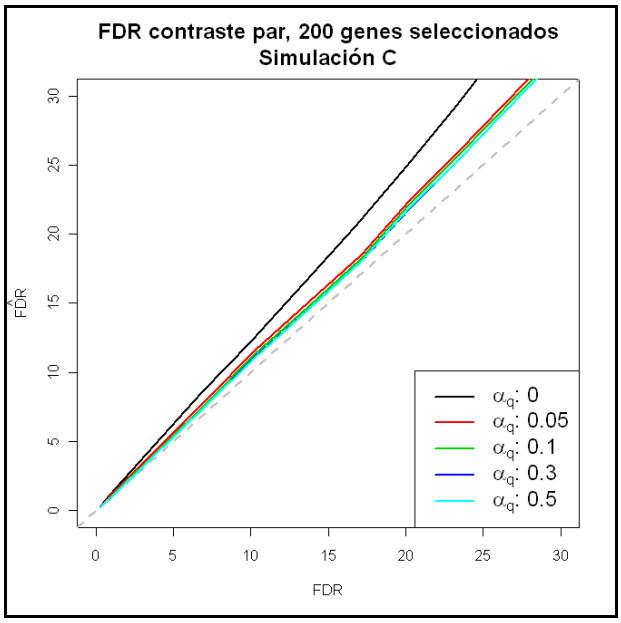


Figura 4.2. Comparación de la FDR real y estimada en el segundo contraste variando $\alpha_q \in \{0, 0.05, 0.1, 0.3, 0.5\}$, simulación C y a partir de un conjunto de 200 genes seleccionados con el contrate global.

En la simulación B (tablas 4.6 y 4.7), caracterizada por una proporción pequeña de genes DE, se obtiene una buena estimación de la FDR, especialmente en el contraste global. En todos los casos, cuando el umbral del número de genes DE se fija por encima del número simulado, la potencia media del contraste supera el 90%.
En cuanto a la simulación C (tablas 4.8 y 4.9), donde prácticamente la mitad de los genes están DE, la estimación de la FDR está sobre-estimando su valor real. Sin embargo, en todos los casos, para FDR pequeñas el método funciona aceptablemente bien. El hecho de que, en la práctica no es habitual fijar un nivel aceptable para la FDR por encima del 20%, resta importancia a este problema.

En el caso del contraste por pares, y puesto que se parte de un sub-conjunto de genes que resultan significativos con el contraste global, puede ser esencial corregir la estimación de la distribución nula a partir del recorte imparcial de nivel α_q del conjunto de estadísticos. En la figura 4.2 se muestra, para la simulación C, y a partir de 200 genes seleccionados con el contraste global por tener algún par significativo, como afecta utilizar o no este ajuste a la hora de estimar la FDR. Cuando $\alpha_q = 0$, el estimador de la FDR sobre-estima su verdadero valor y el error se incrementa a medida que la tasa es mayor. Claramente se observa que la estimación de la FDR es mejor cuando $\alpha_q > 0$. Por otra parte, también destaca que la elección de este parámetro no afecta demasiado a la estimación final, incluso en el caso de niveles de recorte muy pequeños, que necesariamente deben consideran como buenos estadísticos asociados a genes DE.

Un aspecto que también es valorable es el porcentaje de pares gen-clase detectados como DE por el método completo. En la tabla 4.10 se muestran los resultados según el nivel de expresión del par para las simulaciones B y C, aquellas que tienen cierta expresión diferencial. Para decidir si un par es considerado o no como significativo, se utiliza el p-valor ajustado, $\hat{\tilde{p}}_g$, definido en **[Dudoit et al, 2002b]**, fijando un nivel de 0.05. Además, y con el objetivo de evaluar el efecto de s_0, se consideran dos escenarios, en un caso se considera $\delta_0 = 0$ y en otro $\delta_0 = 1$. En general, cuando se fija $\delta_0 = 0$, aparecen más falsos positivos. Aunque en términos de porcentaje la diferencia entre considerar $\delta_0 = 0$ o $\delta_0 = 1$ es pequeña, en la simulación B, se traduce en una diferencia de aproximadamente 60 pares más, declarados como significativos sin serlo cuando $\delta_0 = 0$. En la simulación C, caracterizada por tener la mitad de los genes con algún par DE, esta diferencia es de 149 pares. Los pares con expresión diferencial se detectan bien con cualquiera de las dos opciones.

Tabla 4.10. Pares declarados como DE ($\hat{\tilde{p}}_g \leq 0.05$) en cada nivel de expresión. Parámetros primera etapa: $pk = 15\%$ y $\alpha_q = 0$. Parámetros segunda etapa: $\alpha_q = 0.3$

μ	Simulación B			Simulación C		
	# pares	% pares DE ± DT		# pares	% pares DE ± DT	
		$\delta_0 = 0$	$\delta_0 = 1$		$\delta_0 = 0$	$\delta_0 = 1$
0	19565	0.94 ± 0.10	0.63 ± 0.17	18875	1.69 ± 0.25	0.90 ± 0.26
2	145	68.34 ± 4.40	68.34 ± 6.83	375	67.60 ± 3.76	66.27 ± 3.48
4	145	89.66 ± 3.36	91.10 ± 3.04	375	91.39 ± 1.34	91.95 ± 2.38
6	145	97.59 ± 1.60	98.48 ± 1.37	375	96.61 ± 0.99	97.28 ± 0.80

4.3.2. Aplicación al *dataset* de Tejidos Humanos

Se analiza el conjunto de muestras procedente de 32 tejidos, glándulas y órganos de individuos sanos descrito en el apéndice B (B.1). El objetivo será determinar genes cuyos patrones de expresión sean característicos de tipos de tejidos humanos. Dentro del conjunto de genes DE en uno o varios tipos de tejido, hay que distinguir entre genes selectivos de tejidos (TS, *Tissue-Selective* genes) predominantemente expresados en unos pocos tipos de tejidos simultáneamente, y genes específicos de tejidos (TSp, *Tissue-Specificity* genes), que se expresan en un único tipo de tejido.

Existen muchos métodos utilizados con el objetivo de encontrar genes TS, tradicionalmente basados en el análisis de un único gen. Recientemente, aparecen estudios con datos procedentes de diferentes tecnologías de alto rendimiento, como por ejemplo **[Yu et al, 2006]** que utilizan marcadores de secuencia expresada (EST, *Expressed Sequence Tag*), o **[Siu et al,**

2001] y **[Kouadjo et al, 2007]** que basan sus hallazgos en el análisis en serie de expresión génica (SAGE, *Serial Analysis of Gene Expression*).

Sin embargo, y a pesar de la cantidad de trabajos y metodología específicamente desarrollada para datos procedentes del análisis de microarrays, existen pocos estudios centrados en la resolución de este problema. Alguna metodología utilizada para este objetivo ha sido, el contraste t-Student combinado con el análisis de componentes principales (PCA) utilizados en **[Hsiao et al, 2001]** y **[Misra et al, 2002]**, para analizar 19 y 40 tipos de tejidos humanos respectivamente; el análisis clúster jerárquico utilizado en **[Shyamsundar et al, 2005]** para analizar la expresión en 35 tipos de tejidos; el método de Tukey-Kramer empleado en el problema de las comparaciones múltiples aplicado a 97 tipos de tejidos humanos en **[Liang et al, 2006]**; o distintos *scores* aplicados a meta-análisis, como el propuesto en **[Wang et al, 2010]** aplicado a datos procedentes de 131 conjuntos, o el trabajo de **[Chang et al, 2011]** aplicado a datos de 104 conjuntos en los que están representados 43 tipos de tejidos diferentes.

A continuación se muestran los resultados obtenidos con el método en dos fases propuesto en este capítulo. Originalmente, el procedimiento sirve para la búsqueda de la expresión selectiva, aunque resuelto este problema, y dado que la expresión específica es un caso particular de la expresión selectiva, es posible encontrar también genes específicos utilizando un *score* sencillo. El núcleo de expresión en cada gen se obtiene a partir del recorte imparcial con $\alpha = 0.3$.

4.3.2.1. Primer paso: contraste global

Se contrasta la hipótesis de expresión diferencial en el $pk = 30\%$ de las clases más expresadas respecto al *core*, utilizando el estadístico $F_{g,core}$ definido en (4.4). De la expresión (4.8), y fijando el umbral para la diferencia de medias biológicamente relevante en $\delta_0 = 2$, se obtiene que $s_0^{(global)} = 0.9278$. Se consideran significativos aquellos genes cuyo p-valor ajustado, $\hat{\tilde{p}}_g$, es menor que 0.01, obteniendo una lista de 5101 genes con algún tipo de expresión diferencial, que suponen el 25.29% de los 20172 genes analizados.

En la figura 4.3 se compara este listado inicial, denominado L_n, con el que se obtendría con el filtrado no específico propuesto en **[Prieto et al, 2008]**. Este filtro está basado en la expresión mínima y la variabilidad entre muestras, y en lo que sigue se le denota por L_t. Dado el nivel de expresión de las n observaciones en el gen g, $\{y_{ga}\}_{1\le a\le n}$, se identifican como genes no informativos, aquellos que cumplen dos condiciones: **(i)** genes con una diferencia máxima de expresión menor que la mediana de todas las diferencias máximas, y **(ii)** genes con una media de expresión menor que la mediana de todas las expresiones

medias. En este conjunto de datos, L_t está formado por 5955 genes (29.52% de los genes totales).

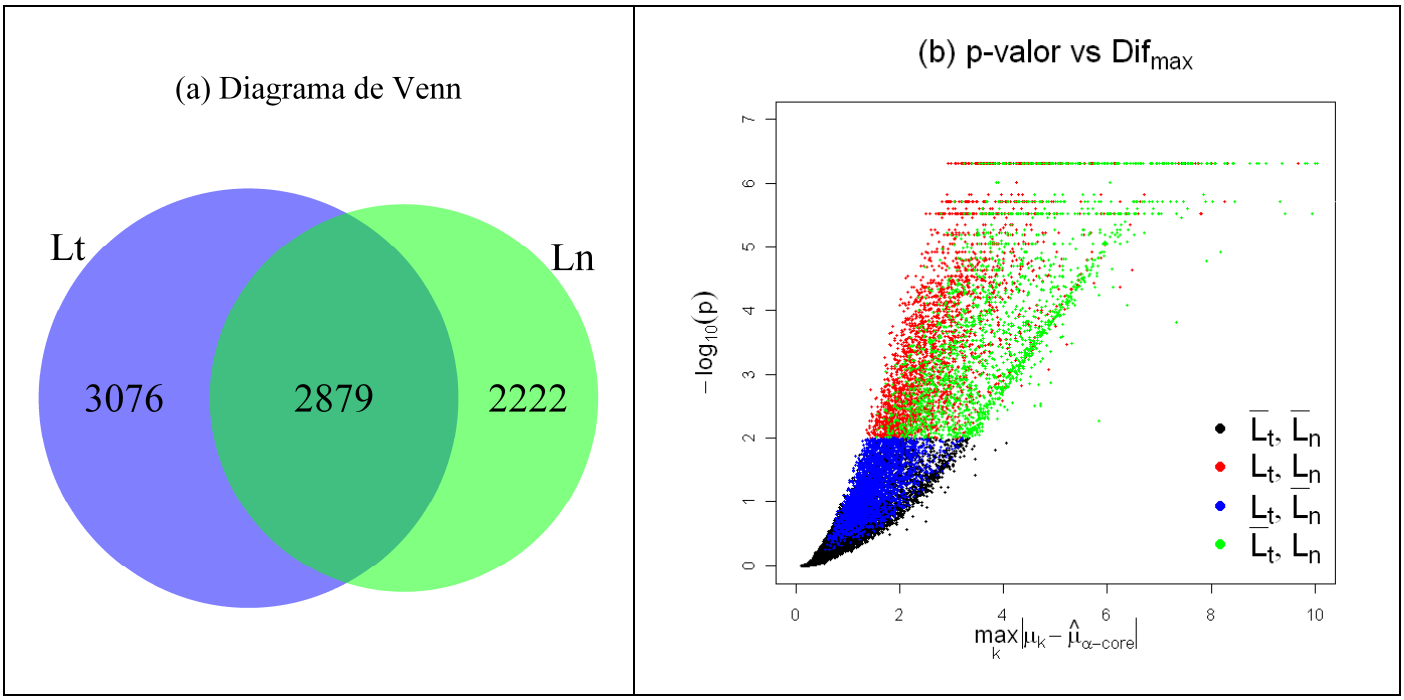


Figura 4.3. Comparación de L_t y L_n. (a) Diagrama de Venn, (b) Diagrama de dispersión que relaciona el p-valor global con la máxima diferencias de medias respecto del core. Se representan en azul los genes incluidos exclusivamente en L_t, en verde los incluidos sólo en L_n, en rojo los incluidos en ambos y en negro los declarados no informativos por ambos métodos.

A pesar de que ambos conjuntos son similares en cuanto al número de genes considerados, apenas coinciden en la mitad de ellos (figura 4.3 (a)). En general, y tal y como se muestra en la figura 4.3 (b), los genes exclusivamente seleccionados por L_t, representados en azul, se caracterizan por tener diferencia de media pequeña, dejando fuera genes con algunos valores muy extremos (representados en verde). En la figura 4.4 se muestran algunos casos extremos. El filtro L_n deja pasar genes del tipo ENSG00000237647 (AC129915.2) (figura 4.4 (a)) caracterizado por un nivel de expresión y variabilidad media, pero que claramente no muestra ningún patrón de expresión diferencial. En cambio, no pasarían genes del tipo ENSG00000176840 (C19orf30) (figura 4.4 (b)) caracterizado por tener un nivel de expresión mayoritario muy bajo con una variabilidad muy pequeña, que sin embargo se expresa claramente en al menos dos clases.

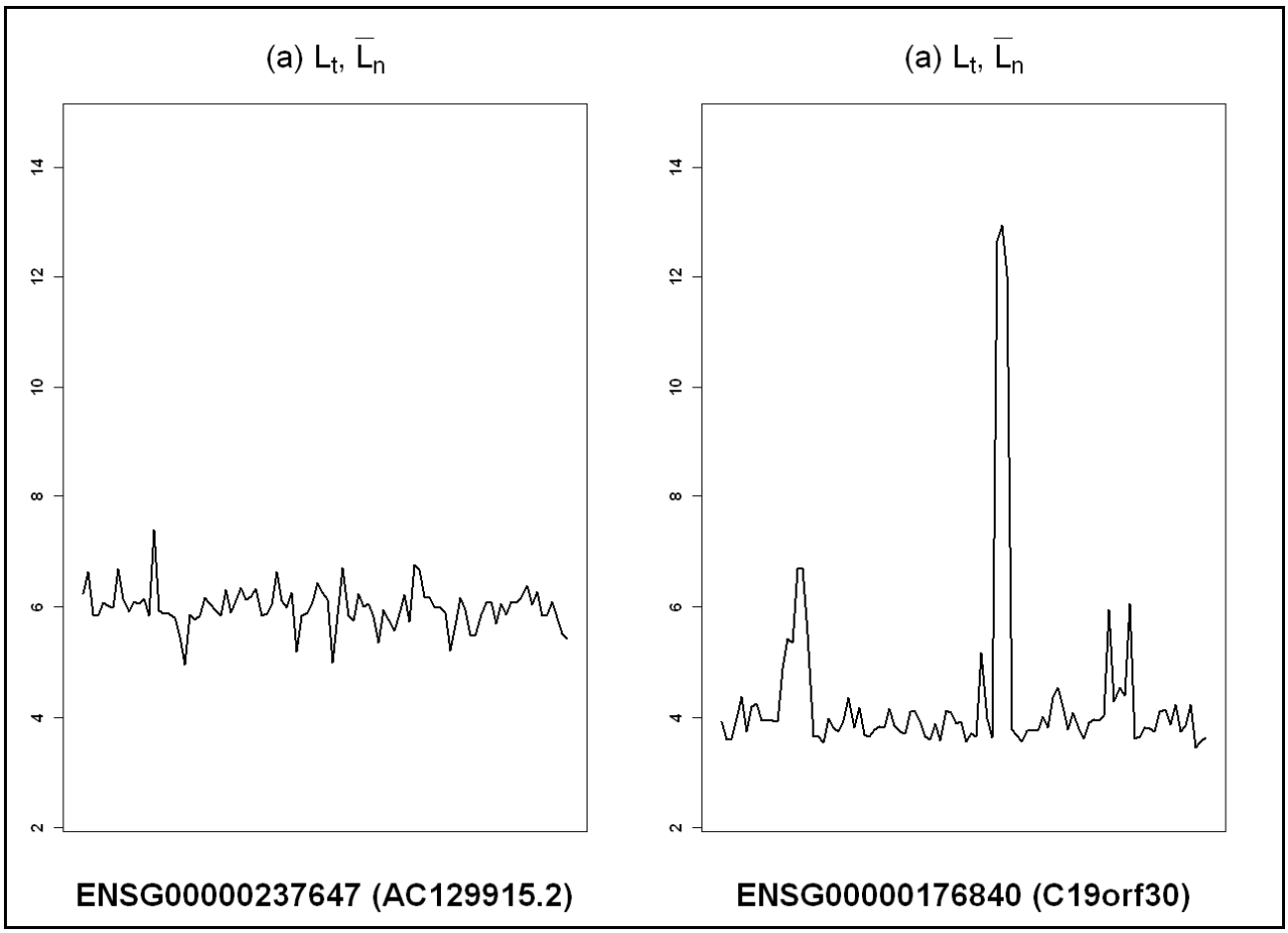


Figura 4.4. (a) Perfil de expresión del gen ENSG00000237647 (AC129915.2), incluido en L_t y excluido de L_n. (b) Perfil de expresión del gen ENSG00000176840 (C19orf30), incluido en L_n y excluido de L_t.

4.3.2.2. Segundo paso: contraste por pares

Se contrasta la hipótesis de igualdad de expresión media en cada par gen-clase respecto del núcleo de expresión correspondiente utilizando el estadístico $t_{gk,core}$ definido en (4.10). Se consideran los genes incluidos en L_n, fijando los parámetros $\delta_0 = 2$ y $\alpha_q = 0.3$. De la expresión (4.12) se obtiene que $s_0^{(par)} = 0.2116$. Se declaran significativos aquellos pares con $\hat{\tilde{p}}_g \leq 0.001$. El resultado es una lista de 8944 pares con algún tipo de expresión diferencial, que suponen el 56% de los 163232 pares analizados. De estos 8944 pares, 8253 (92.27%) están sobre-expresados y sólo 691 (7.73%) infra-expresados. En la tabla 4.11 se muestra el número de genes según el número de tejidos DE, y en la tabla 4.12 el número de pares significativos según el tipo de tejido estudiado. Los tipos de tejidos selectivamente expresados en más genes son testículo, médula ósea, corteza cerebral e hígado con porcentajes sobre el total de pares DE de 9.45%, 5.95%, 5.89% y 5.69%, respectivamente.

Los menos representados son la glándula mamaria, ovario y estómago, que no llegan al 1% del total de pares selectivamente expresados.

Tabla 4.11. Genes según el número de tejidos DE.

	# Tejidos											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Sobre-expresados	1078	1890	1064	537	283	126	61	28	17	8	7	2
Infra-expresados	4715	231	84	36	17	5	6	2	4	1	0	0
DE	814	1986	1124	577	314	140	73	33	21	8	7	4

En la figura 4.5 se representa el porcentaje de concordancia entre tipos de tejidos, en cuanto a los genes expresados en cada uno de ellos. Por encima del 50% de coincidencia están los grupos de los dos tipos de corazón, riñón y estómago, y además los grupos formados por tráquea y bronquio; mucosa oral y esófago; cerebelo, corteza cerebral y médula; nodo linfático, bazo y amígdala; y por último corazón, lengua y músculo esquelético. Estos niveles de concordancia indican que estas clases comparten ciertas funciones biológicas. En el lado opuesto, el tejido más específico, que comparte menos genes con el resto de grupos, es testículo, con un porcentaje de coincidencia máximo del 8%.

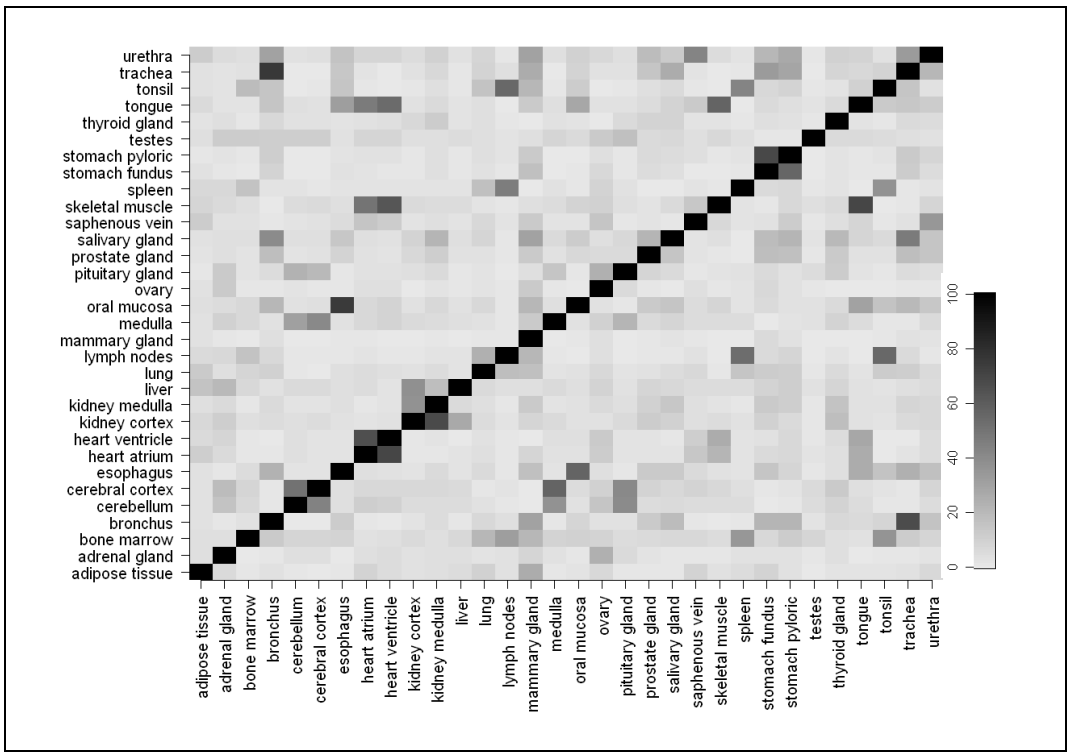


Figura 4.5. Porcentaje de coincidencia entre tipos de tejidos en cuanto a los genes expresados en cada uno de ellos.

Con el grupo de genes TS se ha realizado un análisis de enriquecimiento funcional utilizando las categorías de GO-BP, las vías de señalización de KEGG y las anotaciones en la base de

datos GAD. En la tabla 4.13 se muestran las 3 categorías más sobre-representadas para cada uno de los tejidos. En general, se puede decir que los resultados son coherentes desde el punto de vista biológico, ya que estas categorías están muy relacionadas con las funciones correspondientes a cada tipo de tejido.

Tabla 4.12. Pares DE según el tipo de tejido. Número de pares, mediana y rango del FC respecto del *core* para cada tipo de tejido.

	Sobre-expresados				Infra-expresados			
	# pares	Mediana FC	FC min	FC max	# pares	Mediana FC	FC min	FC max
Adipose tissue omental	165	2.781	1.419	8.263	0	.	.	.
Adrenal gland cortex	118	3.156	1.397	8.872	4	-2.865	-4.246	-2.478
Bone marrow	360	2.839	1.399	9.039	172	-2.972	-5.756	-1.803
Bronchus	153	3.384	1.379	9.127	0	.	.	.
Cerebellum	346	3.218	1.400	7.103	105	-2.916	-7.712	-1.809
Cerebral cortex	458	3.455	1.511	8.202	69	-3.323	-7.654	-1.742
Esophagus	296	3.177	1.314	10.537	0	.	.	.
Heart atrium	194	3.050	1.604	9.210	4	-2.745	-4.125	-1.993
Heart ventricle	172	3.251	1.557	9.755	15	-2.844	-7.386	-1.895
Kidney cortex	355	3.146	1.491	8.468	7	-3.296	-5.318	-3.059
Kidney medulla	194	3.000	1.517	8.798	2	-3.083	-3.482	-2.683
Liver	464	3.779	1.508	9.851	45	-2.851	-5.117	-1.828
Lung	207	2.984	1.552	7.908	1	-4.417	-4.417	-4.417
Lymph nodes	273	2.744	1.452	6.759	4	-3.025	-4.390	-2.177
Mammary gland	20	3.880	1.667	7.174	3	-3.197	-3.828	-2.813
Medulla	341	3.006	1.454	8.424	37	-3.002	-7.445	-1.727
Oral mucosa	378	3.295	1.556	10.525	6	-2.216	-3.808	-1.856
Ovary	55	3.006	1.709	7.640	10	-2.940	-4.741	-2.379
Pituitary gland	238	2.988	1.512	9.466	22	-2.911	-5.078	-2.317
Prostate gland	213	2.985	1.503	9.645	0	.	.	.
Salivary gland	305	3.056	1.450	10.414	10	-2.371	-3.509	-2.053
Saphenous vein	217	2.817	1.542	7.090	3	-4.132	-4.309	-3.738
Skeletal muscle	376	3.311	1.534	9.745	62	-2.598	-6.504	-1.851
Spleen	227	2.917	1.466	7.203	14	-3.759	-4.829	-2.145
Stomach fundus	66	3.177	1.678	9.715	1	-2.418	-2.418	-2.418
Stomach pyloric	82	3.650	1.648	8.780	0	.	.	.
Testes	764	3.777	1.568	9.917	81	-2.670	-4.895	-1.927
Thyroid gland	147	2.932	1.298	8.897	0	.	.	.
Tongue main corpus	359	3.246	1.393	9.708	0	.	.	.
Tonsil	262	3.151	1.454	9.319	14	-2.403	-3.659	-2.108
Trachea	174	3.527	1.626	8.980	0	.	.	.
Urethra	274	2.501	1.414	8.744	0	.	.	.

Tabla 4.13. Análisis de enriquecimiento funcional de los genes TS utilizando la herramienta DAVID. Se muestra la categoría más sobre-representada, o el grupo de categorías con niveles de significación similares.

	# genes anotados	GO biological process		KEGG pathway		Genetic association database	
		Categoría	# genes	Categoría	# genes	Categoría	# genes
Adipose tissue omental	162	GO:0009725~response to hormone stimulus	19	hsa03320:PPAR signaling pathway	11	diabetes, type 2	18
Adrenal gland cortex	116	GO:0008202~steroid metabolic process	15	hsa00100:Steroid biosynthesis	4		
		GO:0006694~steroid biosynthetic process	10				
Bone marrow	521	GO:0006955~immune response	73	hsa04510:Focal adhesion	19		
		GO:0006952~defense response	67				
		GO:0009611~response to wounding	52				
Bronchus	149	GO:0030855~epithelial cell differentiation	9				
Cerebellum	439	GO:0007268~synaptic transmission	38			epilepsy	10
		GO:0019226~transmission of nerve impulse	39			schizophrenia	26
		GO:0007267~cell-cell signaling	45				
Cerebral cortex	510	GO:0030182~neuron differentiation	43				
		GO:0007268~synaptic transmission	35				
		GO:0019226~transmission of nerve impulse	37				
Esophagus	293	GO:0007398~ectoderm development	30				
		GO:0008544~epidermis development	28				
		GO:0030855~epithelial cell differentiation	24				
Heart atrium	195	GO:0006936~muscle contraction	29	hsa05410:Hypertrophic cardiomyopathy (HCM)	18	cardiomyopathy	15
		GO:0003012~muscle system process	29	hsa05414:Dilated cardiomyopathy	17		
		GO:0007517~muscle organ development	26	hsa04260:Cardiac muscle contraction	13		
Heart ventricle	182	GO:0006936~muscle contraction	34	hsa05414:Dilated cardiomyopathy	19	cardiomyopathy	15
		GO:0003012~muscle system process	34	hsa05410:Hypertrophic cardiomyopathy (HCM)	17		
		GO:0006941~striated muscle contraction	19	hsa04260:Cardiac muscle contraction	13		
Kidney cortex	356	GO:0016054~organic acid catabolic process	22	hsa00280:Valine, leucine and isoleucine degradation	12		
		GO:0046395~carboxylic acid catabolic process	22				
		GO:0055114~oxidation reduction	48				
Kidney medulla	191	GO:0009952~anterior/posterior pattern formation	12	hsa04960:Aldosterone-regulated sodium reabsorption	8	hypertension	15
Liver	495	GO:0055114~oxidation reduction	88	hsa04610:Complement and coagulation cascades	34	lipids	12
		GO:0002526~acute inflammatory response	35	hsa00980:Metabolism of xenobiotics by cytochrome P450	21	lipoprotein	12
		GO:0016054~organic acid catabolic process	35	hsa00982:Drug metabolism	21		
Lung	204	GO:0006955~immune response	41	hsa04514:Cell adhesion molecules	14	stroke, ischemic	13
		GO:0006952~defense response	35				
		GO:0006954~inflammatory response	26				
Lymph nodes	262	GO:0006955~immune response	71	hsa04514:Cell adhesion molecules	18	multiple sclerosis	18
		GO:0045321~leukocyte activation	33	hsa04640:Hematopoietic cell lineage	15		
		GO:0001775~cell activation	34	hsa04650:Natural killer cell mediated cytotoxicity	17		

Tabla 4.13. (Continuación)

	# genes anotados	GO biological process		KEGG pathway		Genetic association database	
		Categoría	# genes	Categoría	# genes	Categoría	# genes
Mammary gland	21						
Medulla	367	GO:0030182~neuron differentiation	28				
Oral mucosa	373	GO:0007398~ectoderm development GO:0008544~epidermis development GO:0030855~epithelial cell differentiation	38 35 22				
Ovary	62						
Pituitary gland	244			hsa04080:Neuroactive ligand-receptor interaction	13		
Prostate gland	204	GO:0048732~gland development	13				
Salivary gland	308						
Saphenous vein	216	GO:0007155~cell adhesion GO:0022610~biological adhesion GO:0006936~muscle contraction	34 34 17	hsa04510:Focal adhesion	16		
Skeletal muscle	427	GO:0006936~muscle contraction GO:0003012~muscle system process GO:0007517~muscle organ development	47 47 44	hsa04260:Cardiac muscle contraction hsa05410:Hypertrophic cardiomyopathy (HCM)	21 18 21	cardiomyopathy	9
Spleen	231	GO:0006955~immune response GO:0006952~defense response GO:0001775~cell activation	54 46 29	hsa04060:Cytokine-cytokine receptor interaction	20		
Stomach fundus	65	GO:0060429~epithelium development	8				
Stomach pyloric	80	GO:0060429~epithelium development	8				
Testes	771	GO:0019953~sexual reproduction GO:0007283~spermatogenesis GO:0048232~male gamete generation	77 62 62			infertility, male	9
Thyroid gland	144	GO:0010817~regulation of hormone levels	10				
Tongue main corpus	354	GO:0006936~muscle contraction GO:0003012~muscle system process GO:0007517~muscle organ development	40 40 35	hsa04260:Cardiac muscle contraction hsa05410:Hypertrophic cardiomyopathy (HCM) hsa05414:Dilated cardiomyopathy	16 14 14	cardiomyopathy	9
Tonsil	261	GO:0006955~immune response GO:0045321~leukocyte activation GO:0001775~cell activation	61 27 29	hsa04662:B cell receptor signaling pathway hsa05340:Primary immunodeficiency hsa04062:Chemokine signaling pathway	14 10 18		
Trachea	168	GO:0030855~epithelial cell differentiation	11	hsa00512:O-Glycan biosynthesis	4		
Urethra	268	GO:0001501~skeletal system development GO:0048598~embryonic morphogenesis GO:0007155~cell adhesion	25 24 36	hsa04510:Focal adhesion	14		

4.3.2.3. Expresión específica

Desde un punto de vista biomédico, encontrar genes específicos de ciertos tejidos es muy importante, puesto que serán los primeros candidatos a marcadores de enfermedad y potenciales dianas terapéuticas. Como ya se ha mencionado, la expresión específica podría entenderse como un caso particular de la expresión selectiva y se podría considerar genes TSp aquellos que muestren un único tejido significativo. Sin embargo, en ese listado de 1986 genes, existen casos en los que aparecen otros tejidos al borde de la significación, con niveles de expresión cercanos al que finalmente resulta significativo (figura 4.6 (a)). En estos casos hablar de genes TSp puede no resultar apropiado. También puede darse el caso contrario, y entre aquellos genes que muestran significación en más de una tipo de tejido, aparecer casos en los que las diferencias entre el tejido más expresado y el resto son extremadamente grandes (figura 4.6 (b)). Este tipo de perfiles podrían estar más cerca de la definición de expresión específica.

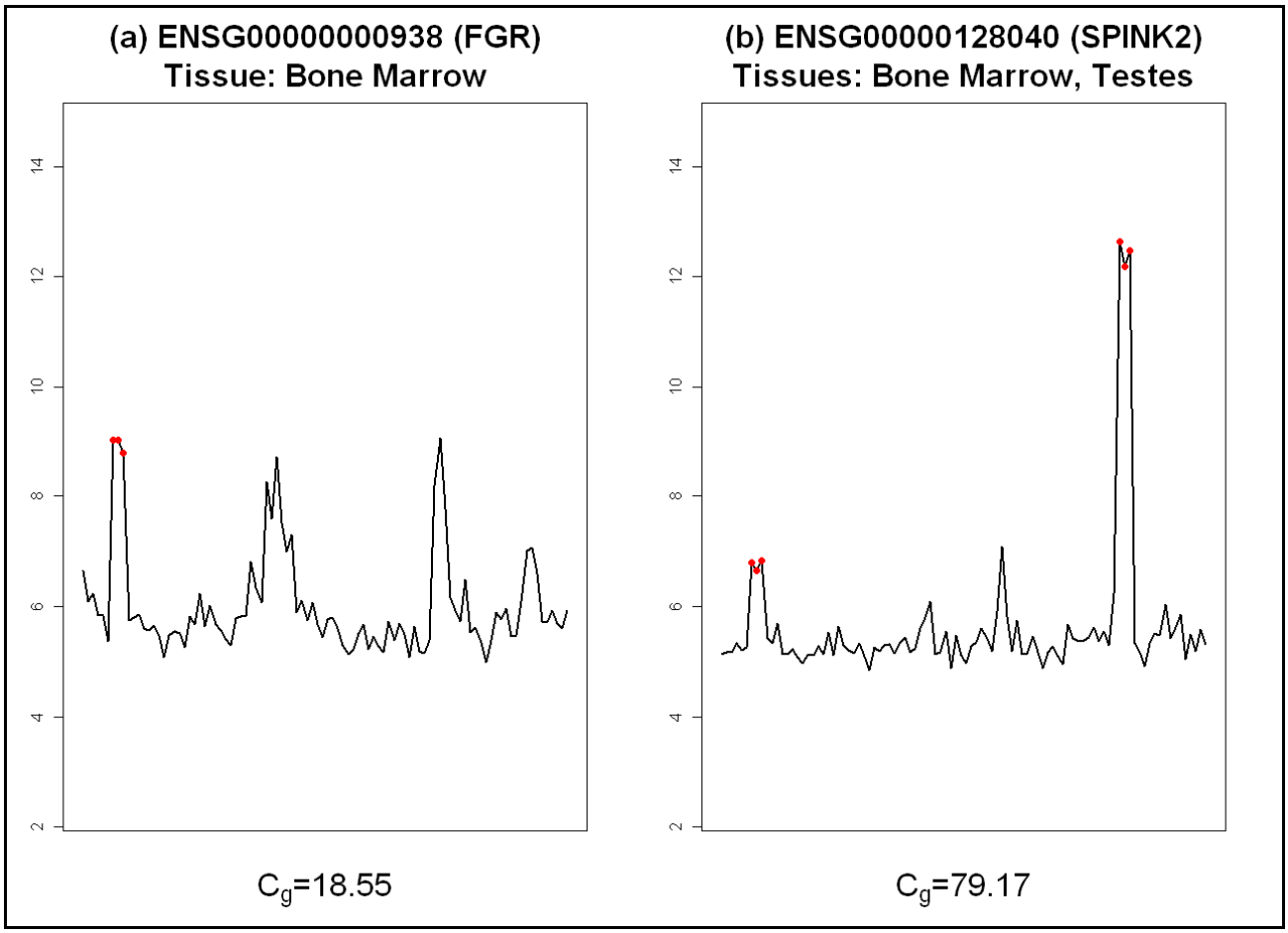


Figura 4.6. (a) Perfil de expresión del gen ENSG00000000938 (FGR), gen TS con un único tejido DE. (b) Perfil de expresión del gen ENSG00000128040 (SPINK2), gen TSp con 2 tejidos DE. Aparecen marcadas en rojo las tres réplicas del(os) grupo(s) que resulta(n) estadísticamente significativo(s).

Tabla 4.14. Genes TSp según el tipo de tejido. Número de genes, mediana y rango del FC respecto del core, y media y desviación típica del índice C_g en cada tipo de tejido.

	Genes TSp				
	# genes	Mediana FC	FC min	FC max	Media C_g± DT
Adipose tissue omental	7	3.524	2.674	5.047	62.894±8.5
Adrenal gland cortex	11	4.607	2.914	8.179	65.479±12.284
Bone marrow	57	4.700	2.521	9.039	76.098±16.773
Bronchus	0	.	.	.	.
Cerebellum	19	5.134	2.959	7.103	61.538±9.077
Cerebral cortex	23	3.621	2.814	5.928	60.191±9.315
Esophagus	2	4.774	3.203	6.345	56.183±2.242
Heart atrium	3	6.866	3.495	7.818	77.893±5.575
Heart ventricle	1	3.815	3.815	3.815	64.429
Kidney cortex	19	3.785	2.752	4.897	59.177±5.268
Kidney medulla	0	.	.	.	.
Liver	152	5.095	2.613	9.851	73.36±12.73
Lung	13	4.059	2.501	7.908	70.696±16.296
Lymph nodes	2	3.792	3.304	4.280	63.058±9.24
Mammary gland	2	5.708	4.380	7.036	68.232±15.279
Medulla	1	3.058	3.058	3.058	53.053
Oral mucosa	3	4.812	3.582	5.135	55.816±1.373
Ovary	3	3.901	3.045	4.115	63.977±7.335
Pituitary gland	27	5.288	2.917	9.466	78.342±13.642
Prostate gland	25	4.474	2.597	8.262	68.524±13.706
Salivary gland	18	4.945	2.870	10.040	64.073±11.63
Saphenous vein	4	3.367	2.915	4.129	59.784±8.674
Skeletal muscle	10	3.435	2.527	6.496	59.721±13.446
Spleen	7	3.805	2.824	5.727	59.374±6.086
Stomach fundus	1	5.224	5.224	5.224	74.026
Stomach pyloric	0	.	.	.	.
Testes	466	4.289	2.489	9.917	79.627±12.515
Thyroid gland	13	5.766	3.294	8.897	74.219±14.175
Tongue main corpus	1	3.104	3.104	3.104	68.018
Tonsil	0	.	.	.	.
Trachea	0	.	.	.	.
Urethra	0	.	.	.	.

Dada la importancia biológica de este tipo de gen, con el objetivo de obtener un listado de genes asociados específicamente a un tipo de tejido, se propone el uso del siguiente *score*,

$$C_g = \frac{\overline{Y}'_{g(1)} - \overline{Y}'_{g(2)}}{\overline{Y}'_{g(1)}} \cdot 100 \tag{4.17}$$

donde

$$\bar{Y}'_{gk} = \left| \bar{Y}_{gk} - \hat{\mu}_{g,0.3-smart} \right| \quad (4.18)$$

e $\bar{Y}'_{g(1)} \geq \bar{Y}'_{g(2)} \geq \ldots \geq \bar{Y}'_{g(K)}$. Esta medida puede interpretarse como el porcentaje de cambio en la expresión diferencial entre las dos clases más expresadas respecto del nivel de expresión mayoritario del gen. Así, valores altos indicarán genes que muestran un tejido claramente expresado respecto del resto.

Utilizando el criterio $C_g \geq 50\%$ se obtiene un listado de 887 genes, resumidos en la tabla 4.14. Los conjuntos de tejidos más específicos son testículo, que supone prácticamente el 53% de la lista completa, e hígado, que cuenta con el 17% del listado. En estos dos casos se realiza el mismo tipo de análisis de enriquecimiento funcional que el llevado a cabo para los genes TS. Las categorías más sobre-representadas se muestran en la tabla 4.15. Tanto los procesos biológicos, como las vías de señalización y enfermedades son biológicamente consistentes con el tipo de tejido.

Tabla 4.15. Análisis de enriquecimiento funcional de los genes TSp en los tejidos testículo e hígado. Se muestran las 10 categorías sobre-representadas.

	# genes anotados	GO biological process		KEGG pathway		Genetic association database	
		Categoría	# genes	Categoría	# genes	Categoría	# genes
Testes	411	GO:0019953~sexual reproduction	58			infertility, male	7
		GO:0007283~spermatogenesis	48			azoospermia; oligospermia	3
		GO:0048232~male gamete generation	48				
		GO:0007276~gamete generation	48				
		GO:0048609~reproductive process in a multicellular organism	50				
		GO:0032504~multicellular organism reproduction	50				
		GO:0048610~reproductive cellular process	25				
		GO:0007286~spermatid development	14				
		GO:0048515~spermatid differentiation	14				
		GO:0009566~fertilization	15				
Liver	145	GO:0009611~response to wounding	36	hsa04610:Complement and coagulation cascades	21	lipids	10
		GO:0002526~acute inflammatory response	19	hsa00982:Drug metabolism	15	lipoprotein	8
		GO:0042060~wound healing	20	hsa00830:Retinol metabolism	13	lipoproteins	7
		GO:0007596~blood coagulation	16	hsa00980:Metabolism of xenobiotics by cytochrome P450	13	cholesterol	8
		GO:0050817~coagulation	16	hsa00591:Linoleic acid metabolism	6	body mass; triglycerides; cholesterol, total; blood pressure; leptin; apoA1; apoA2; fasting blood sugar; fasting blood sugar	6
		GO:0007599~hemostasis	16	hsa03320:PPAR signaling pathway	8	myocardial infarction	11
		GO:0008202~steroid metabolic process	19	hsa00120:Primary bile acid biosynthesis	5	myocardial infarct	15
		GO:0050878~regulation of body fluid levels	16	hsa00140:Steroid hormone biosynthesis	6	thromboembolism, venous	9
		GO:0051605~protein maturation by peptide bond cleavage	13	hsa05020:Prion diseases	5	thrombosis, deep vein	7
		GO:0006954~inflammatory response	21	hsa00983:Drug metabolism	5	hyperlipidemia	6

Capítulo 5

Expresión diferencial en dos clases independientes con respuesta heterogénea: identificación de *outliers*

La investigación en cáncer ha sido uno de los motores del gran desarrollo que ha experimentado la tecnología relacionada con los microarrays en los últimos años. El objetivo fundamental de la búsqueda de expresión diferencial en investigaciones ligadas al cáncer, es identificar genes con capacidad diagnóstica, pronóstica o que constituyan potenciales dianas terapéuticas. Estos marcadores, suelen caracterizarse por presentar niveles de expresión superiores en muestras procedentes de células tumorales que en las muestras control. Se ha observado que en muchas situaciones, los oncogenes muestran patrones de activación heterogéneos **[Tomlins et al, 2005]**, correspondientes a la aparición de niveles de expresión más elevados, solamente en un sub-conjunto del total de la muestra de casos. Un ejemplo clásico es el del oncogen ERBB2 (HER2) **[Slamon et al, 1987]**, en el que la mediana de sus niveles de expresión muestra un modesto 1.3-*fold* respecto a las muestras de tejido sano, frente al 6-*fold* en el que se localiza el percentil 90.

Los métodos tradicionales de detección de expresión diferencial pueden ser menos eficientes en estas situaciones en las que aparece respuesta heterogénea en la muestras de tumores. Estos esperan niveles de expresión homogéneos en cada muestra frente a la mezcla de comportamientos que se puede observar en este tipo de marcadores en la muestra de casos. Por eso, en la última década, se han desarrollado algunos métodos alternativos, muchos de ellos, basados en la idea de robustificar el estadístico t-Student. Los más utilizados son el estadístico COPA (*Cancer Outlier Profile Analysis*) **[Tomlins et al, 2005]**, el OS (*Outilier Sum*) **[Tibshirani y Hastie, 2007]**, el ORT (*Outlier Robust T-statistic*) **[Wu, 2007]** y el MOST (*Maximum Ordered Subset T-statistics*) **[Lian, 2008]**. Todos ellos, tratan de identificar genes que aparecen sobre o infra-expresados solamente en un pequeño porcentaje del grupo de tumores. En esta memoria nos referimos a este tipo de genes utilizando la etiqueta OHE (*Outlier High Expression*). En las simulaciones que aparecen en **[Lian, 2008]**; **[Tibshirani y Hastie, 2007]** y **[Wu, 2007]**, estos métodos muestran un funcionamiento superior al observado

en el estadístico t-Student cuando el porcentaje de muestras sobre-expresadas es inferior al 50%. Existen otras aproximaciones basadas en cuantiles, como el método PPST (*Permutation Percentile Separability Test*) **[Lyoins-Weiker et al, 2004]**, que trata de identificar genes caracterizados por un sub-conjunto grande de muestras caso con valores de expresión por encima de umbrales determinados a partir de las muestras control, o a partir de todas las muestras, como el *score* GTI (*Gene Tissue Index*) propuesto en **[Mpindi et al, 2011]**.

Frente a los genes OHE, recientemente, aparece la idea de genes con expresión predominantemente alta, a los que nos referiremos utilizando la denominación PHE (*Predominantly High Expression*), caracterizados por un número alto de muestras sobre-expresadas en el grupo de casos, aunque no necesariamente todas. Asumiendo que los marcadores de enfermedad van a estar altamente expresados en al menos el 80% de la muestras de casos, **[Gleiss et al, 2011]** proponen el *Adaptive Trimmed t-statistics*. Mientras que en los métodos anteriormente mencionados, los estimadores de localización y escala son sustituidos por medianas o funciones de cuantiles, este último estadístico se corresponde con el estadístico t-Student teniendo en cuenta sólo observaciones no recortadas por reglas basadas en la mediana y en cuantiles.

Toda esta familia de métodos, que tratan de resolver el problema de los patrones de activación heterogéneos propios de los oncogenes, responden, desde el punto de vista estadístico, a dos problemas distintos. Por un lado, la existencia de genes en los que la expresión diferencial se observa en un sub-conjunto, relativamente pequeño, de muestras de tumores. Este se puede analizar como un problema de una única población, en el que interesa comparar los valores observados en la población de casos con la distribución de expresión observada en esta población. El otro problema, que correspondería a los genes con expresión predominantemente alta, se acerca más al problema clásico de contrastar igualdad de localización en dos poblaciones, donde se debe permitir la posibilidad de que alguna de ellas, o las dos, aparezcan contaminadas. De esta forma, se recoge que un porcentaje pequeño de la muestra de casos puede responder como los controles, o incluso que en las dos muestras pueden existir observaciones mal clasificadas o con respuesta atípica.

En ambos casos, la solución al problema está relacionada con encontrar el conjunto de muestras representativo de una o dos poblaciones, es decir su núcleo de expresión. En el primer caso, esta solución se corresponde con la identificación del núcleo común a las dos muestras o al de la muestra de controles, para a continuación evaluar desviaciones de este patrón en la muestra de casos. En el segundo, la solución vendrá dada por contrastar la igualdad de los dos núcleos centrales. Como es esperable que los dos tipos de genes convivan en el mismo conjunto de datos, nuestra recomendación general es aplicar las dos metodologías conjuntamente para detectar el listado de genes diferencialmente expresados en uno u otro sentido.

En este capítulo, proponemos el uso de procedimientos de recorte imparcial para resolver cada uno de estos problemas, en concreto, la utilización del estimador smart presentado en

el capítulo 3 para la identificación de los núcleos de expresión. Los procedimientos basados en el recorte imparcial presentan, en el plano conceptual, una ventaja clara sobre los procedimientos señalados. Todos estos procedimientos, identifican atipicidad como desviación a la posición de la mediana, que ocupa la posición central en el conjunto de observaciones, pero no necesariamente entre las observaciones genuinas, sobre todo cuando el tipo de contaminación que se espera, por la naturaleza del problema, es unilateral. En su aplicación a estas situaciones, los procedimientos de recorte imparcial permiten que nos aproximemos a identificar comportamientos atípicos como desviaciones a la localización del centro de las observaciones genuinas. La correcta identificación de las observaciones atípicas, nos ha permitido evaluar hasta que punto una determinada variable clínica puede ser la responsable de la presencia de observaciones sobre o infra-expresadas. Esto nos lleva a pensar que podría resultar muy útil en otras aplicaciones, fuera del contexto de la expresión diferencial, como por ejemplo en la identificación del *batch effect* **[Johnson y Li, 2007]** o en la evaluación de la homogeneidad de varianzas entre clases, cuando la heterogoneidad se debe a unas pocas observaciones extremas.

Empezamos este capítulo detallando las propuestas, ya mencionadas, disponibles en la literatura, para a continuación detallar nuestra propuesta.

5.1. Métodos disponibles

El estadístico t-Student es el método clásico para contrastar expresión diferencial en cada gen en dos muestras independientes. Cuando solamente un sub-conjunto de las muestras pertenecientes a uno de los grupos muestra expresión diferencial, o equivalentemente cuando al menos una de las muestras aparece contaminada, la media y la varianza muestral van a mostrar un mal funcionamiento por la ausencia de propiedades de robustez. Los procedimientos que se detallan a continuación, fueron construidos para intentar solucionar este problema mediante la robustificación el estadístico t-Student. En el caso de los estadísticos COPA **[Tomlins et al, 2005]**, OS **[Tibshirani y Hastie, 2007]**, ORT **[Wu, 2007]** y MOST **[Lian, 2008]**, se compara la localización en el grupo de controles, representada por la mediana, con los niveles de expresión observados entre los casos utilizando cuantiles o funciones de estos. Cada uno de estos estadísticos tendrá dos versiones, una para el caso de muestras sobre-expresadas, y otra, totalmente análoga, para muestras infra-expresadas. El *Adaptive Trimmed t-statistics* **[Gleiss et al, 2011]** sustituye las medias y las desviaciones típicas por sus correspondientes versiones recortadas, definido el recorte a partir de cuantiles muestrales. Los p-valores se obtienen mediante la estimación de la distribución de los estadísticos en una muestra de permutaciones.

En todos los casos, los datos de partida vienen dados por una matriz de expresión para el grupo de controles x_{ij} con n_G genes y n_1 arrays, y otra independiente, y_{ij}, correspondiente al grupo de los casos con los mismos genes y n_2 arrays.

5.1.1. COPA: *Cancer Outlier Profile Analysis*

Para el gen g, el estadístico COPA **[MacDonald y Ghosh, 2006]**; **[Tomlins et al, 2005]** se calcula como,

$$copa_g^+ = \frac{q_r\left(\{y_{gj}\}_{1\le j\le n_2}\right) - med_g}{mad_g} \tag{5.1}$$

donde, med_g y mad_g son, respectivamente la mediana y el MAD de todas las observaciones, casos y controles, y $q_r(A)$ es el r-ésimo percentil de la muestra global, habitualmente $r = 75,\ 90 \text{ o } 95$.

Como definición del MAD se utiliza la mediana de las desviaciones respecto de la mediana, en valor absoluto, multiplicada por la constante 1.4826, para conseguir que sea un estimador insesgado en el modelo normal para la desviación típica.

De manera similar, se puede definir el estadístico COPA para los genes infra-expresados utilizando los percentiles $r = 25, 10 \text{ o } 5$, como

$$copa_g^- = \frac{med_g - q_r\left(\{y_{gj}\}_{1\le j\le n_2}\right)}{mad_g} \tag{5.2}$$

5.1.2. OS: *Outlier Sums*

Una estrategia más eficiente para detectar genes OHE que la utilizada por el estadístico COPA, corresponde a basar el estadístico, no en un cuantil, sino en todos los valores de expresión de la muestra de casos etiquetados como atípicos. Para la definición de atípico utilizan reglas inspiradas por el diagrama de cajas (*boxplot rules*).

Con esta idea, para un gen g, el estadístico OS **[Tibshirani y Hastie, 2007]** para contrastar sobre-expresión, se calcula como la suma de los valores de expresión considerados como atípicos en la muestra de casos, estandarizados éstos respecto de la mediana y el MAD de todas las observaciones en la muestra. Se definen como atípicos aquellos casos cuya distancia al percentil 75 sea superior al rango intercuartílico.

De forma análoga, sustituyendo el percentil 75 por el 25 se tiene la definición correspondiente para el caso de infra-expresión.

5.1.3. ORT: *Outlier Robust t-statistic*

El estadístico COPA utiliza como estimador de la localización de los niveles de expresión del grupo control, la mediana de todas las muestras. Incorporando las muestras del grupo de

casos, buscan aumentar la eficiencia, pero si este grupo presenta un porcentaje importante de muestras sobre-expresadas, la mediana de la muestra global podría sobre-estimar el verdadero centro del grupo control. El estadístico ORT **[Wu, 2007]** trata de solucionar este problema estandarizando las observaciones atípicas del grupo de casos utilizando la mediana del grupo control, med_{gX} , y la mediana de todas las desviaciones de las observaciones de la muestra respecto de la mediana de su grupo, mad_{gC} . Para cada gen g se calcula, por tanto, como la suma de los valores de expresión considerados como atípicos en la muestra de casos, estandarizados respecto de med_{gX} y mad_{gC} , y con la misma definición de atípicidad que la utilizada en el caso del estadístico OS.

5.1.4. MOST: *Maximum Ordered Subset t-statistic*

En todos los estadísticos anteriores, el nivel a partir del cual se considera una observación como atípica en el grupo de casos se define de una forma arbitraria. El método MOST **[Lian, 2008]**, trata de solucionar este problema intentando que el propio método escoja ese umbral a partir de la información muestral.

Sea $y_{g(1)} \geq y_{g(2)} \geq \ldots \geq y_{g(n_2)}$ el estadístico ordenado de la muestra de casos $\left\{y_{gj}\right\}_{1\leq j\leq n_2}$, y k el número de muestras de este grupo que presentan sobre-expresión, el estadístico $most_g^+$ se define como,

$$most_g^+ = \max_{1\leq k\leq n_2}\left(\frac{\dfrac{\sum_{j=1}^{k}\left(y_{g(j)} - med_{gX}\right)}{mad_{gC}} - \mu_k}{\sigma_k}\right) \tag{5.3}$$

donde $\mu_k = E\left[\sum_{j=1}^{k} z_{(j)}\right]$ y $\sigma_k^2 = Var\left[\sum_{j=1}^{k} z_{(j)}\right]$ con $z_{(1)} \geq z_{(2)} \geq \ldots \geq z_{(n_2)}$ el estadístico ordenado de n_2 muestras generadas de una $N(0,1)$. De forma análoga se puede definir la correspondiente versión para los genes infra-expresados.

5.1.5. *Adaptative Trimmed t-Statistics*

[Gleiss et al, 2011] intentan robustificar el estadístico t-Student para dos muestras independientes, en su aplicación a la identificación de genes PHE, utilizando medias y varianzas recortadas como estimadores de localización y escala respectivamente. Ellos

proponen utilizar un recorte basado en la regla *boxplot*, considerando tres umbrales: $f = 0.5,\ 1$ y 1.5. Una alternativa a esta propuesta para obtener el recorte, corresponde a la búsqueda del nivel de recorte unilateral óptimo, entre una lista de niveles de recortes posibles $\Gamma = \{0,\ 0.1,\ 0.2\}$, en el sentido de que proporcione el menor p-valor del estadístico.

5.2. Método propuesto

Proponemos dos estadísticos para evaluar expresión diferencial en dos clases independientes cuando existen patrones de activación heterogénea en al menos una de ellas. En la utilización del primero asumimos que la expresión diferencial está relacionada con un porcentaje reducido de muestras en una de ellas (búsqueda de genes OHE). En el caso del segundo, suponemos que en las dos clases hay un núcleo de comportamiento típico, pero que puede diferir de una a otra (búsqueda de genes OHE). En los dos casos, los estadísticos están basados en el estimador smart.

5.2.1. Estadístico para detectar genes OHE

Para cada gen g, $g = 1,\ldots,n_G$, del conjunto analizado, el procedimiento propuesto para evaluar si dicho gen presenta un patrón del tipo OHE parte de la construcción del núcleo de expresión común a los dos grupos, casos y controles, fijando un nivel de recorte $\alpha \in [0,1]$ y utilizando las ideas presentadas en el capítulo 3. Consideramos atípicas aquellas observaciones del grupo de casos, que han sido recortadas, bien por arriba, en el caso de que se quiera estudiar sobre-expresión, o por abajo, cuando se evalúa infra-expresión. El estadístico se calcula como la suma de estas observaciones atípicas, estandarizadas por las estimaciones de localización y dispersión correspondientes al smart.

Por tanto, el estadístico que se propone para evaluar si el gen g presenta un patrón OHE se define según la expresión,

$$W^{+}_{g,\alpha} = \sum_{j=1}^{n_2} \frac{y_{gj} - \hat{\mu}_{g,\alpha-smart}}{\hat{\sigma}_{g,\alpha-smart}} I_{\overline{A}^{+}_{g,\alpha}}\left(y_{gj}\right) \tag{5.4}$$

donde $\overline{A}^{+}_{g,\alpha}$ es el conjunto de observaciones recortadas por arriba en el gen g, y $\hat{\mu}_{g,\alpha-smart}$ y $\hat{\sigma}^2_{g,\alpha-smart}$ los estimadores de localización y dispersión del núcleo de expresión del gen g en la muestra global. Análogamente se puede definir la versión correspondiente a los genes infra-expresados.

Para tomar una decisión acerca de si un gen está diferencialmente expresado en el sentido OHE, calculamos un p-valor ajustado siguiendo **[Dudoit et al, 2002b]** y estimamos la FDR utilizando la metodología SAM **[Tusher et al, 2001]** descrita en el capítulo 4.

5.2.2. Estadístico para detectar genes PHE

En este caso, para cada gen g, $g = 1,...,n_G$, del conjunto analizado, se parte de la construcción de un núcleo de expresión en el grupo de controles y otro núcleo en el de casos, fijando los niveles de recorte $\alpha_1 \in [0,1]$ y $\alpha_2 \in [0,1]$, respectivamente.

El estadístico que se propone para evaluar si el gen g presenta un patrón PHE se define según la expresión,

$$t_{g,\alpha_1,\alpha_2} = \frac{\hat{\mu}_{g,\alpha_2-smart} - \hat{\mu}_{g,\alpha_1-smart}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot \sqrt{\frac{n_1\hat{\sigma}^2_{g,\alpha_1-smart} + n_2\hat{\sigma}^2_{g,\alpha_2-smart}}{n_1+n_2-2}}} \qquad (5.5)$$

donde $\hat{\mu}_{g,\alpha_i-smart}$ y $\hat{\sigma}^2_{g,\alpha_i-smart}$ los estimadores de localización y dispersión del núcleo de expresión del gen g en la clase i.

El contraste correspondiente lo llevamos a cabo de forma análoga a la de detectar genes OHE.

5.3. Evaluación del método propuesto

5.3.1. Datos simulados

Generamos matrices de expresión génica por método de Monte Carlo con $n_G\,(=100)$ genes divididos en $K(=2)$ clases del mismo tamaño, $n_j\,(=50)$, $j = 1,2$. Todos los valores de expresión se obtienen a partir de distribuciones normales, y sólo el primero de los 100 genes presentará expresión diferencial de acuerdo a diferentes modelos de contaminación. En todos los casos y, para tener en cuenta la variabilidad de los distintos métodos, se simulan $n_{sim}\,(=50)$ conjuntos de datos de cada uno de los modelos.

5.3.1.1. Simulaciones para genes OHE

El primer gen tendrá expresión diferencial en el 20% de las muestras de tumores correspondiente a una magnitud $\delta = 3$. Estas muestras DE se generan con una desviación típica $\sigma_{DE} = 0.2$, mientras que las observaciones sin expresión diferencial se generan con $\sigma_{EE} = 1$. Además, para el gen DE, y salvo en el **Modelo 0**, sin expresión diferencial, se

añade una proporción $1-\omega=0.2$ de muestras contaminadas en el grupo de controles, de la forma que se describe a continuación.

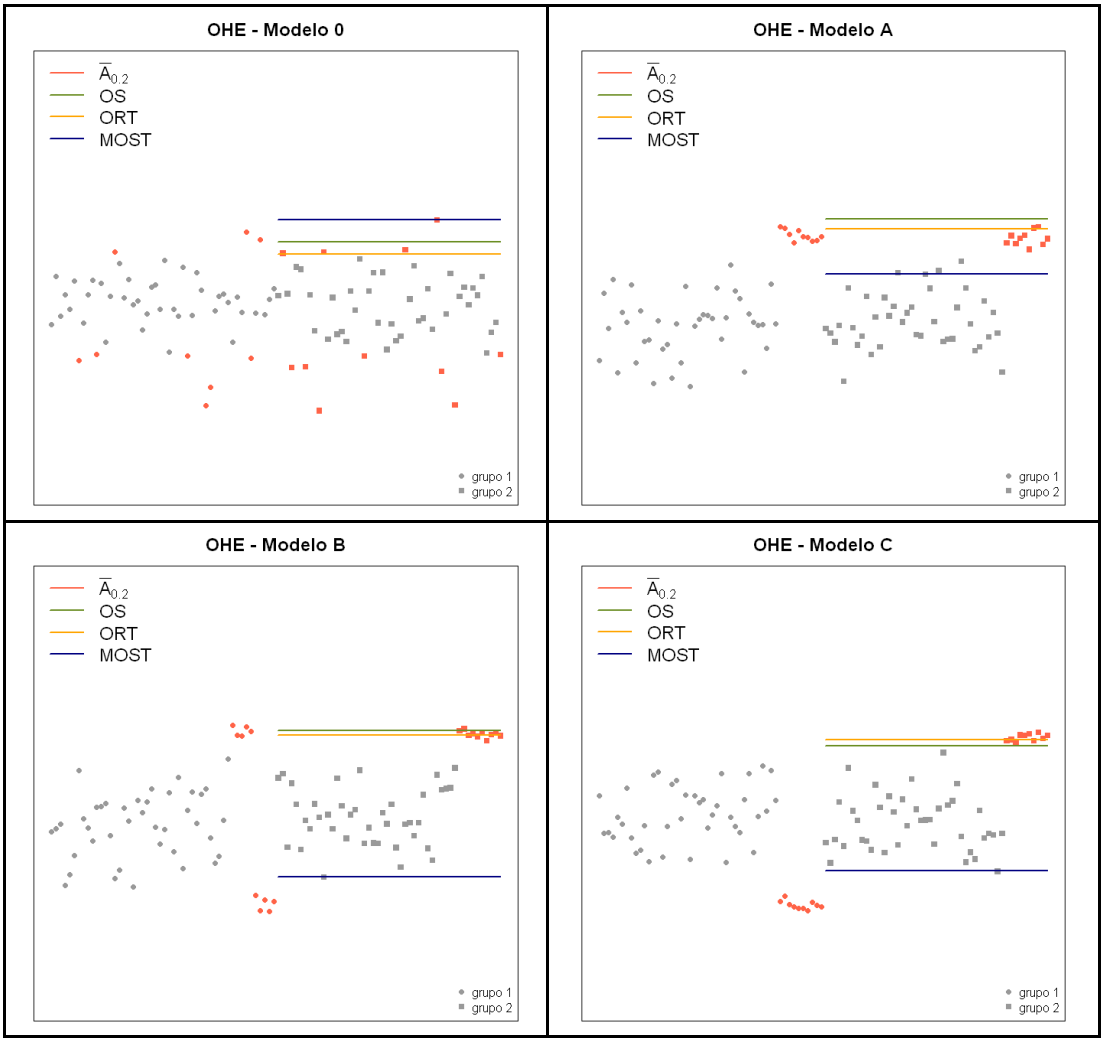


Figura 5.1. Modelos de contaminación para el problema de genes OHE con parámetros $\omega=0.8$ y $\delta=3$. En rojo aparecen marcados los puntos eliminados a partir del recorte imparcial con $\alpha=0.20$. Las líneas delimitan el intervalo de observaciones no contaminadas según los umbrales propuestos por los métodos OS, ORT y MOST.

- **Modelo A**. Las observaciones contaminadas del grupo control se generan a partir de la misma distribución que el grupo de observaciones DE en la muestra de casos: $\mu_C=3$ y $\sigma_C=0.2$.
- **Modelo B.** La mitad de las observaciones contaminadas en el grupo control, se generan a partir de la misma distribución que el grupo de observaciones DE en la muestra de casos: $\mu_C=3$ y $\sigma_C=0.2$. La otra mitad, se coloca infra-expresándose con $\mu_C=-3$ y la misma desviación típica, $\sigma_C=0.2$.

- **Modelo C.** Las observaciones contaminadas del grupo control aparecen infra-expresándose con $\mu_C = -3$ y $\sigma_C = 0.2$.

La figura 5.1 ilustra un ejemplo de un gen simulado con patrón OHE, a partir de los distintos modelos de contaminación propuestos. Además, se marcan las observaciones que se considerarían atípicas según las definiciones de os_g^+, ort_g^+, $most_g^+$ y $W_{g,\alpha}^+$, fijando $\alpha = 0.2$.

En la tabla 5.1 se resumen los resultados obtenidos a partir de cinco medidas: **(i)** Obs_1: valor del estadístico en el gen OHE; **(ii)** Obs_0: valor del estadístico en los genes sin expresión diferencial; **(iii)** O_1: posición que ocupa el gen OHE entre los 100 genes simulados; **(iv)** P_1: porcentaje de veces, de entre las 50 simulaciones, que el gen OHE aparece situado en primer lugar en la muestra ordenada de estadísticos; y **(v)** P_5: porcentaje de veces, de entre las 50 simulaciones, que el gen OHE aparece colocado entre los 5 primeros puestos de la muestra ordenada de estadísticos.

El modelo A, caracterizado por tener en el grupo control un conjunto de muestras expresadas en la misma dirección que en el grupo de casos, es en el que se observa peor funcionamiento para el resto de estadísticos. Este tipo de contaminación, produce que los umbrales utilizados en estas aproximaciones, basados en la regla *boxplot*, incluyan como no atípicos niveles de expresión claramente separados del *verdadero* centro de los datos. Además, las desviaciones respecto de la mediana, bien de todas las muestras o sólo de las muestras control, van a producir una sobre-estimación de la varianza. En el modelo B, el hecho de que la contaminación en la muestra de controles esté sobre e infra-expresada va a dar cierta ventaja al estadístico ORT, que sólo utiliza este grupo para estandarizar las observaciones clasificadas como *outliers*. El tipo de contaminación simulada en el modelo C no genera problemas en ninguno de los métodos. En todos los casos, se observa que $W_{g,0.2}^+$ es superior, salvo respecto a os_g^+, en el modelo C, que al utilizar todas las muestras del grupo control, se ve favorecido por la baja expresión de las observaciones contaminadas en ese grupo en esa dirección. Aún así, incluso en este caso, $W_{g,0.2}^+$ y os_g^+ obtienen resultados similares.

En todos los modelos de contaminación propuestos, $W_{g,0.2}^+$ funciona bien, independientemente del tipo de contaminación que se añada en los controles.

Tabla 5.1. Búsqueda del gen OHE con $\delta = 3$ entre los 99 genes sin expresión diferencial. Métodos COPA, OS, ORT, MOST y W, basado en el recorte imparcial con $\alpha = 0.2$, en cada uno de los modelos de contaminación de la figura 5.1.

Modelo	Medida	Genes OHE				
		$copa^+$	os^+	ort^+	$most^+$	W^+ $\alpha = 0.20$
0	Obs_1 (media ± DT)	1.87 ± 0.38	3.77 ± 4.25	5.79 ± 6.04	0.74 ± 1.29	9.89 ± 4.74
	Obs_0 (media ± DT)	1.86 ± 0.37	4.48 ± 4.43	5.58 ± 5.96	0.98 ± 1.40	9.78 ± 4.48
	O_1 (media ± DT)	52.20 ± 29.92	62.30 ± 27.99	50.70 ± 29.37	55.48 ± 28.98	56.78 ± 29.69
	P_1 (%)	2	0	2	0	2
	P_5 (%)	6	0	4	2	6
A	Obs_1 (media ± DT)	2.94 ± 0.43	6.06 ± 9.48	10.02 ± 12.60	3.12 ± 1.42	27.96 ± 4.04
	Obs_0 (media ± DT)	1.85 ± 0.38	4.40 ± 4.41	5.38 ± 5.76	0.97 ± 1.41	9.59 ± 4.55
	O_1 (media ± DT)	25.70 ± 30.34	42.70 ± 30.49	48.18 ± 32.46	40.42 ± 31.53	11.78 ± 27.00
	P_1 (%)	46	12	14	14	82
	P_5 (%)	46	12	16	18	84
B	Obs_1 (media ± DT)	3.20 ± 0.46	19.43 ± 15.16	30.05 ± 13.75	4.64 ± 1.62	28.15 ± 4.55
	Obs_0 (media ± DT)	1.85 ± 0.37	4.34 ± 4.40	5.33 ± 5.83	0.95 ± 1.41	9.73 ± 4.51
	O_1 (media ± DT)	15.42 ± 25.19	22.10 ± 28.27	13.06 ± 25.10	28.12 ± 32.75	8.72 ± 19.66
	P_1 (%)	64	52	70	44	82
	P_5 (%)	70	54	72	46	82
C	Obs_1 (media ± DT)	3.45 ± 0.43	31.61 ± 10.86	34.87 ± 12.30	5.62 ± 1.36	28.34 ± 4.97
	Obs_0 (media ± DT)	1.85 ± 0.37	4.41 ± 4.46	5.44 ± 5.81	0.99 ± 1.41	9.60 ± 4.48
	O_1 (media ± DT)	13.76 ± 27.54	8.08 ± 19.47	12.16 ± 24.62	11.94 ± 22.77	9.56 ± 22.02
	P_1 (%)	78	84	76	70	82
	P_5 (%)	78	86	76	76	82

5.3.1.2. Simulaciones para genes PHE

El primer gen tendrá expresión diferencial en una cantidad δ y en una proporción mayoritaria de muestras, el 75% en cada grupo.

Se consideran 3 modelos de contaminación, en los que, las observaciones no contaminadas se generan con media $\mu_{1,1} = 0$ para el grupo de controles, y $\mu_{2,1} = \delta$ en los casos, y la misma desviación típica. El resto de observaciones constituyen el grupo de muestras contaminadas y se generan de la siguiente forma:

- **Modelo D**. En cada grupo, una parte pequeña de observaciones son extremas respecto de los dos grupos: en el grupo 1, $\mu_{1,2} = 2\delta$ y en el grupo 2, $\mu_{2,2} = -\delta$. El resto, se

generan con la media del otro grupo, $\mu_{1,3} = \delta$ y $\mu_{2,3} = 0$, respectivamente. Todas las observaciones se generan con la misma desviación típica.

- **Modelo E**. Las observaciones contaminadas se generan, para el grupo control, alrededor del percentil 80 de la distribución de muestras de casos no contaminadas, denotado por $p_{0.80}^{(2)}$, y para el grupo de casos, alrededor del percentil 20 de la distribución de muestras de controles no contaminados, $p_{0.20}^{(1)}$. La desviación típica en estos subgrupos de observaciones es pequeña en relación a la desviación típica de las observaciones no contaminadas.
- **Modelo F.** La contaminación sólo aparece en el grupo de casos, generada alrededor de $-\delta$ con una desviación típica menor que la desviación típica de las observaciones no contaminadas.

Como en el caso anterior, también se considera el **Modelo 0**, en el que no se genera ningún tipo de contaminación.

Se evalúan en dos niveles de expresión diferencial $\delta = \{1, 2\}$, el resultado de los contrastes t-Student, t-Student recortado a partir de la regla *boxplot* considerando tres umbrales, $f = \{0.5, 1, 1.5\}$, y el t-Student recortado utilizando el recorte imparcial con niveles α_1 y α_2 para cada uno de los grupos. La figura 5.2 ilustra un ejemplo de un gen simulado con patrón PHE, a partir de los distintos modelos de contaminación propuestos. Además, se marcan las observaciones que se considerarían atípicas según las definiciones de los dos estadísticos recortados.

En las tabla 5.2 y 5.3 se resumen los resultados variando la cantidad de expresión diferencial, $\delta = \{1, 2\}$. Se utilizan cinco medidas análogas a las utilizadas en la sección anterior: **(i)** t_1: valor del estadístico en el gen PHE; **(ii)** t_0: valor del estadístico en los genes sin expresión diferencial; **(iii)** O_1: posición que ocupa el gen PHE entre los 100 genes simulados; **(iv)** P_1: porcentaje de veces, de entre las 50 simulaciones, que el gen PHE aparece colocado en primer lugar en la muestra ordenada de estadísticos; y **(v)** P_5: porcentaje de veces, de entre las 50 simulaciones, que el gen PHE aparece colocado entre los 5 primeros puestos de la muestra ordenada de estadísticos.

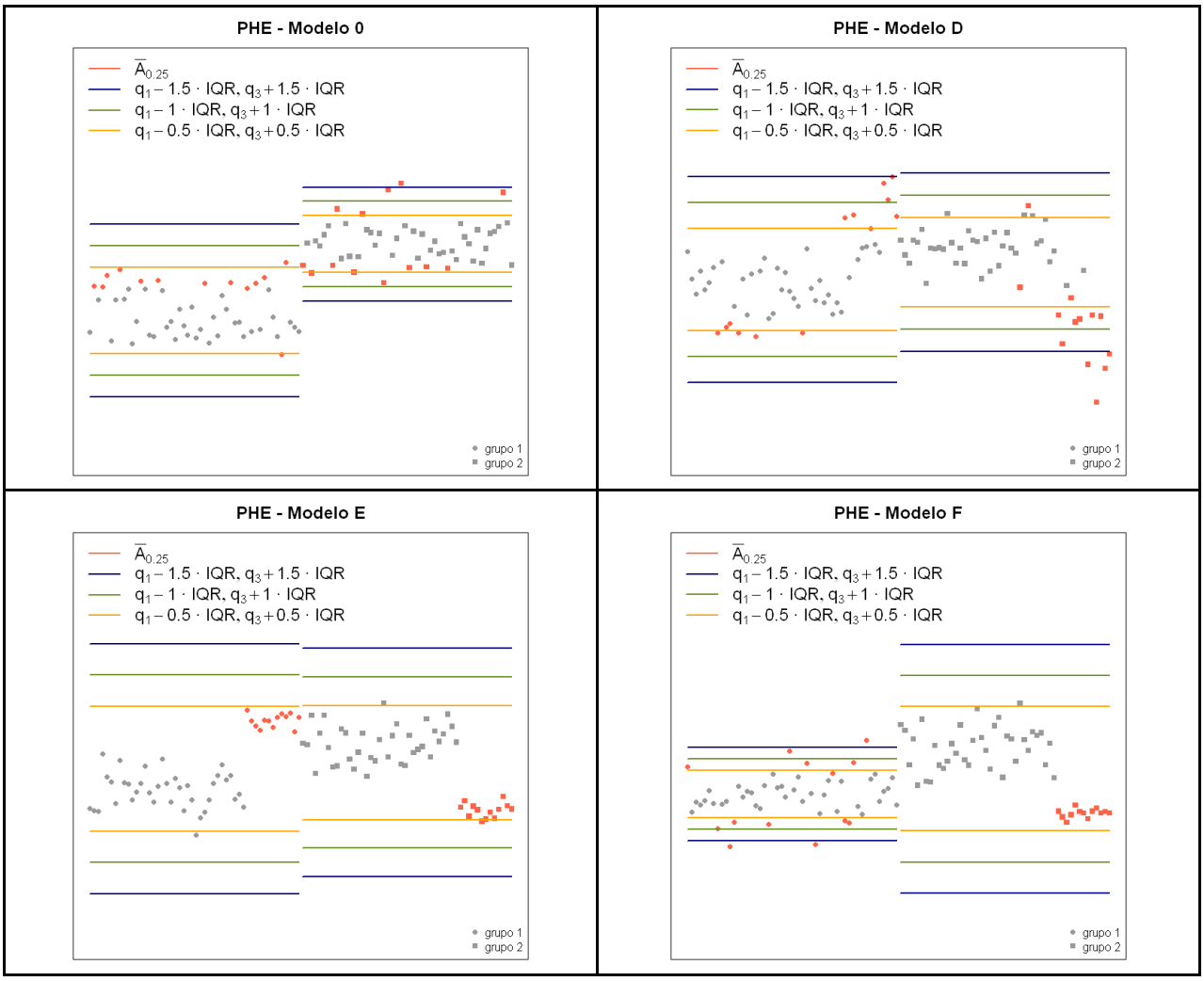


Figura 5.2. Modelos de contaminación para el problema de genes PHE con el 75% de las muestras diferencialmente expresadas en una cantidad $\delta = 2$. En rojo aparecen marcados los puntos eliminados a partir del recorte imparcial con $\alpha_1 = \alpha_2 = 0.25$. Las líneas delimitan el intervalo de observaciones no contaminadas según la regla *boxplot*, considerando $f = \{0.5, 1, 1.5\}$.

El modelo de contaminación E, es en el que funcionan peor los estadísticos obtenidos a partir de los recortes basados en la regla *boxplot*, al menos cuando las diferencias entre grupos son grandes, $\delta = 2$. Incluso, utilizando el estadístico t-Student clásico, se obtienen mejores resultados que con este tipo de recorte. Con $\delta = 2$, la contaminación del modelo F apenas tiene efecto, consiguiendo buenos resultados a partir de cualquiera de los *scores* evaluados. Sin embargo, con distancias más pequeñas, $\delta = 1$, es precisamente este modelo el que muestra mayores diferencias según el tipo de recorte, debido a que utilizando el recorte imparcial se consigue aumentar la potencia de los contrastes. En contrapartida, el modelo sin

contaminación muestra peores resultados con este tipo de recorte, debido a la pérdida de eficiencia que supone trabajar con un conjunto reducido cuando no existe contaminación.
Hay un caso en el que todos los estadísticos utilizados muestran un mal funcionamiento, el modelo D con $\delta = 1$. Las observaciones contaminantes son difíciles de detectar independientemente del tipo de recorte utilizado, ya que, la mayoría de ellas, se generan con una media que dista menos de una desviación típica de la media de las observaciones consideradas buenas, y sólo un conjunto muy pequeño, formado por cuatro observaciones en cada grupo, se genera con una media que dista dos desviaciones típicas.

Tabla 5.2. Búsqueda del gen PHE con $\delta = 1$ mezclado con 99 genes sin expresión diferencial. Se comparan los métodos basados en el estadístico t recortado en cada uno de los modelos de contaminación considerados.

Modelo	Medida	t-Statistic	Adaptative trimmed t-Statistics			Nuestra propuesta $\alpha_1 = \alpha_2 = 0.25$
			Boxplot $f = 1.5$	Boxplot $f = 1$	Boxplot $f = 0.5$	
0	t_1 (media ± DT)	6.77 ± 1.21	7.10 ± 1.40	7.63 ± 1.70	9.12 ± 2.31	7.28 ± 2.24
	t_0 (media ± DT)	0.00 ± 0.99	-0.01 ± 1.09	-0.01 ± 1.28	-0.02 ± 1.65	-0.01 ± 1.66
	O_1 (media ± DT)	1.00 ± 0.00	1.00 ± 0.00	2.10 ± 7.78	1.00 ± 0.00	7.32 ± 20.08
	P_1 (%)	100	100	98	100	90
	P_5 (%)	100	100	98	100	90
D	t_1 (media ± DT)	1.09 ± 0.65	1.58 ± 0.81	2.25 ± 1.17	2.95 ± 1.48	3.22 ± 1.63
	t_0 (media ± DT)	-0.03 ± 1.09	-0.02 ± 1.27	-0.03 ± 1.00	-0.03 ± 1.65	-0.04 ± 1.66
	O_1 (media ± DT)	54.80 ± 29.99	43.76 ± 32.32	38.42 ± 33.64	36.88 ± 30.51	33.76 ± 31.39
	P_1 (%)	2	12	24	20	28
	P_5 (%)	2	12	26	22	30
E	t_1 (media ± DT)	3.11 ± 0.55	3.11 ± 0.55	3.17 ± 0.61	3.38 ± 0.58	6.49 ± 1.53
	t_0 (media ± DT)	0.01 ± 1.00	0.02 ± 1.10	0.02 ± 1.28	0.01 ± 1.63	0.00 ± 1.66
	O_1 (media ± DT)	9.00 ± 20.86	15.42 ± 27.70	21.36 ± 31.99	37.68 ± 29.71	10.90 ± 27.15
	P_1 (%)	82	70	56	22	86
	P_5 (%)	82	72	58	22	86
F	t_1 (media ± DT)	2.71 ± 0.77	2.74 ± 0.76	2.71 ± 0.83	2.66 ± 0.94	5.42 ± 1.59
	t_0 (media ± DT)	0.00 ± 1.01	-0.01 ± 1.11	-0.03 ± 1.29	-0.02 ± 1.65	-0.03 ± 1.65
	O_1 (media ± DT)	23.96 ± 33.34	28.54 ± 35.09	35.98 ± 33.20	48.14 ± 33.68	12.40 ± 23.62
	P_1 (%)	58	46	30	10	74
	P_5 (%)	60	54	34	14	74

Tabla 5.3. Posición del gen DE con $\delta = 2$ entre los 100 genes generados en cada uno de los modelos de contaminación considerados.

Modelo	Medida	t-Statistic	Adaptative trimmed t-Statistics			Nuestra propuesta $\alpha_1 = \alpha_2 = 0.25$
			Boxplot $f = 1.5$	Boxplot $f = 1$	Boxplot $f = 0.5$	
0	t_1 (media ± DT)	13.62 ± 1.20	14.11 ± 1.41	15.28 ± 1.69	18.38 ± 2.46	14.37 ± 2.42
	t_0 (media ± DT)	0.00 ± 1.01	0.01 ± 1.11	0.00 ± 1.28	0.03 ± 1.67	0.01 ± 1.69
	O_1 (media ± DT)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	P_1 (%)	100	100	100	100	100
	P_5 (%)	100	100	100	100	100
D	t_1 (media ± DT)	1.87 ± 0.47	3.25 ± 0.90	5.06 ± 1.10	7.66 ± 1.74	9.70 ± 2.17
	t_0 (media ± DT)	0.01 ± 1.00	0.01 ± 1.10	0.01 ± 1.29	0.03 ± 1.67	0.00 ± 1.69
	O_1 (media ± DT)	45.02 ± 35.90	18.22 ± 30.93	4.30 ± 16.11	2.76 ± 12.45	1.00 ± 0.00
	P_1 (%)	18	70	94	98	100
	P_5 (%)	24	72	96	98	100
E	t_1 (media ± DT)	1.90 ± 0.41	1.90 ± 0.41	1.90 ± 0.41	1.71 ± 0.54	10.83 ± 1.47
	t_0 (media ± DT)	-0.01 ± 1.00	-0.01 ± 1.09	-0.01 ± 1.28	0.00 ± 1.65	0.00 ± 1.67
	O_1 (media ± DT)	43.10 ± 30.96	44.96 ± 27.55	53.88 ± 27.22	51.48 ± 25.6	1.00 ± 0.00
	P_1 (%)	14	8	2	0	100
	P_5 (%)	20	10	4	8	100
F	t_1 (media ± DT)	5.74 ± 0.64	5.74 ± 0.64	5.74 ± 0.65	5.62 ± 0.74	12.32 ± 1.93
	t_0 (media ± DT)	0.00 ± 1.02	-0.01 ± 1.10	-0.01 ± 1.29	-0.02 ± 1.67	-0.03 ± 1.70
	O_1 (media ± DT)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	3.42 ± 13.02	1.00 ± 0.00
	P_1 (%)	100	100	100	94	100
	P_5 (%)	100	100	100	94	100

5.3.2. Aplicación al *dataset* de Cáncer de Pulmón

Analizamos el *dataset* de Cáncer de Pulmón, descrito en la sección B.2 del apéndice B. El conjunto cuenta con 20172 genes en un total de 91 muestras, 45 de ellas procedentes de tejido sano y 46 de células tumorales divididas en dos sub-grupos: 14 adenocarcinomas y 32 squamous-cell carcinomas. Asumiendo que se desconoce el tipo de cáncer, esta mezcla de dos poblaciones en la muestra de casos, sirve como un ejemplo de patrones de activación heterogéneos.

Se consideran un nivel de recorte de $\alpha = 0.35$ para la búsqueda de genes OHE a partir del estadístico definido en la expresión (5.4). Para los genes PHE se consideran los niveles $\alpha_1 = 0.1$ y $\alpha_2 = 0.3$ en el estadístico t-Student recortado definido en (5.5). Estos niveles de recorte se eligen atendiendo a la proporción de cada tipo de cáncer en el grupo de casos.

En la tabla 5.4 se resumen los resultados obtenidos a partir de los métodos propuestos. En total se declaran significativos, a nivel 0.01, 9684 genes, aproximadamente el 48% de los genes evaluados.

Tabla 5.4. Genes significativos, p-valor≤0.01, entre los 20172 genes estudiados

Contraste	# genes	% genes	% genes únicos	Concordancia entre contrastes (%)		
				$W^{+}_{0.35}$	$W^{-}_{0.35}$	$t_{0.10,0.30}$
$W^{+}_{0.35}$	4171	20.68	43.87	100	0.22	55.93
$W^{-}_{0.35}$	1624	8.05	19.95	0.55	100	79.56
$t_{0.10,0.30}$	7522	37.29	51.82	31.02	17.18	100

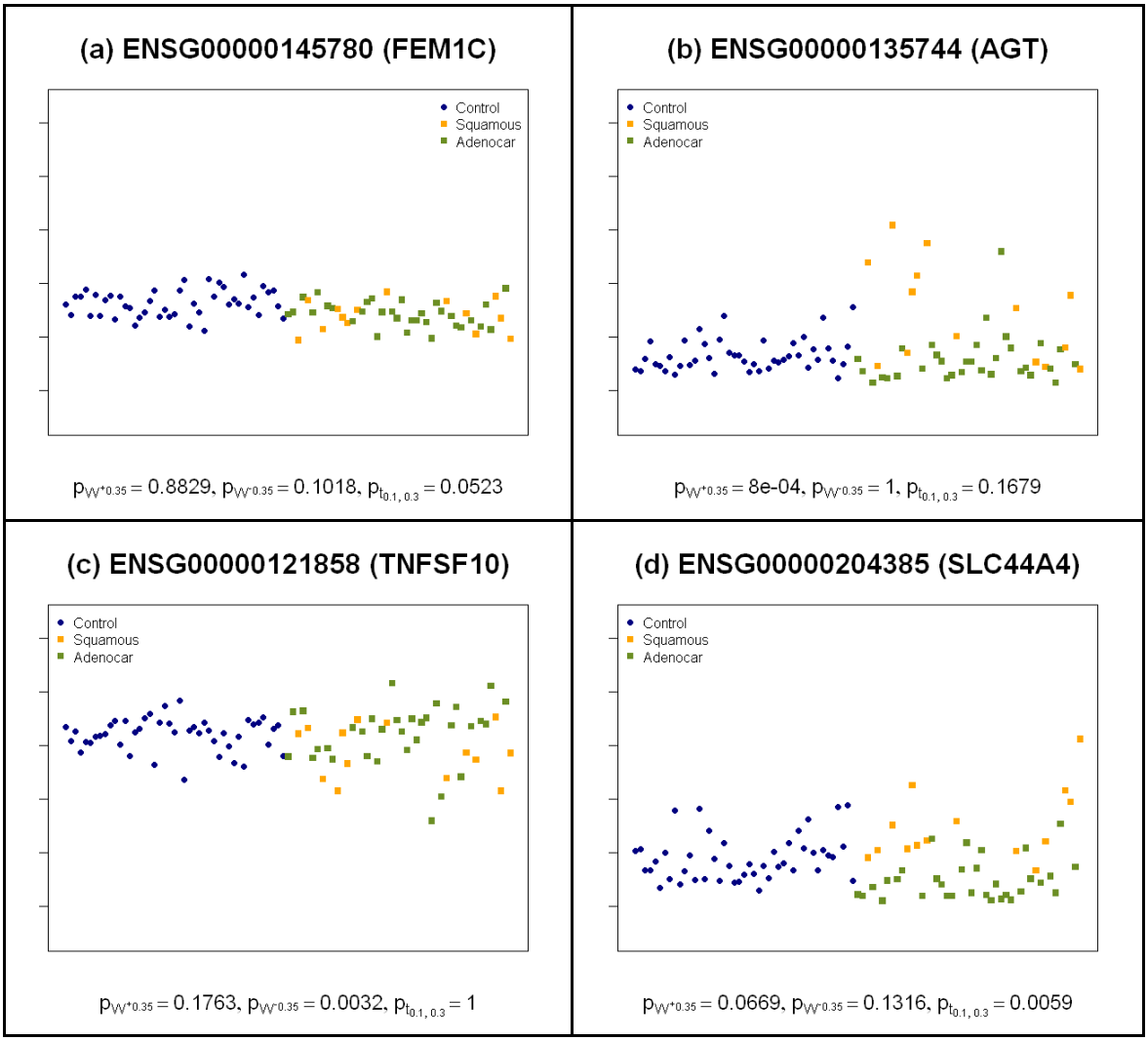


Figura 5.3. Niveles de expresión de genes (a) no declarados como significativos, a nivel 0.01, por ninguno de los métodos basados en el recorte imparcial, y (b)-(d) significativos a ese nivel para (b) $W^{+}_{0.35}$, (c) $W^{-}_{0.35}$ y (d) $t_{0.1,0.3}$, respectivamente.

En la figura 5.3 se representan algunos ejemplos de genes que se declaran significativos por alguno de los métodos. Salvo en ENSG00000145780 (FEM1C), representado en el panel superior derecho, que no resulta significativo con ninguno de los estadísticos propuestos en este capítulo, se observa cierto nivel de contaminación.

Para comparar los distintos métodos descritos en este capítulo, se consideran dos sub-conjuntos de genes, que por sus características sirven de muestras de validación,

- **Muestra I**, formada por los 170 genes con diferencia de medianas entre los dos sub-grupos de cáncer mayor o igual que 2. Son genes en los que, presumiblemente, existen dos sub-poblaciones en la muestra de tumores.
- **Muestra II**, formada por 90 genes y 70 muestras propuesta como muestra independiente de validación en el trabajo original del que se han obtenido los datos **[Sanchez-Palencia et al, 2011]**. Se evalúa la expresión de 92 genes seleccionados utilizando qRT-PCR en 70 muestras de las mismas características que las muestras originales pero procedentes de distintos individuos. Al hacer la conversión de conjuntos de sondas de *Affymetrix* a Ensemb se pierden 2 genes.

Hay una coincidencia de 41 genes entre las dos muestras de validación.

A continuación y para cada una de estas muestras, se presentan los resultados obtenidos a partir de los distintos métodos descritos en este capítulo. En todos los casos, se establece un nivel de significación de 0.01.

5.3.2.1. Resultados en la muestra de validación I

Los 170 genes que forman parte de la muestra de validación I, se dividen en dos sub-conjuntos según la mínima diferencia de medianas al grupo control. El primer conjunto, formado por 126 genes (63% de la muestra I), se caracteriza por que la mínima diferencia de medianas de cada tipo de cáncer al grupo control es menor que 0.5. Estos genes tendrán un comportamiento de tipo OHE, con una parte mayoritaria de las observaciones expresadas al mismo nivel y el resto sobre o infra expresadas. Los 74 genes restantes (37%), con mínima diferencia de medianas mayor que 0.5, se aproximan mejor a un comportamiento tipo PHE, con un tipo de cáncer diferencialmente expresado respecto de las muestras control. Por lo tanto, el primer sub-conjunto se utiliza para evaluar los estadísticos $copa$, os, ort, $most$ y $W_{0.35}$. En todos los casos, puesto que existen dos versiones, una para evaluar sobre-expresión y otra para la infra-expresión, se considera el máximo entre las dos posibilidades como resumen global. El segundo sub-conjunto se utiliza para evaluar los estadísticos SAM, t_{IQR} y $t_{0.10,0.30}$.

Tabla 5.5. Muestra I. Genes significativos, p-valor≤0.01, con cada uno de los métodos en el sub-conjunto de genes correspondiente.

Genes OHE						Genes PHE			
copa	*os*	*ort*	*most*	$W_{0.35}$	Total	*SAM*	t_{IQR}	$t_{0.10,0.30}$	Total
89	58	122	98	124	**126**	57	58	63	**74**

En la tabla 5.5 se muestra el número de genes declarados significativos a nivel 0.01 por cada uno de los métodos evaluados en cada sub-conjunto de genes. Claramente, en cuanto a número de genes detectados como significativos, los métodos basados en los recortes imparciales son superiores al resto.

En la figura 5.4 se representa la coincidencia entre métodos, en términos del porcentaje de concordancia de los genes declarados como significativos, con cada uno de los estadísticos evaluados. En todos los casos se observan porcentajes de concordancia por encima del 60%, salvo con el estadístico *os* , que sólo consigue clasificar como significativos a 58 genes de los 126 genes OHE considerados en esta muestra.

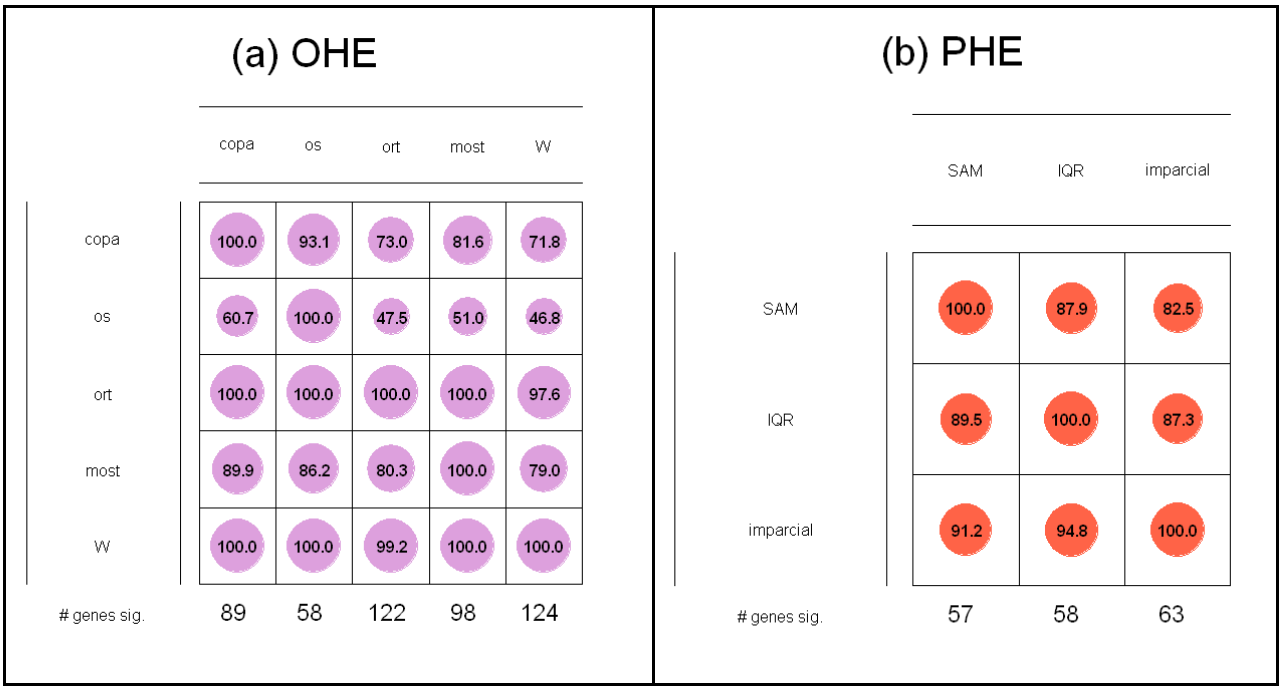


Figura 5.4. Muestra I. Porcentaje de concordancia entre los genes declarados significativos, a nivel 0.01, con cada uno de los métodos evaluados.

Las figuras 5.5 y 5.6 muestran, respectivamente, los niveles de expresión del gen no declarado como OHE y los 6 genes no declarados como PHE por ninguno de los métodos evaluados.

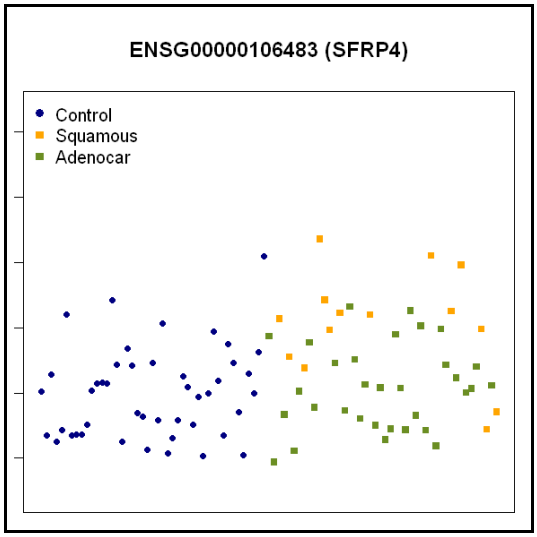


Figura 5.5. Muestra I. Nivel de expresión del gen no declarado como OHE, a nivel 0.01, por ninguno de los métodos.

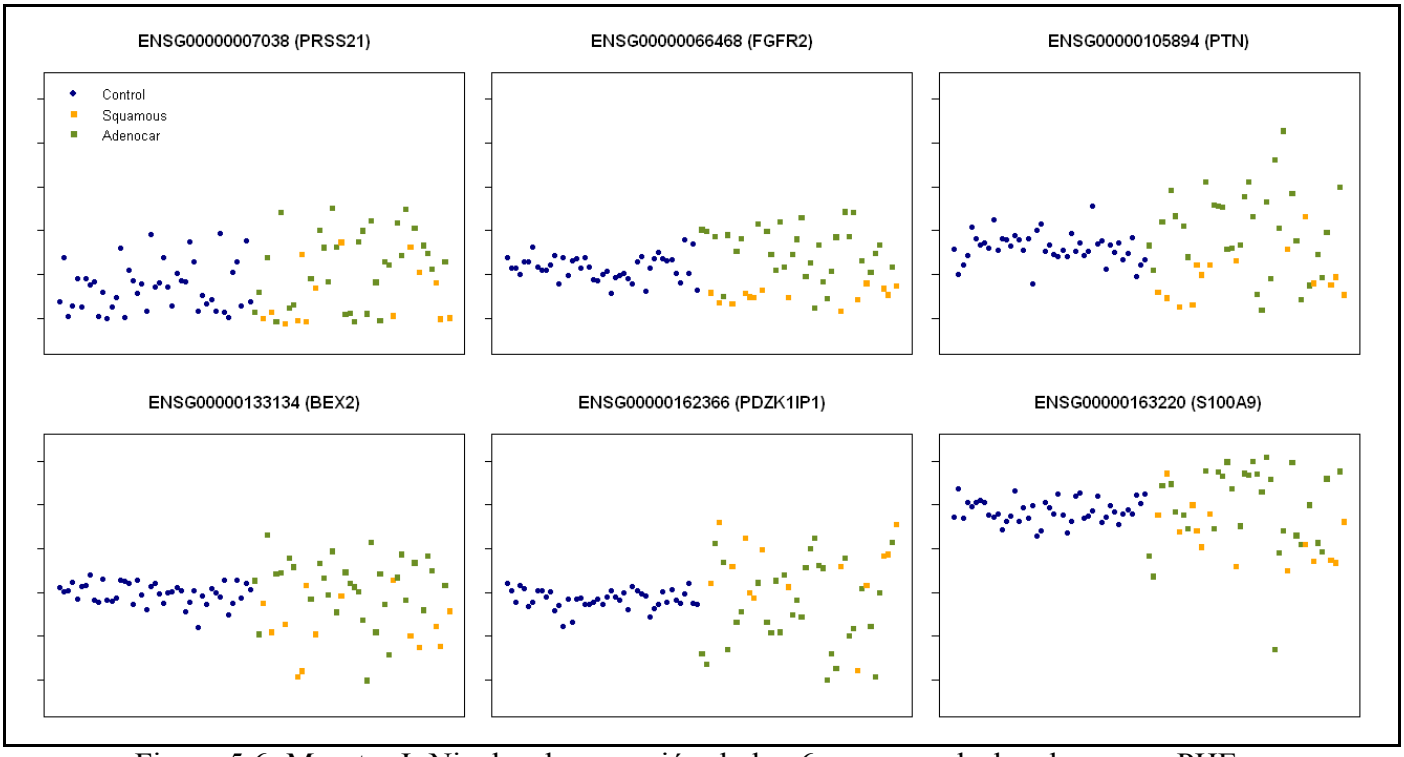


Figura 5.6. Muestra I. Niveles de expresión de los 6 genes no declarados como PHE, a nivel 0.01, por ninguno de los métodos.

Para comparar los resultados obtenidos con los métodos propuestos en este capítulo y los métodos alternativos, se considera, una primera sección en la que se evalúan los genes que no resultan significativos según los métodos basados en el recorte imparcial, y una segunda sección en la que se analizan los genes que si resultan significativos con estos métodos y no con el resto.

5.3.2.1.1. Genes no significativos según los métodos basados en el recorte imparcial

A) Genes OHE

Además del gen representado en la figura 5.5 que no se clasifica como OHE por ninguno de los métodos, el gen ENSG00000167183 (PRR15L), representado en la figura 5.7, no resulta significativo a partir del estadístico $W_{0.35}$, pero sí a partir del ort. Sin embargo, el p-valor obtenido para este gen con $W_{0.35}$ es 0.0111, muy próximo al nivel de significación fijado. Con el resto de estadísticos, $copa$, os y $most$, los p-valores obtenidos son mayores que 0.05. Por otra parte, a la vista del patrón de expresión de este gen, uno podría pensar que se trata de un gen PHE. De hecho, este gen resulta significativo utilizando el estadístico $t_{0.10,0.30}$.

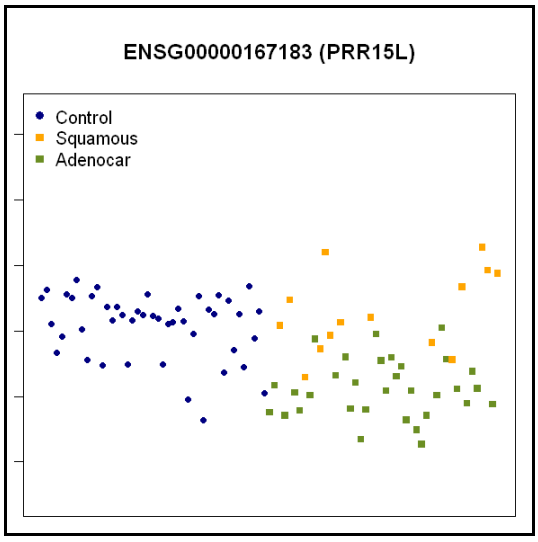


Figura 5.7. Muestra I. Niveles de expresión del gen no declarados como OHE, a nivel 0.01, por el estadístico $W_{0.35}$, y sí por alguno de los métodos alternativos para la búsqueda de esta clase de genes.

B) Genes PHE

Además de los genes representados en la figura 5.6 que no se clasifican como PHE por ninguno de los métodos, hay 5 genes más no significativos a partir del estadístico $t_{0.10,0.30}$, pero sí por alguno de los otros métodos alternativos. Estos genes se representan en la figura 5.8. Todos ellos son clasificados como genes OHE por el estadístico $W_{0.35}$.

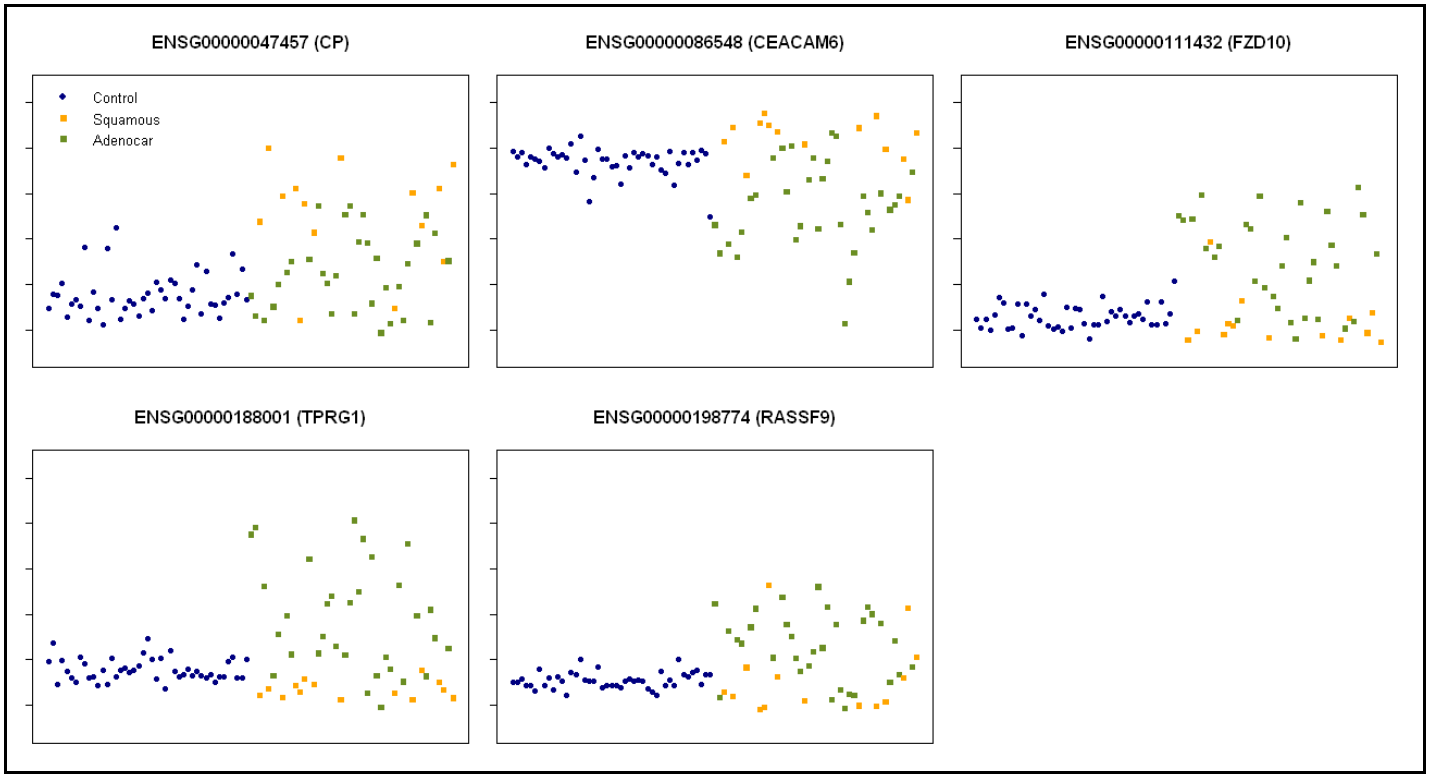


Figura 5.8. Muestra I. Niveles de expresión de los genes no declarados como PHE, a nivel 0.01, por el estadístico t-Student basado en el recorte imparcial, $t_{0.10,0.30}$, y sí por alguno de los otros métodos evaluados cuyo objetivo es encontrar genes PHE.

5.3.2.1.2. Genes significativos según los métodos basados en el recorte imparcial

En la tabla 5.6 se muestran los p-valores de 7 genes significativos con alguno de los estadísticos basados en los recortes imparciales, $t_{0.10,0.30}$ o $W_{0.35}$, pero no por los métodos alternativos que persiguen el mismo objetivo, bien encontrar genes PHE, bien genes OHE. Los tres primeros serían identificados como OHE sólo por $W_{0.35}$, mientras que los cuatro últimos sólo serían identificados como PHE por $t_{0.10,0.30}$. Los niveles de expresión de estos genes se representan en la figura 5.9. El gen ENSG00000165215 (CLDN3), representado en el panel de abajo a la izquierda, sólo sería clasificado como significativo por los métodos basados en el recorte imparcial.

Tabla 5.6. Muestra I. P-valores significativos, a nivel 0.01, según alguno de los estadísticos basados en el recorte imparcial y no por los métodos alternativos. $\log FC$ denota la diferencia de medianas entre subgrupos de cáncer.

GEN	log *FC*	Genes OHE					Genes PHE		
		copa	*os*	*ort*	*most*	$W_{0.35}$	*SAM*	t_{IQR}	$t_{0.10,0.30}$
ENSG00000070526 (ST6GALNAC1)	2.36	0.0522	0.0428	0.0109	0.7557	**0.0043**	**0.0025**	**0.0072**	0.2527
ENSG00000163993 (S100P)	2.69	0.0441	0.0544	0.0159	0.7557	**0.0054**	**0.0079**	0.0201	0.2392
ENSG00000171557 (FGG)	4.20	0.3834	1.0000	0.1100	0.1498	**0.0079**	**0.0001**	**<0.0001**	**<0.0001**
ENSG00000070731 (ST6GALNAC2)	-3.00	**0.0035**	**0.0008**	**0.0004**	0.8781	**0.0063**	1.0000	1.0000	**0.0014**
ENSG00000089356 (FXYD3)	-3.23	**0.0009**	**0.0002**	**0.0001**	0.5013	**0.0020**	0.9318	1.0000	**0.0002**
ENSG00000121552 (CSTA)	-4.26	0.0223	0.0257	**0.0019**	0.7557	**0.0037**	0.0350	0.0261	**<0.0001**
ENSG00000165215 (CLDN3)	3.30	0.0281	0.0346	0.0210	0.4761	**0.0065**	0.7327	0.4337	**0.0014**

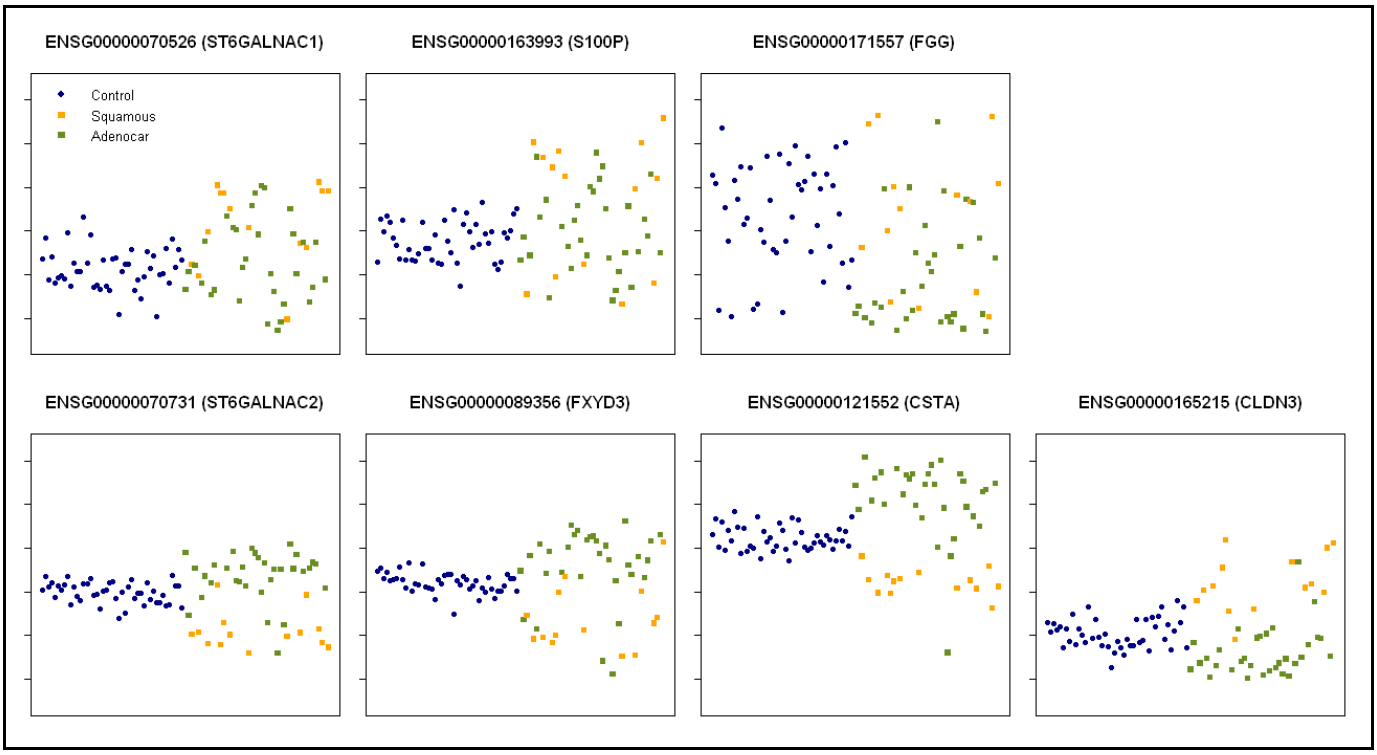


Figura 5.9. Muestra I. Niveles de expresión de los genes declarados como significativos, a nivel 0.01, por alguno de los estadísticos basados en el recorte imparcial, y no por alguno los métodos alternativos.

5.3.2.2. Resultados en la muestra de validación II

En la tabla 5.7 se muestran los genes encontrados por cada uno de los métodos evaluados según la comparación que resultó significativa a partir del análisis de referencia (qRT-PCR).

Hay 5 genes que no resultan significativos utilizando el contraste $t_{0.10,0.30}$ y 11 que no lo serían a partir de $W_{0.35}$. Los resultados se muestran en la tabla 5.8, y como se puede comprobar, sólo 2 de ellos no son clasificados como diferencialmente expresados por ninguno de los métodos (figura 5.10). Estos dos casos, son los únicos no detectados por los métodos basados en el recorte imparcial.

Tabla 5.7. Muestra II. Genes significativos, p-valor≤0.01, con cada método según la comparación que resulta significativa a partir del análisis de referencia qRT-PCR.

	qRT-PCR	Genes PHE			Genes OHE					$t_{0.10,0.30} \cup W_{0.35}$
		SAM	t_{IQR}	$t_{0.10,0.30}$	*copa*	*os*	*ort*	*most*	$W_{0.35}$	
Tumour vs Control	**22**	22	22	22	2	0	22	21	16	22
Adenocarcinoma vs Squamous	**52**	50	48	48	26	7	50	43	47	51
Adenocarcinoma vs Control	**18**	18	18	18	2	0	18	17	13	18
Squamous vs.Control	**57**	56	55	55	18	4	56	48	47	56
Total	**75**	72	70	70	30	9	72	62	64	73

Tabla 5.8. Muestra II. P-valores no significativos, a nivel 0.01, según alguno de los estadísticos basados en el recorte imparcial.

GEN	Genes PHE			Genes OHE				
	SAM	t_{IQR}	$t_{0.10,0.30}$	*copa*	*os*	*ort*	*most*	$W_{0.35}$
ENSG00000247993 (FOXD1)	0.2688	1.0000	0.2692	1.0000	1.0000	0.7811	0.2961	0.1664
ENSG00000124785 (NRN1)	**0.0001**	0.0143	0.9410	**0.0009**	**0.0002**	**0.0001**	0.0985	**0.0014**
ENSG00000148053 (NTRK2)	**<0.0001**	**0.0013**	0.3919	**0.0064**	**0.0031**	**0.0051**	0.0220	**0.0001**
ENSG00000213599 (SULT1A3)	0.9881	0.1654	0.5973	1.0000	1.0000	0.5586	0.1338	0.0564
ENSG00000072274 (TFRC)	**0.0081**	0.0223	0.1300	0.0327	0.0600	**0.0078**	0.7557	**0.0021**
ENSG00000204305 (AGER)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**0.0000**	**0.0001**	1.0000
ENSG00000068650 (ATP11A)	**0.0012**	**<0.0001**	**<0.0001**	0.3069	0.2914	**0.0001**	0.2044	0.1325
ENSG00000108821 (COL1A1)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**0.0006**	**0.0001**	1.0000
ENSG00000169031 (COL4A3)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**<0.0001**	**0.0013**	0.7391
ENSG00000164932 (CTHRC1)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**0.0005**	**<0.0001**	1.0000
ENSG00000168309 (FAM107A)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**<0.0001**	**0.0003**	1.0000
ENSG00000179348 (GATA2)	**<0.0001**	**<0.0001**	**<0.0001**	0.5654	1.0000	**<0.0001**	0.0899	1.0000
ENSG00000144791 (LIMD1)	**<0.0001**	**<0.0001**	**<0.0001**	0.3412	1.0000	**<0.0001**	0.0709	0.0213
ENSG00000186340 (THBS2)	**<0.0001**	**<0.0001**	**<0.0001**	1.0000	1.0000	**0.0003**	**<0.0001**	0.2380

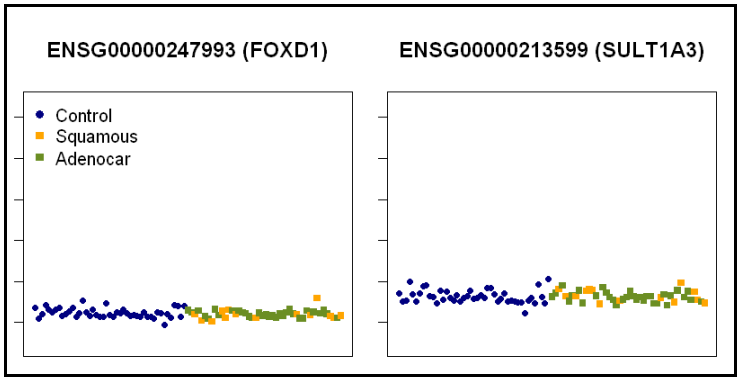


Figura 5.10. Muestra II. Niveles de expresión de los genes no declarados como significativos, a nivel 0.01, por ninguno de los métodos evaluados.

En las siguientes figuras se muestran los niveles de expresión de, figura 5.11, los 3 genes no significativos según $t_{0.10,0.30}$, y, figura 5.12, de los 9 genes no clasificados como OHE a partir de $W_{0.35}$. En ambos casos, los patrones de contaminación responden al otro tipo de gen, OHE y PHE respectivamente.

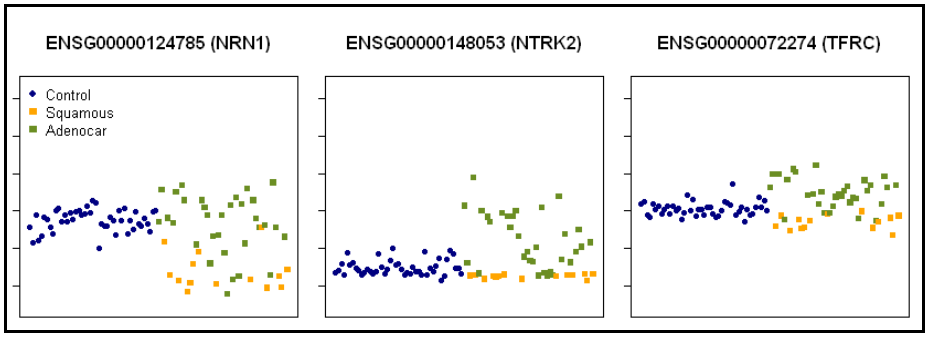


Figura 5.11. Muestra II. Niveles de expresión de los genes no declarados como significativos, a nivel 0.01, por el estadístico $t_{0.10,0.30}$.Todos resultan significativos a ese nivel utilizando $W_{0.35}$.

5.3.3. Aplicación al *dataset* de Cáncer de Mama I

En este conjunto de datos, descrito en **[Hatzis et al, 2011]** y en la sección B.3 del apéndice B, se analizan 12576 genes en 310 muestras. Estas muestras se dividen en dos grupos, 113 pacientes sensibles a un tratamiento de quimioterapia y 197 no sensibles a dicho tratamiento.

Se considera un nivel de recorte de $\alpha = 0.45$ para la búsqueda de genes OHE a partir del estadístico definido en la expresión (5.4). Para los genes PHE se consideran los niveles

$\alpha_1 = 0.25$ y $\alpha_2 = 0.25$ en el estadístico t-Student recortado cuya expresión se especifica en (5.5). En la tabla 5.9 se resumen los resultados obtenidos. En total se declaran significativos, a nivel 0.01, 3273 genes, aproximadamente el 26% de los genes evaluados.

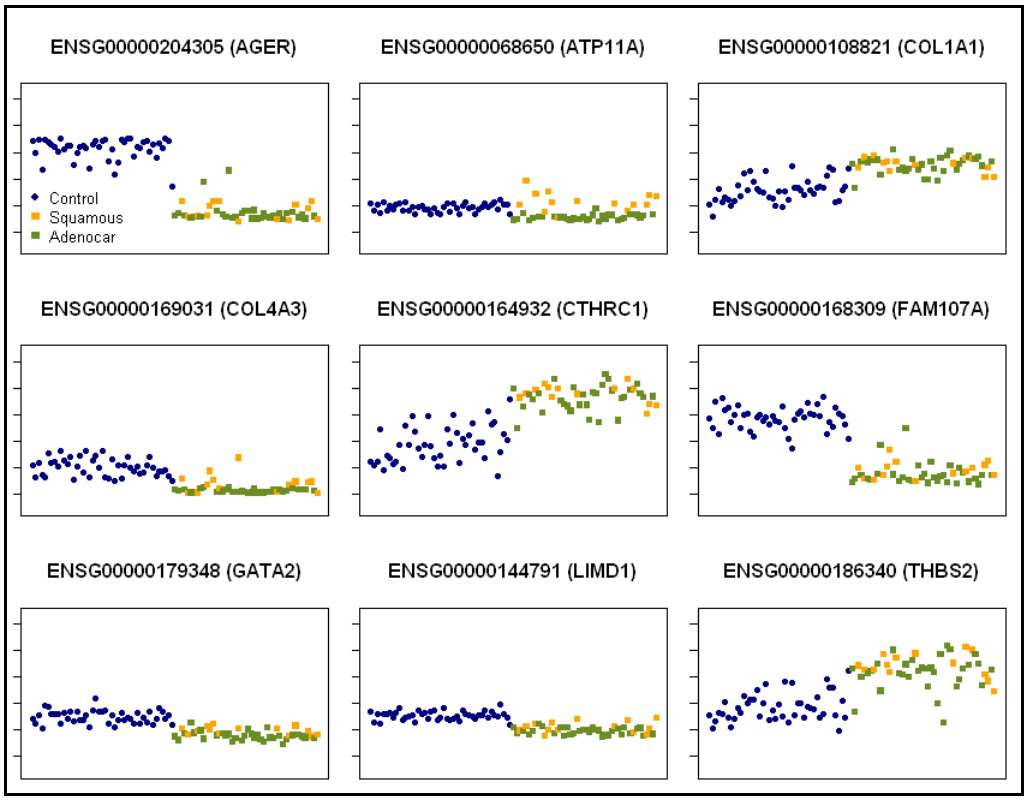


Figura 5.12. Muestra II. Niveles de expresión de los genes no declarados como significativos, a nivel 0.01, por el estadístico $W_{0.35}$. Todos resultan significativos a ese nivel utilizando $t_{0.10,0.30}$.

Evaluando los perfiles de expresión de algunos genes que mostraban discrepancias en los resultados obtenidos con los estadísticos propuestos en este trabajo y los alternativos, llama la atención cierto comportamiento que aparece en algún caso como los mostrados en la figura 5.13. Claramente aparece un sub-conjunto de muestras en cada grupo que podrían responder a cierto patrón de contaminación. Analizando las posibles causas descubrimos que la razón de estos cambios en los niveles de expresión podía corresponder a la procedencia de las muestras. Esta variable, etiquetada como *source*, toma dos posibles valores: ISPY, 83 muestras procedentes del consorcio I-SPY-1 (*Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis*), y MDACC, 227 muestras procedentes del departamento de patología del M.D. *Anderson Cancer Center*, Houston, Texas.

Tabla 5.9. Genes significativos, p-valor≤0.01, entre los 12576 genes evaluados en el *dataset* de Cáncer de Mama I.

Contraste		# genes	% genes	% genes únicos	Concordancia entre contrastes (%)		
					$W^{+}_{0.45}$	$W^{-}_{0.45}$	$t_{0.25,0.25}$
OHE	$W^{+}_{0.45}$	742	5.9	68.85	100	7.21	23.93
	$W^{-}_{0.45}$	1903	15.13	90.84	8.89	100	0.27
PHE	$t_{0.25,0.25}$	915	7.28	88.39	11.51	0.11	100

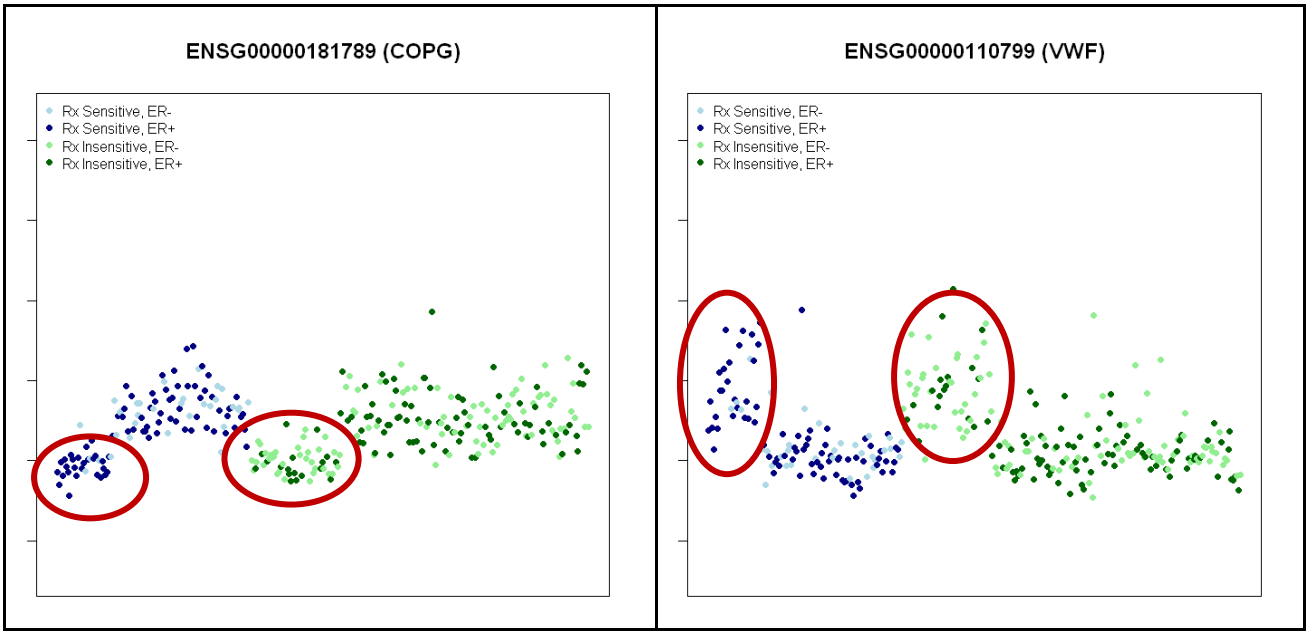


Figura 5.13. Ejemplos de genes con patrones de contaminación debido a la procedencia de la muestra en el *dataset* de Cáncer de Mama I.

Con el fin de obtener un listado de genes posiblemente afectados por esta variable se llevó a cabo el siguiente procedimiento,

1) Utilizar $W_{0.45}$ como filtro previo. Los genes buscados se caracterizan por un conjunto de muestras *outlier* respecto de la tendencia general de expresión de cada uno de ellos, tanto en el grupo de muestras de pacientes sensibles al tratamiento como en el grupo de pacientes insensibles, por tanto, el contraste basado en el estadístico $W_{0.45}$ debería resultar significativo. Se establece un nivel de significación de 0.01.

2) El estadístico $W_{0.45}$ clasifica cada muestra en cada gen como *outlier* o no-*outlier*. Cruzando este resultado con la procedencia de la muestra, es posible definir una medida de la sensibilidad del estadístico para clasificar a una muestra procedente de ISPY como atípica, y una medida de la especificidad para clasificar a una muestra procedente de MDACC como no atípica.

3) Para dar una idea de la significación de estas medidas de acierto, se calcula el p-valor *bootstrap* como la frecuencia de veces que B permutaciones aleatorias de los valores de la variable *source*, proporcionan valores de sensibilidad y especificidad mayores que los observados en la muestra original. Se ha utilizado $B = 1000$.

A nivel 0.01, hay 412 genes cuyo patrón de contaminación podría deberse a la variable *source*.

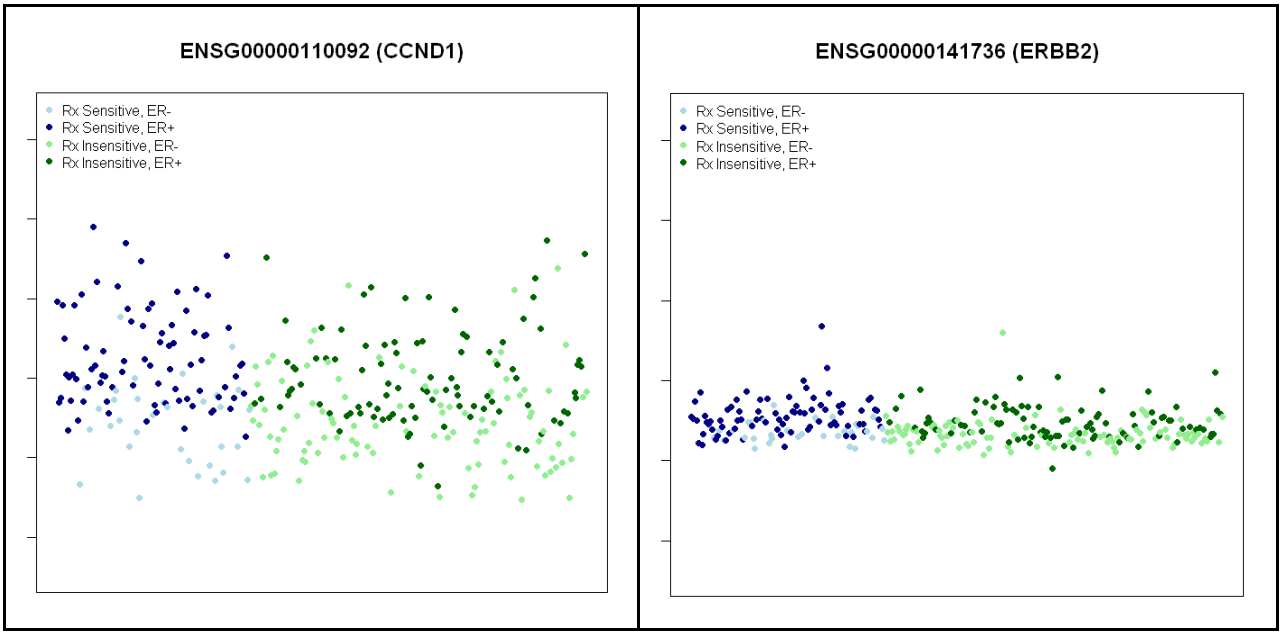


Figura 5.14. Ejemplos de genes sin el efecto de la variable *source* en el *dataset* de Cáncer de Mama I. Estos dos genes se sabe que están muy relacionados con el cáncer de mama.

La hipótesis que se baraja, es que las diferencias en los niveles de expresión debidas a la variable *source*, tienen que responder a una causa biológica. Si no fuese así, esas diferencias deberían observarse en todos los genes. Sin embargo, existen genes como los representados en la figura 5.14, muy relacionados con el cáncer de mama, en los que no se observa este patrón. En la tabla 5.10 se muestran las categorías funcionales sobre-representadas en términos de GO-BP y KEGG. Todas ellas tienen que ver con el movimiento celular y la liberación de energía, y representan funciones básicas en la supervivencia de las células.

Tabla 5.10. Análisis de enriquecimiento funcional de los 412 genes con un efecto de la variable *source*. Se muestran las categorías sobre-representadas (p-valor FDR≤0.05).

# genes anotados	GO biological process		KEGG pathway	
	Categoría	# genes	Categoría	# genes
401	GO:0006119~oxidative phosphorylation	14	hsa04510:Focal adhesion	20
	GO:0006091~generation of precursor metabolites and energy	24		
	GO:0015980~energy derivation by oxidation of organic compounds	15		
	GO:0006928~cell motion	29		

Capítulo 6

Biclustering para identificar patrones de co-expresión

El término *biclustering* fue utilizado inicialmente en el análisis de expresión de genes por **[Cheng y Church, 2000]**, aunque la técnica fue introducida originariamente mucho antes, primero por **[Morgan y Sonquistz, 1963]** y posteriormente por **[Hartigan, 1972]**. Revisiones clásicas de esta metodología son **[Busygin et al, 2008]**; **[Madeira y Oliveira, 2004]**; **[Van Mechelen et al, 2004]** y **[Tanay et al, 2002]**. Su objetivo es agrupar simultáneamente genes y arrays, permitiendo encontrar patrones de expresión en grupos de genes que sólo se presentan bajo un determinado conjunto de condiciones experimentales. En realidad, los genes tienden a estar co-regulados, y en consecuencia co-expresados, no en todos los arrays, sino en sub-conjuntos de ellos. Por lo tanto, las técnicas de *clustering* proporcionan soluciones insuficientes a los problemas de clasificación relacionados con las matrices de expresión, haciéndose esto más patente a medida que es posible manejar matrices con un mayor número de muestras que contienen distintos tipos de condiciones o clases.

El *biclustering*, por tanto, engloba procedimientos de clasificación no supervisada que, dada una matriz de datos $\mathrm{Y} = (G, A)$, agrupan conjuntamente filas y columnas. En adelante, Y representará a la matriz de expresión donde las filas son genes, $G = \{g_1, g_2, ..., g_{n_G}\}$, y las columnas muestras biológicas o arrays, $A = \{a_1, a_2, ..., a_{n_A}\}$. Sean $I \subset G$ y $J \subset A$ sub-conjuntos de genes y arrays respectivamente, la sub-matriz $\mathrm{Y}_{IJ} = (I, J)$ será un biclúster si satisface ciertas condiciones de homogeneidad. Típicamente, esta homogeneidad puede estar relacionada con que todos los valores de expresión en el biclúster sean similares, o bien, con que la expresión varíe de la misma forma en todos los arrays.

Actualmente existen un gran número de propuestas para la búsqueda de biclústers partiendo de una matriz de expresión. El problema presenta una complejidad muy superior al de la búsqueda de clústers, del que ya es conocida su complejidad. Estas propuestas para la obtención de biclústers incorporan algoritmos iterativos basados en heurísticas o en modelos estadísticos. Adicionalmente, se recomienda realizar un pre-procesado previo de los datos, con el fin de reducir el problema y hacer más evidentes los patrones de interés.

Buscando que los procedimientos propuestos agrupen patrones de expresión similares y no simplemente expresión, algunos de los métodos disponibles incluyen algún tipo de estandarización en los algoritmos. Por ello, muchos procedimientos de *biclustering*, toman como punto de partida una matriz de expresión estandarizada como por ejemplo, el ITWC (*Interrelated Two-way Clustering)* **[Tang et al, 2001]**, el SAMBA (*Statistic-Algorithmic Method for Bicluster Analysis*) **[Tanay et al, 2002]**, el propuesto en **[Farcomeni, 2009]** o en **[Ayadi et al, 2012]**. El procedimiento BBC (*Bayesian BiClustering model*) de **[Gu y Liu, 2008]**, propone utilizar la media y desviación típica recortada (no-imparcial) con el objetivo de robustificar esta estandarización. Otros procedimientos como el ISA (*Iterative Signature Algorithm*) **[Bergmann et al, 2003]**, el ISA-$Q_{1/2}$ **[Freitas et al, 2011]** y el SMOB (*Sequential Multi-Objetive Biclustering*) de **[Divina et al, 2012]** incorporan esta normalización en el algoritmo aplicado, dentro de las iteraciones.

Nuestra aproximación al problema de *biclustering* se encuentra dentro de los procedimientos basados en modelos estadísticos, y más concretamente dentro de las respuestas que buscan funcionamiento robusto mediante la aplicación de recortes imparciales. Este tipo de aproximación ha sido aplicada con éxito en la búsqueda de clústers. La primera propuesta en este ámbito corresponde a las k-medias recortadas de **[Cuesta-Albertos et al, 1997]**, que robustifican las k-medias al buscar k centros a la vez y eliminar un porcentaje de individuos de la muestra. Más recientemente, el TCLUST de **[García-Escudero et al, 2008]** y la propuesta de **[Gallegos y Ritter, 2009]**, aportan mayor flexibilidad a la búsqueda de clústers incorporando la posibilidad de diferentes matrices de covarianza y pesos en las poblaciones. Como aplicación de estas ideas al *biclustering*, **[Farcomeni, 2009]** propone un procedimiento, basado en recortes imparciales, que generaliza las k-medias recortadas al problema de clusterizar simultáneamente filas (genes) y columnas (arrays), dejando sin clasificar un porcentaje de genes y de arrays determinado por la muestra de forma automática. Algunos procedimientos que buscan biclústers de forma secuencial pueden dejar sin clasificar algunos genes y arrays, como el δ-biclúster de **[Cheng y Chuch,2000]** o el ISA **[Bergmann et al, 2003]**, y de esta forma conseguir cierta robustez asociada a los recortes. En cualquier caso, la robustez asociada a este forma de aplicar los recortes, que no fija el tamaño de los mismos, no se puede cuantificar y podría, en algunos casos extremos, llegar a ser muy débil.

El primero de los procedimientos que proponemos en este capítulo busca clústers de co-expresión. El objetivo es encontrar una partición de la matriz en grupos de genes homogéneos en el sentido de la co-expresión. Pre-fijamos una proporción de genes a no ser clasificados y también dejamos fuera de la clasificación, para cada grupo de genes, una proporción de arrays, que puede diferir de unos grupos a otros. De esta forma no es obligatorio que los patrones de co-expresión encontrados tengan que afectar a todos los arrays. El procedimiento propuesto tiene conexiones con el utilizado en las aproximaciones de *clustering* robusto realizadas por **[García-Escudero et al, 2008]** o **[Gallegos y Ritter, 2009]**, para el problema de clasificar individuos en un espacio de *p* dimensiones. Las diferencias con aquella metodología corresponden a que las matrices de covarianzas, aquí, están restringidas a ser diagonales, y a que se elimina de la estimación, no solo un porcentaje de

individuos (en este caso, genes), sino también un porcentaje de variables (arrays), posiblemente diferentes para cada clúster. Además, en nuestra aproximación, los datos de cada gen, necesitan de un re-alineamiento o estandarización, que se realiza en cada paso del algoritmo.

El segundo de los procedimientos que proponemos busca simultáneamente clústers de genes y arrays. Más concretamente, el objetivo es encontrar agrupaciones de genes con distribuciones de expresión, que una vez estandarizadas, permitan agrupaciones similares para los arrays. La estandarización, por gen, se realiza respecto al grupo de arrays mayoritario. Este procedimiento estaría conectado al ofrecido por **[Farcomeni, 2009]** anteriormente mencionado. En nuestra aproximación, buscamos flexibilizar su modelo incluyendo parámetros correspondientes a la variabilidad y al peso de las poblaciones estimadas, además de aplicar el alineamiento de los genes, necesario para encontrar co-expresión, en cada paso del algoritmo en vez de realizarlo previamente a la aplicación de la metodología. Esto tiene la ventaja de que el re-alineamiento solo está basado en las estimaciones disponibles para el resto de parámetros y en los arrays incluidos en el análisis.

Nuestra última aportación dentro de las técnicas de *biclustering*, se encuadra en la búsqueda de patrones de atipicidad en matrices de datos de expresión. Los procedimientos aplicados en el capítulo 5 ofrecían una selección de genes que contenían algún array atípico. Este tipo de listados puede contener numerosos genes, motivados por la atipicidad de conjuntos de arrays diferentes para cada uno de ellos. Esto los puede hacer difícilmente interpretables y, por tanto, su utilidad puede ser dudosa. Nuestra propuesta está relacionada con post-procesar estas listas buscando agrupaciones de genes que muestren comportamientos de atipicidad similares en términos de los arrays implicados. Esto permitiría reducir estos listados a unos pocos grupos de genes, cada uno de ellos caracterizado por el comportamiento atípico de un grupo de arrays. Por tanto, la realización de este tipo de análisis, a nuestro juicio, es recomendable para conseguir interpretabilidad de los *outputs* correspondientes a los procedimientos que identifican comportamientos atípicos en la matriz de expresión. De hecho, tras la realización de estos análisis, es posible relacionar el reducido conjunto de patrones de atipicidad con las variables clínico-patológicas disponibles, para intentar explicar esos patrones observados. Incluso, en situaciones en las que los patrones de *outliers* aparezcan muy claramente definidos en un grupo de genes biológicamente coherente y no puedan ser explicados por ese grupo de variables clínicas, podrían estar identificando la presencia de una variable *hidden* responsable de ese comportamiento observado. Como punto de partida, para la búsqueda de estos patrones de atipicidad, se utiliza una matriz de datos binaria, transformación de la matriz de datos de expresión original, en la que aparecen con un 1 aquellos pares gen-condición identificados como *outliers* en un análisis previo. Este punto de partida es similar al propuesto en el algoritmo Bimax **[Prelic et al, 2006]**, que utiliza también una matriz binaria, donde, en su caso, los valores 1 aparecen asociados a condiciones que superaran un 2-*fold* respecto de un experimento control. Para la búsqueda de patrones de *outliers*, proponemos una modificación del algoritmo de biclúster ofrecido en nuestra segunda propuesta de este capítulo. De esta

forma, obtenemos ventaja de la situación simplificada a la que nos enfrentamos en esta ocasión, ya que, no es necesario aplicar ningún tipo de estandarización, y, gracias a la naturaleza binaria de la matriz de expresión, la localización de los clústers es conocida.

6.1. Métodos propuestos

6.1.1. Búsqueda de clústers de co-expresión: $K_G x 1$

El objetivo de este análisis es encontrar grupos de genes que co-expresan, particionando el conjunto de genes en grupos disjuntos y permitiendo que un porcentaje de ellos quede sin clasificar. La co-expresión que buscamos debe estar determinada por una mayoría de arrays, pero no necesariamente tiene porque implicar a todos, ni deben ser los mismos en cada uno de los grupos de genes. Así, un porcentaje de arrays no participará en la identificación de cada clúster de genes, y estos arrays no tienen porque ser los mismos en cada uno de los clústers. Suponemos que los genes que co-expresan, tras ser estandarizados, muestran niveles de expresión similares. Aplicamos este método de clasificación a matrices de expresión génica, de las que, previamente, se han eliminado genes que muestran insuficiente variabilidad en su núcleo de expresión. Un caso particular, interesante para la aplicación de este procedimiento, sería el correspondiente a la determinación del mejor sub-conjunto de sondas, en un conjunto de sondas disponibles, para representar la expresión de un gen. En este caso, la aplicación del método requeriría fijar el número de sondas a buscar y el porcentaje de arrays que no participarían en esa identificación.

Asumimos que la matriz de expresión está dividida en K_G grupos de co-expresión de genes, que una proporción $\alpha_G \in [0,1]$ de genes no pertenece a ninguno de los K_G grupos, y que una proporción de arrays, $\alpha_A \in [0,\ 1]$, no necesariamente los mismos en cada grupo de genes, no contribuye al patrón de co-expresión existente en el grupo de genes correspondiente. Tanto el número de clústers de genes, K_G, como los supuestos niveles de recorte para las dos dimensiones, α_G para genes y α_A para arrays, deberán ser establecidos antes de aplicar esta metodología.

De esta forma, suponemos un modelo en el que el vector de expresión observado para cada gen compartiendo grupo, procede de la misma normal multivariante. La correspondiente matriz de covarianzas de esta distribución será diagonal, por suponer que los datos de expresión de diferentes arrays son independientes. La inclusión de un parámetro peso, para cada grupo de genes en la mezcla, permite no restringir, en algún sentido, los tamaños de los grupos a ser iguales. Como se ha señalado, admitimos que una proporción de genes no provienen de esa mezcla y que dentro de cada grupo de genes una proporción de arrays, específica de cada grupo, no sigue el modelo. Adicionalmente, asumimos que los genes se

observan en una escala que es una transformación lineal desconocida, posiblemente diferente para cada unos de ellos, de su escala tipificada original.

El modelo necesita, para poder ser estimado, que todos los parámetros correspondientes a la variabilidad de los arrays estén acotados inferiormente, o que el tamaño relativo de los parámetros que estiman esta variabilidad esté acotado. Para datos de expresión provenientes de microarrays, cuya log-escala es conocida a *priori*, la restricción más simple para ser aplicada, corresponde a fijar un umbral inferior para las varianzas. En el caso de elegir controlar el tamaño relativo de la variabilidad puede adaptarse la metodología desarrollada en **[Fritz et al, 2012]** para ser incorporada al correspondiente algoritmo. De la misma forma, será necesario restringir los parámetros que estiman la variabilidad de los genes. Como en el caso de los arrays, recomendamos la elección de una cota inferior para restringir la varianza de los genes.

6.1.1.1. Parámetros a estimar

Los parámetros del modelo serán,

(i) Parámetros característicos de cada clúster de genes:

- perfil medio de expresión, $\mu_k = \left(\mu_{1k},...,\mu_{ak},...,\mu_{n_A k}\right)$,
- desviación típica, $\sigma_k = \left(\sigma_{1k},...,\sigma_{ak},...,\sigma_{n_A k}\right)$ con $1 \le a \le n_A$ y $1 \le k \le K_G$
- proporción de genes o peso, denotado por π_k. Estos pesos verifican $0 \le \pi_k \le 1$ y $\sum_{k=1}^{K_G} \pi_k = 1$.

Nos referiremos con $\Theta_k = \left\{\pi_k,\ \mu_k, \sigma_k\right\}$ al conjunto de parámetros del clúster k y por $\Theta = \left\{\Theta_k\right\}_{1\le k\le K_G}$ a todos los parámetros relacionados con estos clústers de genes.

(ii) Característicos de cada gen: patrón de expresión medio de cada gen y su desviación típica, m_g y s_g, $1 \le g \le n_G$. Estos dos parámetros se utilizan para la estandarización de los datos de expresión del gen y se nombran utilizando $\mathrm{E}_g = \left\{m_g, s_g\right\}$. Con $E = \left\{E_g\right\}_{1\le g\le n_G}$ nos referiremos a todos los parámetros de este tipo.

(iii) Conjuntos de genes y arrays activos, no recortados. Se denota por B^G_{k,α_G} al conjunto de genes asignado al clúster de genes k, y B^A_{k,α_A} representa el conjunto de arrays no recortados correspondiente al grupo de genes k. $B^G_{\alpha_G} = \left\{B^G_{k,\alpha_G}\right\}_{1\le k\le K_G}$ y $B^A_{\alpha_A} = \left\{B^A_{k,\alpha_A}\right\}_{1\le k\le K_G}$ representan los conjuntos de genes y arrays no recortados en toda la partición.

6.1.1.2. Función objetivo

Dada $Y=\left\{y_{ga}\right\}_{1\le g\le n_G,\ 1\le a\le n_A}$ la matriz de expresión, y fijado un nivel de recorte $\alpha_G\in[0,1]$ en la dimensión de genes y un nivel, $\alpha_A\in[0,\ 1]$ en la dimensión de arrays, la función objetivo viene dada por la expresión,

$$\Psi\left(Y;B^G_{\alpha_G},B^A_{\alpha_A},\Theta,E\right)=$$

$$=\frac{1}{2\cdot n_G\cdot n_A}\sum_{k=1}^{K_G}\sum_{g=1}^{n_G}\sum_{a=1}^{n_A}I_{B^G_{k,\alpha_G}}(g)\cdot I_{B^A_{k,\alpha_A}}(a)\left(\left(\frac{\frac{y_{ga}-m_g}{s_g}-\mu_{a,k}}{\sigma_{a,k}}\right)^2+\log\left(\sigma^2_{a,k}\right)\right) \qquad (6.1)$$

$$-\frac{1}{n_G\cdot n_A}\sum_{k=1}^{K_G}\sum_{g=1}^{n_G}I_{B^G_{k,\alpha_G}}(g)\cdot\log(\pi_k)$$

con las restricciones,

$$R=\left\{\left(\sigma_1,...,\sigma_{K_G}\right):\ \sigma_{ak}\ge c_R,\ 1\le a\le n_A\right\} \qquad (6.2)$$

y

$$S=\left\{s_g\ge c_S,\ 1\le g\le n_G\right\} \qquad (6.3)$$

Suponiendo conocidos los parámetros necesarios para la estandarización, la obtención de estimaciones se corresponde con la búsqueda de los valores de los parámetros Θ, $B^G_{\alpha_G}$ y $B^A_{\alpha_A}$, que minimizan la función objetivo anterior, Ψ, en el espacio paramétrico restringido por la restricción R. Para obtener estas estimaciones se puede utilizar un algoritmo tipo EM que, en pasos E obtenga, dados los parámetros Θ, las asignaciones a los clúster o a recorte de los genes y los arrays dentro de cada clúster de genes; y en los pasos M optimizar Θ dadas las asignaciones. Estos pasos E y M mejorarán, o al menos no empeorán, el valor de la función objetivo. Como los parámetros de la distribución de los genes necesarios para la estandarización, son desconocidos en los pasos E, incluiremos la estimación de estos parámetros, que permiten realizar el re-alineamiento de los valores de expresión observados. En esta estimación se aplican restricciones equivalentes a la restricción R, dadas por la restricción S definida en (6.3), para evitar que las varianzas estimadas se aproximen a 0.

6.1.1.3. Algoritmo

El algoritmo que proponemos está basado en las iteraciones EM mencionadas. Aquí incluimos nuestra propuesta para la obtención de soluciones iniciales y para las iteraciones. Se tomará como estimación de los parámetros la correspondiente al mejor valor de la función objetivo

obtenido tras aplicar a un número predeterminado de comienzos iniciales y un número predeterminado de iteraciones.

6.1.1.3.1. Solución inicial

Se eligen al azar 2 genes por cada uno de los K_G clúster de genes y 2 arrays, también de manera aleatoria, en cada uno de ellos. Dados $B_{k,\alpha_G}^{G(0)}$, el conjunto de genes representantes del clúster k y $B_{k,\alpha_A}^{A(0)}$, el conjunto de 2 arrays en ese mismo grupo, se eligen como valores iniciales de los parámetros para la estandarización, $\hat{E}_g^{(0)} = \left\{ \hat{m}_g^{(0)}, \hat{s}_g^{(0)} \right\}$, los que se obtienen a partir de las expresiones,

$$\hat{m}_g^{(0)} = \frac{\sum_{a=1}^{n_A} I_{B_{k,\alpha_A}^{A(0)}}(a) \cdot y_{ga}}{\sum_{a=1}^{n_A} I_{B_{k,\alpha_A}^{A(0)}}(a)} \tag{6.4}$$

$$\hat{s}_g^{(0)} = \sqrt{\frac{\sum_{a=1}^{n_A} I_{B_{k,\alpha_A}^{A(0)}}(a) \cdot \left(y_{ga} - \hat{m}_g^{(0)} \right)^2}{\sum_{a=1}^{n_A} I_{B_{k,\alpha_A}^{A(0)}}(a)}} \tag{6.5}$$

Como valores iniciales para las estimaciones de los parámetros del clúster k , se utilizan los resultantes de las siguientes expresiones,

$$\hat{\pi}_k^{(0)} = \frac{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g)}{\sum_{k=1}^{K_G} \sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g)} \tag{6.6}$$

$$\hat{\mu}_{a,k}^{(0)} = \frac{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g) \cdot \left(\frac{y_{ga} - \hat{m}_g^{(0)}}{\hat{s}_g^{(0)}} \right)}{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g)} \tag{6.7}$$

$$\hat{\sigma}_{a,k}^{(0)} = \sqrt{\frac{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g) \cdot \left(\frac{y_{gj} - \hat{m}_g^{(0)}}{\hat{s}_g^{(0)}} - \hat{\mu}_{a,k}^{(0)} \right)^2}{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(0)}}(g)}} \tag{6.8}$$

Con $\hat{\Theta}_k^{(0)} = \left\{ \hat{\pi}_k^{(0)}, \hat{\mu}_{\mathrm{a,k}}^{(0)}, \hat{\sigma}_{a,k}^{(0)} \right\}$ se representa al conjunto de estimaciones iniciales para los parámetros del clúster k .

Input: Y : matriz de expresión con n_G genes y n_A arrays

Input: K_G : número de grupos en la dimensión de genes

Input: α_G : nivel de recorte en la dimensión de genes

Input: α_A : nivel de recorte en la dimensión de arrays para cada grupo de genes

Input: nr : número de comienzos aleatorios

Input: nl : número máximo de iteraciones EM

Mientras $r \leq nr$ **hacer**

Definir $B_{\alpha_G}^{G(0)}$, asignación aleatoria de 2 genes a cada uno de los K_G grupos

Establecer $B_{\alpha_A}^{A(0)}$, extrayendo de manera aleatoria 2 arrays en cada grupo

Estimar $\hat{E}^{(0)} = \left\{\hat{E}_g^{(0)}\right\}_{1\leq g\leq n_G}$ con $\hat{E}_g^{(0)} = \left\{\hat{m}_{\mathrm{g}}^{(0)}, \hat{s}_g^{(0)}\right\}$ según (6.4) y (6.5)

Estimar $\hat{\Theta}^{(0)} = \left\{\hat{\Theta}_k^{(0)}\right\}_{1\leq k\leq K_G}$ con $\hat{\Theta}_k^{(0)} = \left\{\hat{\pi}_k^{(0)},\ \hat{\mu}_{\mathrm{a,k}}^{(0)}, \hat{\sigma}_{a,k}^{(0)}\right\}$ según (6.6)- (6.8)

Mientras $l \leq nl$ o $parada$ **hacer**

E.l.a: Calcular $\psi_{g,k}^{(a)}\left(y_{ga}; B_{k,\alpha_G}^{G(l-1)}, B_{k,\alpha_A}^{A(l-1)}, \hat{\Theta}_k^{(l-1)}, \hat{E}_g^{(l-1)}\right)\ 1 \leq g \leq n_G$ según (6.9)

Asignar el gen g al grupo $k^{(l)}$, $1 \leq g \leq n_G$

Construir $B_{\alpha_G}^{G(l)}$

E.l.b: Calcular $\psi_{a,k}^{(b)}\left(y_{ga}; B_{k,\alpha_G}^{G(l)}, B_{k,\alpha_A}^{A(l-1)}, \hat{\Theta}_k^{(l-1)}, \hat{E}_g^{(l-1)}\right)\ 1 \leq a \leq n_A$ según (6.10)

Construir $B_{\alpha_A}^{A(l)}$

E.l.c: Estimar $\hat{E}^{(l)} = \left\{\hat{E}_g^{(l)}\right\}_{1\leq g\leq n_G}$ con $\hat{E}_g^{(l)} = \left\{\hat{m}_g^{(l)}, \hat{s}_g^{(l)}\right\}$ según (6.4) y (6.5)

M.l: Estimar $\hat{\Theta}^{(l)} = \left\{\hat{\Theta}_k^{(l)}\right\}_{1\leq k\leq K_G}$ con $\hat{\Theta}_k^{(l)} = \left\{\hat{\pi}_k^{(l)},\ \hat{\mu}_{\mathrm{a,k}}^{(l)}, \hat{\sigma}_{a,k}^{(l)}\right\}$ según (6.6) - (6.8)

Si $\Psi\left(Y; B_{\alpha_G}^{G(l-1)}, B_{\alpha_A}^{A(l-1)}, \hat{\Theta}^{(l-1)}, \hat{E}^{(l-1)}\right) = \Psi\left(Y; B_{\alpha_G}^{G(l)}, B_{\alpha_A}^{A(l)}, \hat{\Theta}^{(l)}, \hat{E}^{(l)}\right)$ **entonces**

$parada = Verdadero$

Fin si

Calcular $\Psi_r = \Psi\left(Y; B_{\alpha_G}^{G(l)}, B_{\alpha_A}^{A(l)}, \hat{\Theta}^{(l)}, \hat{E}^{(l)}\right)$

Establecer $B_{\alpha_G}^{Gr} = B_{\alpha_G}^{G(l)}$, $B_{\alpha_A}^{Ar} = B_{\alpha_A}^{A(l)}$, $\hat{E}^r = \hat{E}^{(l)}$ y $\hat{\Theta}^r = \hat{\Theta}^{(l)}$

Fin mientras

Si $\Psi_r < \Psi_{r-1}$ **entonces**

Establecer $B_{\alpha_G}^{Gopt} = B_{\alpha_G}^{Gr}$, $B_{\alpha_A}^{Aopt} = B_{\alpha_A}^{Ar}$, $\hat{E}^{opt} = \hat{E}^r$ y $\hat{\Theta}^{opt} = \hat{\Theta}^r$

Fin si

Fin mientras

Output: $B_{\alpha_G}^{G} = B_{\alpha_G}^{Gopt}$: asignación de los $\lceil (1-\alpha_G)\cdot n_G \rceil$ genes no recortados a los K_G grupos

Output: $\left\{B_{k,\alpha_A}^{A}\right\}_{1\leq k\leq K_G} = \left\{B_{k,\alpha_A}^{Aopt}\right\}_{1\leq k\leq K_G}$: conjunto de $\lceil (1-\alpha_A)\cdot n_A \rceil$ arrays no recortados en cada grupo

Output: $\hat{E} = \left\{\hat{E}_{\mathrm{g}}^{opt}\right\}_{1\leq g\leq n_G}$: parámetros de la estandarización estimados

Output: $\hat{\Theta} = \left\{\hat{\Theta}_k^{opt}\right\}_{1\leq k\leq K_G}$: parámetros del clúster k estimados

Figura 6.1. Algoritmo para encontrar K_G clústers de genes altamente correlados.

6.1.1.3.2. Iteraciones

En la iteración l, se tienen dos pasos: el paso E, dividido a su vez en tres sub-pasos, y el paso M.

$E.l$. Se distinguen tres sub-pasos,

$E.l.a.$ Re-definir la clasificación de genes y el recorte en esta dimensión.

Para cada gen g, $g=1,...,n_G$, y cada grupo de genes k, $k=1,...,K_G$, se calcula,

$$\psi_{g,k}^{(a)}\left(y_{ga};B_{k,\alpha_G}^{G(l-1)},B_{k,\alpha_A}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)= \\ =\frac{1}{2}\sum_{a=1}^{n_A} I_{B_{k,\alpha_A}^{A(l-1)}}(a)\left(\left(\frac{\frac{y_{ga}-\hat{m}_g^{(l-1)}}{\hat{s}_g^{(l-1)}}-\hat{\mu}_{a,k}^{(l-1)}}{\hat{\sigma}_{a,k}^{(l-1)}}\right)^2+\log\left(\hat{\sigma}_{a,k}^{(l-1)2}\right)\right)-\log\left(\hat{\pi}_k^{(l-1)}\right) \tag{6.9}$$

El gen g pertenecerá al grupo $k^{(l)}$, si se verifica que,

$$k^{(l)}=\underset{1\le m\le K_G}{\arg\min}\left(\psi_{g,m}^{(a)}\left(y_{ga};B_{k,\alpha_G}^{G(l-1)},B_{k,\alpha_A}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)\right) \tag{6.10}$$

Además, sea $\psi_{(g)}^{(a)}$ el estadístico ordenado de la muestra de valores $\psi_{g,k^{(l)}}^{(a)}\left(y_{ga};B_{k,\alpha_G}^{G(l-1)},B_{k,\alpha_G}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)$, tal que $\psi_{(1)}^{(a)}\le\psi_{(2)}^{(a)}\le...\le\psi_{(n_G)}^{(a)}$, el gen g no será recortado si se verifica que,

$$\psi_{g,k^{(l)}}^{(a)}\left(y_{ga};B_{k,\alpha_G}^{G(l-1)},B_{k,\alpha_A}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)\le\psi_{(\lceil(1-\alpha_G)\cdot n_G\rceil)}^{(a)} \tag{6.11}$$

El gen g pertenecerá al conjunto $B_{k,\alpha_G}^{G(l)}$, si se verifican las condiciones (6.10) y (6.11) simultáneamente.

$E.l.b.$ Para cada grupo de genes, determinar el conjunto condiciones no recortadas. El array a, $a=1,...,n_A$, pertenecerá al conjunto $B_{k,\alpha_A}^{A(l)}$ si,

$$\psi_{a,k}^{(b)}\left(y_{ga};B_{k,\alpha_G}^{G(l)},B_{k,\alpha_A}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}^{(l-1)}\right)\le\psi_{(\lceil(1-\alpha_{Ak})\cdot n_A\rceil),k}^{(b)} \tag{6.12}$$

donde,

$$\psi_{a,k}^{(b)}\left(y_{ga};B_{k,\alpha_G}^{G(l)},B_{k,\alpha_A}^{A(l-1)},\hat{\Theta}_k^{(l-1)},\hat{E}^{(l-1)}\right)= \\ =\frac{1}{2}\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot\left(\left(\frac{\frac{y_{ga}-\hat{m}_g^{(l-1)}}{\hat{s}_g^{(l-1)}}-\hat{\mu}_{a,k}^{(l-1)}}{\hat{\sigma}_{a,k}^{(l-1)}}\right)^2+\log\left(\hat{\sigma}_{a,k}^{(l-1)2}\right)\right) \tag{6.13}$$

$$-\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot\log\left(\hat{\pi}_k^{(l-1)}\right)$$

y $\psi_{(1)}^{(b)} \leq \psi_{(2)}^{(b)} \leq \ldots \leq \psi_{(n_A)}^{(b)}$.

$E.l.c.$ A partir de los conjuntos $B_{\alpha_G}^{G(l)} = \left\{ B_{k,\alpha_G}^{G(l)} \right\}_{1 \leq k \leq K_G}$ y $B_{\alpha_A}^{A(l)} = \left\{ B_{k,\alpha_A}^{A(l)} \right\}_{1 \leq k \leq K_G}$, re-estimar los nuevos parámetros para la estandarización $\hat{E}^{(l)} = \left\{ \hat{E}_g^{(l)} \right\}_{1 \leq g \leq n_G}$ según (6.4) y (6.5).

$M.l.$ A partir de los conjuntos $B_{\alpha_G}^{G(l)}$, $B_{\alpha_A}^{A(l)}$ y $\hat{E}^{(l)}$, re-estimar los nuevos parámetros, $\hat{\Theta}^{(l)} = \left\{ \hat{\Theta}_k^{(l)} \right\}_{1 \leq k \leq K_G}$ utilizando las expresiones (6.6), (6.7) y (6.8). Si el valor de alguna de las varianzas no excede el umbral inferior propuesto, c_R, se elige como valor de la correspondiente estimación el umbral c_R.

La figura 6.1 contiene un esquema del algoritmo propuesto.

6.1.2. Búsqueda de Biclústers: $K_G x K_A$

Pretendemos encontrar en la matriz de expresión grupos de genes que compartan agrupaciones de arrays en el sentido de co-expresión. En este caso, la agrupación buscada clasificará cada gen a uno de los grupos o a ser recortado, y dentro de cada grupo de genes clasificará a los arrays a una de las agrupaciones posibles o a permanecer sin clasificar. En este caso, también suponemos que los valores de expresión observados para los genes son una transformación lineal de su escala original, en la que valores de expresión similares corresponderán a co-expresión.

Partimos de la hipótesis de que la matriz de expresión está dividida en K_G grupos de genes y de que, en cada uno de ellos, existen K_A grupos de arrays no establecidos *a priori*. Asumimos que una proporción $\alpha_G \in [0,1]$ de genes no pertenece a ninguno de los K_G clústers, y que una proporción de arrays, $\alpha_A \in [0,1]$, no pertenece a ningún biclúster, no necesariamente afectando a los mismos arrays en cada uno de ellos. El número de clústers de genes, K_G, el número de grupos de arrays en cada clúster de genes, K_A, y el nivel de recorte para las dos dimensiones, α_G y α_A, se fijan *a priori*.

De esta forma, el modelo asumido para los datos de expresión corresponde a que genes en el mismo grupo comparten el mismo modelo de mezcla de normales generador de las expresiones observadas en los arrays. Para cada grupo de genes, las varianzas de las poblaciones normales en la mezcla se suponen iguales, aunque esta varianza común podría

diferir de un grupo de genes a otro. Para este modelo, como en el procedimiento de agrupación de genes anteriormente presentado, las varianzas deberían estar acotadas inferiormente o su tamaño relativo acotado. Se incluye un parámetro peso para cada grupo de genes y, adicionalmente, dentro de cada grupo de genes, se incluyen pesos para cada grupo de arrays. Suponemos que los valores observados para los genes corresponden a una transformación de localización y escala de una escala original estandarizada.

Al igual que en el procedimiento de *clustering* propuesto anteriormente, este análisis biclúster podría aplicarse a matrices de expresión pre-procesadas en las que se hubieran eliminado genes cuyo núcleo de expresión muestren poca variabilidad. De la misma forma que en el caso anterior, es necesario aplicar restricciones a los parámetros varianza que aparecen en el modelo. Escogemos para esta aplicación, como en el procedimiento anterior, acotar inferiormente la variabilidad.

6.1.2.1. Parámetros a estimar

Los parámetros del modelo serán,

(i) Característicos del clúster de genes y de cada biclúster de genes y arrays:

- la proporción de genes de cada clúster de genes, π_k. Estos pesos verifican que $0 \le \pi_k \le 1$ y $\sum_{k=1}^{K_G} \pi_k = 1$.
- nivel medio de expresión, μ_{kh} con $1 \le k \le K_G$ y $1 \le h \le K_A$.
- desviación típica común para todos los clúster de arrays dentro del clúster de genes k, σ_k con $1 \le k \le K_G$. Asumimos varianza común en todos los grupos de arrays del mismo clúster de genes.
- la proporción de arrays en cada biclúster, π_{kh}, $0 \le \pi_{kh} \le 1$ y $\sum_{h=1}^{K_A} \pi_{kh} = 1$ $\forall\ 1 \le k \le K_G$

Se denota por $\Theta_k = \left\{\pi_k, \{\pi_{kh}\}_{1 \le h \le K_A}, \{\mu_{kh}\}_{1 \le h \le K_A}, \sigma_k\right\}$ al conjunto de parámetros del clúster de genes k y por $\Theta = \{\Theta_k\}_{1 \le k \le K_G}$ al conjunto de parámetros de toda la clasificación.

(ii) Característicos de cada gen: patrón de expresión medio de cada gen y su desviación típica, m_g y s_g, $1 \le g \le n_G$. Estos dos parámetros se utilizan para estandarizar la expresión de cada gen, y se denota por $E_g = \{m_g, s_g\}$ al conjunto de parámetros del gen g, y por $E = \{E_g\}_{1 \le g \le n_G}$ a los correspondientes a todos los genes.

(iii) Conjuntos de genes y arrays activos, no recortados. Se denota por B^G_{k,α_G} al conjunto de genes asignado al clúster de genes k, y B^A_{kh,α_A} representa el conjunto de arrays no recortados correspondiente al biclúster kh. $B^G_{\alpha_G} = \left\{B^G_{k,\alpha_G}\right\}_{1\le k\le K_G}$ y $B^A_{\alpha_A} = \left\{B^A_{kh,\alpha_A}\right\}_{1\le k\le K_G, 1\le h\le K_A}$ representan respectivamente los conjuntos de genes y arrays no recortados en la clasificación.

6.1.2.2. Función objetivo

Dada una matriz de expresión $Y = \left\{y_{ga}\right\}_{1\le g\le n_G,\ 1\le a\le n_A}$, fijado un nivel de recorte $\alpha_G \in [0,1]$ en la dimensión de genes y un nivel $\alpha_A \in [0,\ 1]$ en la dimensión de arrays, la función objetivo viene dada por la expresión,

$$\Gamma\left(Y; B^G_{\alpha_G}, B^A_{\alpha_A}, \Theta, E\right) =$$

$$= \frac{1}{2\cdot n_G \cdot n_A} \sum_{k=1}^{K_G}\sum_{h=1}^{K_A}\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B^G_{k,\alpha_G}}(g)\, I_{B^A_{kh,\alpha_A}}(a)\left(\left(\frac{\frac{y_{gj}-m_g}{s_g}-\mu_{kh}}{\sigma_k}\right)^2 + \log\left(\sigma_k^2\right)\right) \qquad (6.14)$$

$$-\frac{1}{n_G\cdot n_A}\sum_{k=1}^{K_G}\sum_{g=1}^{n_G} I_{B^G_{k,\alpha_G}}(g)\cdot\log(\pi_k) - \frac{1}{n_G\cdot n_A}\sum_{k=1}^{K_G}\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B^A_{kh,\alpha_A}}(a)\cdot\log(\pi_{kh})$$

con las restricciones,

$$R = \left\{\sigma_k \ge c_R,\ 1\le k\le K_G\right\} \qquad (6.15)$$

y S definida en (6.3).

Como en el procedimiento anterior, suponiendo conocidos los parámetros que definen las transformaciones de localización y escala correspondientes a cada gen, se pretende conocer los valores de los parámetros que minimizan la función anterior. Como allí, el algoritmo, bajo esta suposición, al menos no empeorará a cada paso, y como allí, como estas transformaciones son desconocidas, se incluye su estimación en las iteraciones del algoritmo. El algoritmo propuesto será de tipo EM, obteniendo, en los pasos E, dados los valores actuales de los parámetros Θ, los valores óptimos para las agrupaciones, el recorte y la estandarización, y en los pasos M, se obtienen los óptimos para los parámetros Θ, en el espacio restringido, dadas las asignaciones y los parámetros de la estandarización.

6.1.2.3. Algoritmo

6.1.2.3.1. Solución inicial

Se propone inicializar el algoritmo mediante la asignación de un gen al azar a cada grupo de genes, y la aplicación de las k-medias recortadas **[Cuesta-Albertos et al, 1997]** a cada uno de

estos genes, para determinar los grupos de arrays dentro de cada gen. La doble búsqueda de genes y arrays incrementa mucho la complejidad del problema de agrupamiento en relación con lo que sería la búsqueda en una sola de las dimensiones. Con el tipo de inicialización propuesto se pretende disminuir esta complejidad, inherente al biclúster, favoreciendo la aparición de soluciones iniciales de mayor calidad. El procedimiento se resume en la figura 6.2.

En cada comienzo del procedimiento de *biclustering*, se eligen, de manera aleatoria, K_G genes, representantes de los K_G grupos. Sea $\{g_k\}_{1\le k\le K_G}$ este conjunto de genes inicial. Del procedimiento de *clustering* en la dimensión de los arrays (figura 6.2), se obtiene la asignación de los arrays para cada uno de los genes seleccionados, $B_{\alpha_A}^{A(0)} = \left\{B_{kh,\alpha_A}^{A(0)}\right\}_{1\le k\le K_G, 1\le h\le K_A}$ con

$$B_{kh,\alpha_A}^{A(0)} = \left\{a : 1 \le a \le n_A \ con\ B_{kh,\alpha_A}^{A(0)}(a) = B_h(a)\cdot I_{\tilde{B}_{h,\alpha_A}}(a)\right\} \tag{6.16}$$

y las siguientes estimaciones iniciales,

$$\hat{\mu}_{kh}^{(0)} = \frac{\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)\cdot y_{g_k a}}{\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)} \tag{6.17}$$

$$\hat{\sigma}_k^{(0)} = \sqrt{\frac{\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)\cdot\left(y_{g_k a} - \hat{\mu}_{kh}^{(0)}\right)^2}{\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)}} \tag{6.18}$$

$$\hat{\pi}_{kh}^{(0)} = \frac{\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)}{\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(0)}}(a)} \tag{6.19}$$

Inicialmente se asume que todos los clústers de genes son del mismo tamaño, es decir que,

$$\hat{\pi}_k^{(0)} = \frac{1}{K_G} \tag{6.20}$$

Además se estiman los parámetros de estandarización a partir de las expresiones,

$$\hat{m}_g^{(0)} = \frac{\sum_{a=1}^{n_A} I_{B_{kk_a^{\max},\alpha_A}^{A(0)}}(a)\cdot y_{ga}}{\sum_{a=1}^{n_A} I_{B_{kk_a^{\max},\alpha_A}^{A(0)}}(a)} \tag{6.21}$$

$$\hat{s}_g^{(0)} = \sqrt{\frac{\sum_{a=1}^{n_A} I_{B^{A(0)}_{kk_a^{\max},\alpha_A}}(a)\cdot\left(y_{ga}-\hat{m}_g^{(0)}\right)^2}{\sum_{a=1}^{n_A} I_{B^{A(0)}_{kk_a^{\max},\alpha_A}}(a)}} \tag{6.22}$$

donde,

$$k_a^{\max} = \arg\max_{1\le h\le K_A}\left\{\hat{\pi}_{kh}^{(0)}\right\} \tag{6.23}$$

es decir, que se estandariza utilizando el clúster de arrays más numeroso.

Se denota por $\hat{\Theta}_k^{(0)} = \left\{\hat{\pi}_k^{(0)}, \left\{\hat{\pi}_{kh}^{(0)}\right\}_{1\le h\le K_A}, \left\{\hat{\mu}_{kh}^{(0)}\right\}_{1\le h\le K_A}, \hat{\sigma}_k^{(0)}\right\}$ al conjunto de estimaciones iniciales en el clúster de genes k y $\hat{E}_g^{(0)} = \left\{\hat{m}_g^{(0)}, \hat{s}_g^{(0)}\right\}$ a las estimaciones iniciales de los parámetros de estandarización.

6.1.2.3.2. Iteraciones

En la iteración l, se tienen dos pasos: el paso E, dividido a su vez en tres sub-pasos, y el paso M.

$E.l$. Se distinguen tres sub-pasos,

$E.l.a.$ Re-definir la clasificación de genes y el recorte en esta dimensión.

Para cada gen g, $g = 1,\ldots,n_G$, y cada grupo de genes k, $k = 1,\ldots,K_G$, se calcula,

$$\begin{aligned}
&\upsilon_{g,k}^{(a)}\left(y_{ga};\left\{B_{kh,\alpha_a}^{A(l-1)}\right\}_{1\le h\le K_A},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right) = \\
&= \frac{1}{2}\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(l-1)}}(a)\cdot\left(\left(\frac{\dfrac{y_{ga}-\hat{m}_g^{(l-1)}}{\hat{s}_g^{(l-1)}}-\hat{\mu}_{kh}^{(l-1)}}{\hat{\sigma}_k^{(l-1)}}\right)^2 + \log\left(\hat{\sigma}_k^{(l-1)2}\right)\right) \\
&\quad -\sum_{h=1}^{K_A}\sum_{a=1}^{n_A} I_{B_{kh,\alpha_A}^{A(l-1)}}(a)\cdot\log\left(\hat{\pi}_{kh}^{(l-1)}\right) - \log\left(\hat{\pi}_k^{(l-1)}\right)
\end{aligned} \tag{6.24}$$

El gen g pertenecerá al grupo $k_g^{(l)}$, si se verifica que,

$$k_g^{(l)} = \arg\min_{1\le m\le K_G}\left(\upsilon_{g,m}^{(a)}\left(y_{ga};\left\{B_{kh,\alpha_A}^{A(l-1)}\right\}_{1\le h\le K_A},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)\right) \tag{6.25}$$

Además, sea $\upsilon_{(g)}^{(a)}$ el estadístico ordenado de la muestra de valores $\upsilon_{g,k_g^{(l)}}^{(a)}\left(y_{ga};\left\{B_{kh,\alpha_A}^{A(l-1)}\right\}_{1\le h\le K_A},\hat{\Theta}_k^{(l-1)},\hat{E}_g^{(l-1)}\right)$, tal que $\upsilon_{(1)}^{(a)} \le \upsilon_{(2)}^{(a)} \le \ldots \le \upsilon_{(n_G)}^{(a)}$, el gen g no será recortado si se verifica que,

$$\upsilon^{(a)}_{g,k_g^{(l)}}\left(y_{ga};\left\{B^{A(l-1)}_{kh,\alpha_A}\right\}_{1\le h\le K_A},\hat{\Theta}^{(l-1)}_k,\hat{E}^{(l-1)}_g\right)\le \upsilon^{(a)}_{\left(\lceil(1-\alpha_G)\cdot n_G\rceil\right)} \tag{6.26}$$

El gen g pertenecerá al conjunto $B^{G(l)}_{k_g^{(l)},\alpha_G}$, si se verifican las condiciones (6.25) y (6.26) simultáneamente.

E.l.b. Para cada grupo de genes, $1\le k\le K_G$, re-definir la clasificación de arrays y el recorte en esta dimensión.

Para cada array a, $a=1,...,n_A$, y cada grupo de arrays h, $h=1,...,K_A$, en el grupo de genes k, $k=1,...,K_G$, se calcula,

$$\upsilon^{(b)}_{a,h}\left(y_{ga};B^{G(l)}_{k,\alpha_G},\hat{\mu}^{(l-1)}_{kh},\hat{\pi}^{(l-1)}_{kh},\hat{\sigma}^{(l-1)}_k,\hat{E}^{(l-1)}\right)=$$

$$=\frac{1}{2}\sum_{g=1}^{n_G} I_{B^{G(l)}_{k,\alpha_G}}(g)\cdot\left(\left(\frac{\frac{y_{ga}-\hat{m}^{(l-1)}_g}{\hat{s}^{(l-1)}_g}-\hat{\mu}^{(l-1)}_{kh}}{\hat{\sigma}^{(l-1)}_k}\right)^2+\log\left(\hat{\sigma}^{(l-1)2}_k\right)\right)-\log\left(\hat{\pi}^{(l-1)}_{kh}\right) \tag{6.27}$$

El array a pertenecerá al grupo $k_a^{(l)}$, si se verifica que,

$$k_a^{(l)}=\arg\min_{1\le m\le K_A}\left(\upsilon^{(b)}_{a,m}\left(y_{ga};B^{G(l)}_{k,\alpha_G},\hat{\mu}^{(l-1)}_{kh},\hat{\pi}^{(l-1)}_{kh},\hat{\sigma}^{(l-1)}_k,\hat{E}^{(l-1)}\right)\right) \tag{6.28}$$

Además, sea $\upsilon^{(b)}_{(a)}$ el estadístico ordenado de la muestra de valores $\upsilon^{(b)}_{a,k_a^{(l)}}\left(y_{ga};B^{G(l)}_{k,\alpha_G},\hat{\mu}^{(l-1)}_{kh},\hat{\pi}^{(l-1)}_{kh},\hat{\sigma}^{(l-1)}_k,\hat{E}^{(l-1)}\right)$, tal que $\upsilon^{(b)}_{(1)}\le\upsilon^{(b)}_{(2)}\le...\le\upsilon^{(b)}_{(n_G)}$, el array a no será recortado si se verifica que,

$$\upsilon^{(b)}_{a,k_a^{(l)}}\left(y_{ga};B^{G(l)}_{k,\alpha_G},\hat{\mu}^{(l-1)}_{kh},\hat{\pi}^{(l-1)}_{kh},\hat{\sigma}^{(l-1)}_k,\hat{E}^{(l-1)}\right)\le\upsilon^{(b)}_{\left(\lceil(1-\alpha_A)\cdot n_A\rceil\right)} \tag{6.29}$$

El array a pertenecerá al conjunto $B^{A(l)}_{k_a^{(l)}h,\alpha_A}$, si se verifican las condiciones (6.28) y (6.29) simultáneamente.

E.l.c. A partir de los conjuntos $B^{G(l)}_{\alpha_G}=\left\{B^{G(l)}_{k,\alpha_G}\right\}_{1\le k\le K_G}$ y $B^{A(l)}_{\alpha_A}=\left\{B^{A(l)}_{kh,\alpha_A}\right\}_{1\le k\le K_G,1\le h\le K_A}$, re-estimar los nuevos parámetros para la estandarización $\hat{E}^{(l)}=\left\{\hat{\mathrm{E}}^{(l)}_{\mathrm{g}}\right\}_{1\le g\le n_G}$ utilizando las expresiones (6.21) y (6.22).

Input: Y_g : valores de expresión del gen g con n_A condiciones

Input: K_A : número de grupos

Input: α_A : nivel de recorte en la dimensión de arrays

Input: nr : número de comienzos aleatorios

Input: nl : número máximo de iteraciones EM

Mientras $r \leq nr$ **hacer**

Extraer muestras aleatorias de 2 arrays para cada uno de los K_A grupos

Parámetros iniciales: media y desviación típica de esas 2 observaciones por grupo. Inicialmente se asume que $\pi_k^{(0)} = 1/k$. Sea $\hat{\theta}^{(0)} = \left\{\hat{\theta}_k^{(0)}\right\}_{1\leq k\leq K_A}$ con $\hat{\theta}_k^{(0)} = \left\{\hat{\pi}_k^{(0)},\ \hat{\mu}_k^{(0)}, \hat{\sigma}_k^{(0)}\right\}$ el conjunto inicial.

Mientras $l \leq nl$ o $parada$ **hacer**

E.I: Calcular $d_{a,k}^{(l)}\left(y_{ga}\right) = \frac{1}{2}\left(\left(\frac{y_{ga} - \hat{\mu}_k^{(l-1)}}{\hat{\sigma}_k^{(l-1)}}\right)^2 + \log\left(\hat{\sigma}_k^{(l-1)2}\right)\right) - \log\left(\hat{\pi}_k\right),\ 1 \leq k \leq K_A$

Asignar el array j al grupo $k^{(l)} = \arg\min_{1\leq k\leq K_A}\left(d_{a,k}^{(l)}\left(y_{ga}\right)\right)$. Se denota por $B_{K_A}^{(l)}$ a la asignación de cada una de las observaciones al grupo con $d_{a,k}^{(l)}\left(y_{ga}\right)$ mínima y por $d_a^{(l)}$ a dicho mínimo.

Sea $d_{(a)}^{(l)}$ el estadístico ordenado de la muestra de valores $d_a^{(l)}\left(y_{ga}\right)$, tal que $d_{(1)}^{(l)} \leq d_{(2)}^{(l)} \leq \ldots \leq d_{(n_A)}^{(l)}$, construir $\tilde{B}_{\alpha_A}^{(l)}$ como el conjunto de arrays que verifican $d_a^{(l)}\left(y_{ga}\right) \leq d_{(\lceil(1-\alpha_A)\cdot n_A\rceil)}^{(l)}$

M.I: Estimar $\hat{\theta}^{(l)} = \left\{\hat{\theta}_k^{(l)}\right\}_{1\leq k\leq K_A}$ con $\hat{\theta}_k^{(l)} = \left\{\hat{\pi}_k^{(l)},\ \hat{\mu}_k^{(l)}, \hat{\sigma}_k^{(l)}\right\}$ como la proporción, la media y la desviación típica de las observaciones no recortadas asignadas al clúster $k^{(l)}$

Función objetivo:

$$D\left(Y_g; B_{K_A}^{(l)}, \tilde{B}_{\alpha_A}^{(l)}, \hat{\theta}^{(l)}\right) = \frac{1}{2n_A}\sum_{k=1}^{K_A}\sum_{a=1}^{n_A} I_{\tilde{B}_{k,\alpha_A}^{(l)}}(a)\left(\left(\frac{y_{ga} - \hat{\mu}_k^{(l)}}{\hat{\sigma}_k^{(l)}}\right)^2 + \log\left(\hat{\sigma}_k^{(l)2}\right)\right) - \frac{1}{n_A}\sum_{k=1}^{K_A}\log\left(\hat{\pi}_k^{(l)}\right)$$

Si $D\left(Y_g; B_{K_A}^{(l-1)}, \tilde{B}_{\alpha_A}^{(l-1)}, \hat{\theta}^{(l-1)}\right) = D\left(Y_g; B_{K_A}^{(l)}, \tilde{B}_{\alpha_A}^{(l)}, \hat{\theta}^{(l)}\right)$ **entonces**

$parada = Verdadero$

Fin si

Calcular $D_r = D\left(Y_g; B_{K_A}^{(l)}, \tilde{B}_{\alpha_A}^{(l)}, \hat{\theta}^{(l)}\right)$

Establecer $B_{K_A}^r = B_{K_A}^{(l)}$, $\tilde{B}_{\alpha_A}^r = \tilde{B}_{\alpha_A}^{(l)}$ y $\hat{\theta}^r = \hat{\theta}^{(l)}$

Fin mientras

Si $D_r < D_{r-1}$ **entonces**

Establecer $B_{K_A}^{opt} = B_{K_A}^r$, $\tilde{B}_{\alpha_A}^{opt} = \tilde{B}_{\alpha_A}^r$ y $\hat{\theta}^{opt} = \hat{\theta}^r$

Fin si

Fin mientras

Output: $B_{K_A} = B_{K_A}^r$: asignación de los n_A arrays a uno de los K_A grupos

Output: $\tilde{B}_{\alpha_A}^{opt} = \tilde{B}_{\alpha_A}^r$: $\lceil(1-\alpha_A)\cdot n_A\rceil$ arrays no recortados

Output: $\hat{\theta} = \left\{\hat{\theta}_k^{opt}\right\}_{1\leq k\leq K_A}$: parámetros del clúster estimados

Figura 6.2. Algoritmo para encontrar K_A clústers de arrays en cada gen.

$M.l$. A partir de los conjuntos $B_{\alpha_G}^{G(l)}$, $B_{\alpha_A}^{A(l)}$ y $\hat{E}^{(l)}$, re-estimar los nuevos parámetros, $\hat{\Theta}^{(l)} = \left\{\hat{\Theta}_k^{(l)}\right\}_{1\le k\le K_G}$ utilizando (6.19) para estimar $\hat{\pi}_{kh}^{(l)}$ y el resto de parámetros según las expresiones,

$$\hat{\pi}_k^{(l)} = \frac{\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(l)}}(g)}{\sum_{k=1}^{K_G}\sum_{g=1}^{n_G} I_{B_{k,\alpha_G}^{G(l)}}(g)} \tag{6.30}$$

$$\hat{\mu}_{kh}^{(l)} = \frac{\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot I_{B_{kh,\alpha_a}^{A(l)}}(a)\cdot\left(\frac{y_{ga}-\hat{m}_g^{(l)}}{\hat{s}_g^{(l)}}\right)}{\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot I_{B_{kh,\alpha_A}^{A(l)}}(a)} \tag{6.31}$$

$$\hat{\sigma}_k^{(l)} = \sqrt{\frac{\sum_{h=1}^{K_A}\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot I_{B_{kh,\alpha_A}^{A(l)}}(a)\cdot\left(\frac{y_{ga}-\hat{m}_g^{(l)}}{\hat{s}_g^{(l)}}-\hat{\mu}_{kh}^{(l)}\right)^2}{\sum_{h=1}^{K_A}\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B_{k,\alpha_G}^{G(l)}}(g)\cdot I_{B_{kh,\alpha_A}^{A(l)}}(a)}} \tag{6.32}$$

En la figura 6.3 se muestra un resumen del algoritmo.

6.1.3. Búsqueda de patrones de atipicidad: $K_G x2$

En esta sección presentamos un procedimiento para identificar patrones de atipicidad en una matriz de expresión. En el capítulo anterior ofrecíamos una metodología para detectar genes con comportamientos atípicos. Como resultado de este tipo de procedimientos, aparecen largas listas de genes, costosas de evaluar desde un punto de vista biológico. Nuestra propuesta está relacionada con realizar un post-procesado de este tipo de listas buscando grupos de genes y de arrays que, conjuntamente, expliquen estos comportamientos atípicos observados. En una situación ideal, sería posible reducir estos listados a grupos de genes, con un comportamiento atípico en el mismo conjunto de arrays. De esta forma, analizando la relación entre los grupos de arrays identificados como atípicos con las variables clínicas disponibles se podría encontrar explicación a los comportamientos observados, no realizándolo gen a gen, sino para grupos de genes.

El punto de partida para este post-análisis de los datos es una matriz binaria, de la misma dimensión que la matriz de expresión, en la que los valores sospechosos de atipicidad, identificados por algún análisis previo aparecerán con un 1 y los no sospechosos con un 0. Solo se incluirían en este análisis, aquellos genes que hubieran aparecido como sospechosos de contener atípicos.

Input: Y : matriz de expresión con n_G genes y n_A arrays

Input: K_G : número de grupos en la dimensión de genes

Input: K_A : número de grupos en la dimensión de arrays

Input: α_G : nivel de recorte en la dimensión de genes

Input: α_A : nivel de recorte en la dimensión de arrays

Input: nr : número de comienzos aleatorios

Input: nl : número máximo de iteraciones EM

K_A -medias recortadas en cada gen según procedimiento de la figura 6.2

Mientras $r \le nr$ **hacer**

Extraer una muestra aleatoria de K_G genes, $\{g_k\}_{1\le k\le K_G}$

Establecer $B_{\alpha_A}^{A(0)} = \left\{B_{kh,\alpha_A}^{A(0)}\right\}_{1\le k\le K_G, 1\le h\le K_A}$ según (6.16)

Estimar $\hat{E}^{(0)} = \left\{\hat{E}_{\mathrm{g}}^{(0)}\right\}_{1\le g\le n_G}$ con $\hat{E}_g^{(0)} = \left\{\hat{m}_{\mathrm{g}}^{(0)}, \hat{s}_g^{(0)}\right\}$ según (6.21) y (6.22)

Estimar $\hat{\Theta}^{(0)} = \left\{\hat{\Theta}_k^{(0)}\right\}_{1\le k\le K_G}$ con $\hat{\Theta}_k^{(0)} = \left\{\hat{\pi}_k^{(0)}, \left\{\hat{\pi}_{kh}^{(0)}\right\}_{1\le h\le K_A}, \left\{\hat{\mu}_{kh}^{(0)}\right\}_{1\le h\le K_A}, \hat{\sigma}_k^{(0)}\right\}$ según (6.17) – (6.20)

Mientras $l \le nl$ o $parada$ **hacer**

E.l.a: Calcular $\upsilon_{g,k}^{(a)}\left(y_{ga}; \left\{B_{kh,\alpha_a}^{A(l-1)}\right\}_{1\le h\le K_A}, \hat{\Theta}_k^{(l-1)}, \hat{E}_g^{(l-1)}\right)$ según (6.24)

Asignar el gen g al grupo $k^{(l)}$ según (6.25)

Construir $B_{\alpha_G}^{G(l)}$ según (6.25) y (6.26)

E.l.b: Para $1 \le k \le K_G$, calcular $\upsilon_{a,h}^{(b)}\left(y_{ga}; B_{k,\alpha_G}^{G(l)}, \hat{\mu}_{kh}^{(l-1)}, \hat{\pi}_{kh}^{(l-1)}, \hat{\sigma}_k^{(l-1)}, \hat{E}^{(l-1)}\right)$ según (6.27)

Asignar el array a al grupo $k_a^{(l)}$ según (6.28)

Construir $B_{\alpha_A}^{A(l)}$ según (6.28) y (6.29)

E.l.c: Estimar $\hat{E}^{(l)} = \left\{\hat{E}_{\mathrm{g}}^{(l)}\right\}_{1\le g\le n_G}$ con $\hat{E}_g^{(l)} = \left\{\hat{m}_g^{(l)}, \hat{s}_g^{(l)}\right\}$ según (6.21) y (6.22)

M.l: Estimar $\hat{\Theta}^{(l)} = \left\{\hat{\Theta}_k^{(l)}\right\}_{1\le k\le K_G}$ con $\hat{\Theta}_k^{(l)} = \left\{\hat{\pi}_k^{(l)}, \left\{\hat{\pi}_{kh}^{(l)}\right\}_{1\le h\le K_A}, \left\{\hat{\mu}_{kh}^{(l)}\right\}_{1\le h\le K_A}, \hat{\sigma}_k^{(l)}\right\}$ según (6.19), (6.30) - (6.32)

Si $\Gamma\left(Y; B_{\alpha_G}^{G(l-1)}, B_{\alpha_A}^{A(l-1)}, \hat{\Theta}^{(l-1)}, \hat{E}^{(l-1)}\right) = \Gamma\left(Y; B_{\alpha_G}^{G(l)}, B_{\alpha_A}^{A(l)}, \hat{\Theta}^{(l)}, \hat{E}^{(l)}\right)$ **entonces**

$parada = Verdadero$

Fin si

Calcular $\Gamma_r = \Gamma\left(Y; B_{\alpha_G}^{G(l)}, B_{\alpha_A}^{A(l)}, \hat{\Theta}^{(l)}, \hat{E}^{(l)}\right)$

Establecer $B_{\alpha_G}^{Gr} = B_{\alpha_G}^{G(l)}$, $B_{\alpha_A}^{Ar} = B_{\alpha_A}^{A(l)}$, $\hat{\Theta}^r = \hat{\Theta}^{(l)}$ y $\hat{E}^r = \hat{E}^{(l)}$

Fin mientras

Si $\Gamma_r = \Gamma_{r-1}$ **entonces**

Establecer $B_{\alpha_G}^{Gopt} = B_{\alpha_G}^{Gr}$, $B_{\alpha_A}^{Aopt} = B_{\alpha_A}^{Ar}$, $\hat{\Theta}^{opt} = \hat{\Theta}^r$ y $\hat{E}^{opt} = \hat{E}^r$

Fin si

Fin mientras

Output: $B_{\alpha_G}^{G} = B_{\alpha_G}^{Gopt}$: asignación de los $\lceil (1-\alpha_G)\cdot n_G \rceil$ genes no recortados

Output: $\left\{B_{kh,\alpha_A}^{A}\right\}_{1\le k\le K_G} = \left\{B_{kh,\alpha_A}^{Aopt}\right\}_{1\le k\le K_G}$: asignación de las $\lceil (1-\alpha_A)\cdot n_A \rceil$ arrays no recortados en cada clúster de genes

Output: $\hat{E} = \left\{\hat{E}_{\mathrm{g}}^{opt}\right\}_{1\le g\le n_G}$: parámetros de la estandarización estimados

Output: $\hat{\Theta} = \left\{\hat{\Theta}_k^{opt}\right\}_{1\le k\le K_G}$: parámetros del clúster estimados

Figura 6.3. Algoritmo para encontrar $K_G x K_A$ biclústers.

6.1.3.1. Definición de *outlier*

Sea y_{ga} el nivel de expresión del gen $g = 1,...,n_G$ en el array $a = 1,...,n_A$, el array a será definido como *outlier* en el sentido de sobre-expresión para el gen g si se verifica que,

$$y_{ga} \geq \hat{\mu}_{g,\alpha-smart} + f \cdot \hat{\sigma}_{g,\alpha-smart} \tag{6.33}$$

donde $\hat{\mu}_{g,\alpha-smart}$ y $\hat{\sigma}_{g,\alpha-smart}$ son las estimaciones de localización y dispersión, respectivamente, del núcleo de expresión del gen g, obtenidos a partir del recorte imparcial de nivel $\alpha \in [0,1]$, y f es el umbral en puntuación tipificada que determina qué nivel de alejamiento de la media consideramos como atípico.

De forma totalmente análoga, el array a será definido como *outlier* en el sentido de Infra-expresión para el gen g si se verifica que,

$$y_{ga} \leq \hat{\mu}_{g,\alpha-smart} - f \cdot \hat{\sigma}_{g,\alpha-smart} \tag{6.34}$$

Estas definiciones de *outlier* identifican arrays posiblemente sobre o infra-expresados respecto de su núcleo de expresión.

6.1.3.2. Parámetros a estimar

En este caso, el número de grupos de arrays en cada clúster de genes será fijo, $K_A = 2$, mientras que el usuario deberá decidir el número de clústers de genes, K_G, y los niveles de recorte, α_G para genes y α_A para arrays.

Este problema de búsqueda de biclústers puede verse como una versión simplificada de la búsqueda de biclústers $K_G x K_A$. En este caso, no hay parámetros de estandarización, y suponemos conocidas las medias de los arrays, 0 y 1. El conjunto de parámetros viene dado por,

(i) para el biclúster kh, $1 \leq k \leq K_G$ y $1 \leq h \leq K_A$, se reduce a $\Theta_{kh} = \{\pi_k\}$ con $0 \leq \pi_k \leq 1$ y $\sum_{k=1}^{K_G} \pi_k = 1$,

(ii) $K_G x 2$ sub-matrices formadas por $\{\pi_k\}_{1 \leq k \leq K_G}$ genes pertenecientes al conjunto de $\lceil (1-\alpha_G) \cdot n_G \rceil$ genes no recortados y $\{\pi_{kh}\}_{1 \leq k \leq K_G, 1 \leq h \leq 2}$ arrays del conjunto de $\lceil (1-\alpha_A) \cdot n_A \rceil$ arrays no recortados en el grupo k de genes.

6.1.3.3. Función objetivo

Fijados los niveles de recorte $\alpha_G \in [0,1]$ para genes y $\alpha_A \in [0,1]$ para arrays, se pretende estimar los parámetros que minimizan la función objetivo,

$$\begin{aligned}&\Gamma^*\left(O; B^G_{\alpha_G}, B^A_{\alpha_A}, \pi\right) = \\ &= \frac{1}{2n_G n_a}\sum_{k=1}^{K_G}\sum_{h=1}^{2}\sum_{g=1}^{n_G}\sum_{a=1}^{n_A} I_{B^G_{k,\alpha_G}}(g)\, I_{B^A_{kh,\alpha_A}}(a)\left(o_{ga}-\mu_{kh}\right)^2 \\ &\quad -\frac{1}{n_G n_a}\sum_{k=1}^{K_G}\sum_{g=1}^{n_G} I_{B^G_{k,\alpha_G}}(g)\cdot\log(\pi_k)\end{aligned} \tag{6.35}$$

donde $o_{ga} = \begin{cases} 0 & si\ a\ outlier\ en\ el\ gen\ g \\ 1 & si\ a\ no-outlier\ en\ el\ gen\ g \end{cases}$, y $B^G_{\alpha_G}$ y $B^A_{\alpha_A}$ representan los conjuntos de asignación de genes y arrays, respectivamente.

6.1.3.4. Solución inicial

De manera aleatoria se asigna cada gen a uno de los K_G clústers de genes y se determina el conjunto de genes no recortados, como los $\lceil (1-\alpha_G)\cdot n_G \rceil$ con menor distancia al grupo asignado. Sea $B^{G(0)}_{\alpha_G} = \left\{B^{G(0)}_{k,\alpha_G}\right\}_{1\le k\le K_G}$ el conjunto de asignación de los genes no recortados, el peso de cada clúster se estima según la expresión,

$$\hat{\pi}^{(0)}_k = \frac{\sum_{g=1}^{n_G} I_{B^{G(0)}_{k,\alpha_G}}(g)}{\sum_{k=1}^{K_G}\sum_{g=1}^{n_G} I_{B^{G(0)}_{k,\alpha_G}}(g)} \tag{6.36}$$

Para cada grupo de genes, $1 \le k \le K_G$, se clasifican los arrays y se determina el recorte en esta dimensión. Para cada array a, $a = 1,\ldots,n_A$, y cada grupo de arrays h, $h = 1,2$, en el grupo de genes k se calcula,

$$\upsilon^{(b)*}_{a,h}\left(o_{ga}; B^{G(0)}_{\alpha_G}\right) = \frac{1}{2}\sum_{g=1}^{n_G} I_{B^{G(l)}_{k,\alpha_G}}(g)\cdot\left(o_{ga}-\mu_{kh}\right)^2 \tag{6.37}$$

El array a pertenecerá al grupo $k^{(0)}_a$, si se verifica que,

$$k^{(0)}_a = \underset{1\le m\le K_A}{\arg\min}\left(\upsilon^{(b)*}_{a,m}\left(o_{ga}; B^{G(0)}_{\alpha_G}\right)\right) \tag{6.38}$$

Además, sea $\upsilon^{(b)*}_{(a)}$ el estadístico ordenado de la muestra de valores $\upsilon^{(b)*}_{a,k^{(0)}_a}\left(o_{ga}; B^{G(0)}_{\alpha_G}\right)$, tal que $\upsilon^{(b)*}_{(1)} \le \upsilon^{(b)*}_{(2)} \le \ldots \le \upsilon^{(b)*}_{(n_G)}$, el array a no será recortado si se verifica que,

$$\upsilon^{(b)*}_{a,k_g^{(0)}}\left(o_{ga};B^{G(0)}_{k,\alpha_G}\right)\leq \upsilon^{(b)*}_{\left(\lceil(1-\alpha_A)\cdot n_A\rceil\right)} \tag{6.39}$$

El array a pertenecerá al conjunto $B^{A(0)}_{k_g^{(0)}h,\alpha_A}$, si se verifican las condiciones (6.38) y (6.39) simultáneamente.

6.1.3.5. Iteraciones

En la iteración l,

$E.l$. Se distinguen dos sub-pasos,

$E.l.a.$ Re-definir la clasificación de genes y el recorte en esta dimensión.

Para cada gen g, $g=1,...,G$, y cada grupo de genes k, $k=1,...,K_G$, se calcula,

$$\upsilon^{(a)*}_{g,k}\left(o_{ga};\left\{B^{A(l-1)}_{kh,\alpha_A}\right\}_{h=1,2},\hat{\pi}^{(l-1)}_k\right)=\frac{1}{2}\sum_{h=1}^{2}\sum_{a=1}^{n_A} I_{B^{A(l-1)}_{kh,\alpha_A}}(a)\cdot\left(o_{ga}-\mu_{kh}\right)^2-\log\left(\hat{\pi}^{(l-1)}_k\right) \tag{6.40}$$

El gen g pertenecerá al grupo $k_g^{(l)}$, si se verifica que,

$$k_g^{(l)}=\underset{1\leq m\leq K_G}{\arg\min}\left(\upsilon^{(a)*}_{g,m}\left(o_{ga};\left\{B^{A(l-1)}_{kh,\alpha_A}\right\}_{h=1,2},\hat{\pi}^{(l-1)}_k\right)\right) \tag{6.41}$$

Además, sea $\upsilon^{(a)*}_{(g)}$ el estadístico ordenado de la muestra de valores $\upsilon^{(a)*}_{g,k_g^{(l)}}\left(o_{ga};\left\{B^{A(l-1)}_{kh,\alpha_A}\right\}_{h=1,2},\hat{\pi}^{(l-1)}_k\right)$, tal que $\upsilon^{(a)*}_{(1)}\leq\upsilon^{(a)*}_{(2)}\leq...\leq\upsilon^{(a)*}_{(n_G)}$, el gen g no será recortado si se verifica que,

$$\upsilon^{(a)*}_{g,k_g^{(l)}}\left(o_{ga};\left\{B^{A(l-1)}_{kh,\alpha_A}\right\}_{h=1,2},\hat{\pi}^{(l-1)}_k\right)\leq\upsilon^{(a)*}_{\left(\lceil(1-\alpha_G)\cdot n_G\rceil\right)} \tag{6.42}$$

El gen g pertenecerá al conjunto $B^{G(l)}_{k_g^{(l)},\alpha_G}$, si se verifican las condiciones (6.41) y (6.42) simultáneamente.

$E.l.b.$ Para cada grupo de genes, $1\leq k\leq K_G$, re-definir la clasificación de arrays y el recorte en esta dimensión, tal y como se ha detallado en el apartado 6.1.3.4. Para cada array a, $a=1,...,n_A$, y cada grupo de arrays h, $h=1,2$, en el grupo de genes k calcular $\upsilon^{(b)*}_{a,h}\left(o_{ga};B^{G(l)}_{\alpha_G}\right)$ según (6.37). El array a pertenecerá al conjunto $B^{A(l)}_{kh,\alpha_A}$ si se verifican las condiciones (6.38) y (6.39) simultáneamente.

$M.l$. A partir de los conjuntos $B^{G(l)}_{\alpha_G}$ y $B^{A(l)}_{\alpha_A}$, re-estimar $\left\{\hat{\pi}^{(l)}_k\right\}_{1\leq k\leq K_G}$ según la expresión (6.36).

6.2. Evaluación de los métodos propuestos

Evaluaremos los procedimientos propuestos utilizando datos obtenidos por simulación y datos reales.

En el caso de los datos simulados, cuando busquemos un único biclúster, se cuantificará el porcentaje de coincidencia entre el grupos real y el encontrado, en términos de sensibilidad (porcentaje de individuos del clúster verdadero encontrados por la solución) y de especificidad (porcentaje de individuos no en el clúster que son recortados). Estas medidas se pueden calcular tanto en la dimensión de genes como en la de arrays. En el caso de soluciones con K biclústers, para medir el nivel de acierto aplicaremos el *Gene Match Score (*GMS) de **[Prelic et al, 2006]**. Siendo M^{opt} la configuración de biclústers verdadera y M el resultado obtenido, la relevancia de la solución vendrá determinada por $S\left(M, M^{opt}\right)$, que refleja el grado en el que biclústers encontrados representan verdaderos grupos en la dimensión de genes. Por otra parte la cobertura de la solución será, $S\left(M^{opt}, M\right)$, que cuantifica como de bien cada uno de los biclústers reales es cubierto por los grupos encontrados. Ambas medidas tendrán un valor máximo de 1 cuando $M = M^{opt}$.

Para datos reales la forma más habitual de evaluar la bondad de las clasificaciones obtenidas está relacionada con el grado de coincidencia entre éstas y anotaciones funcionales del tipo de GO o KEGG.

6.2.1. Datos simulados

6.2.1.1. $K_G x 1$ clústers de co-expresión con $K_G = 1$

Simulamos datos de expresión génica en matrices con $n_G\left(=100\right)$ genes y $n_A\left(=50\right)$ arrays con un único grupo de genes co-expresados. Las medias $\mu = \left(\mu_1, ..., \mu_g, ..., \mu_{n_G}\right)$ y varianzas $\sigma^2 = \left(\sigma_1^2, ..., \sigma_g^2, ..., \sigma_{n_G}^2\right)$ de los genes se obtienen a partir de realizaciones de uniformes $U\left(4, 14\right)$ y de exponenciales $\exp\left(2\right)$, respectivamente, siendo todas ellas independientes. Suponemos que la correlación entre parejas de genes, ρ_{ij}, en el grupo de co-expresión se supone igual a un valor fijo ρ, y cuando uno de los genes en la pareja no esté en ese grupo, $\rho = 0$. Los datos de expresión de los genes se generan como n_A realizaciones de una normal n_G-dimensional de media μ y matriz de covarianzas Σ, cuyo elemento ij viene dado por $\sigma_i \sigma_j \rho$, $1 \le i, j \le n_G$. Los datos de expresión de los arrays que no contribuyen al patrón de co-expresión en genes del grupo se sustituyen por realizaciones, de normales con media y varianza la correspondiente a cada gen, independientes de todas las demás realizaciones. Para tener en cuenta la variabilidad del método, se simulan $n_{sim}\left(=50\right)$ conjuntos de datos.

En la tabla 6.1 se muestran los resultados obtenidos tras la aplicación del método a datos simulados correspondientes a diferentes elecciones para la correlación en el grupo de co-expresión $\rho = \{0, 0.3, 0.5, 0.8\}$ y para diferentes tamaños del clúster de genes. En estas aplicaciones se fijaron los niveles de recorte, α_G y α_A, en los correspondientes, respectivamente, a la contaminación incluida en el modelo de generación de los datos. En todos los casos se utilizan 50 comienzos aleatorios.

Para cada conjunto de simulaciones se muestran las medias y las desviaciones típicas de las 50 réplicas correspondientes a: (a) el coeficiente de correlación medio en el clúster estimado, $\overline{\rho}(\mathrm{Y}_{IJ})$, utilizando solamente los genes y los arrays no recortados; (b) el coeficiente de correlación medio en el clúster, considerando todos los arrays, $\overline{\rho}(\mathrm{Y}_{IA})$; (c) el coeficiente de correlación medio considerando todos los genes de la matriz y sólo los arrays no recortados, $\overline{\rho}(\mathrm{Y}_{GJ})$; (d) la sensibilidad en la identificación de genes, proporción de genes recortados entre los no pertenecientes al cluster, o equivalentemente, proporción de acierto en genes a eliminar; (e) la especificidad en la identificación de genes, o equivalentemente, proporción de acierto en el recorte de genes pertenecientes al cluster o, equivalentemente, acierto en genes a no eliminar; (f) y (g) los correspondientes valores de sensibilidad y especificidad para la dimensión de los arrays.

Tabla 6.1. $K_G x 1$ clústers de co-expresión con $K_G = 1$. Resultados variando el número de genes y arrays del clúster y la correlación entre genes.

IxJ	ρ	$\overline{\rho}(\mathrm{Y}_{IJ})$	$\overline{\rho}(\mathrm{Y}_{IA})$	$\overline{\rho}(\mathrm{Y}_{GJ})$	$\mathrm{Sens_g}$	$\mathrm{Esp_g}$	$\mathrm{Sens_a}$	$\mathrm{Esp_a}$
40x40	0	0.06 ± 0.01	0.042 ± 0.008	0.011 ± 0.003	0.404 ± 0.063	0.603 ± 0.042	0.8 ± 0.028	0.198 ± 0.112
	0.3	0.284 ± 0.042	0.232 ± 0.041	0.051 ± 0.009	0.948 ± 0.034	0.966 ± 0.023	0.897 ± 0.037	0.588 ± 0.149
	0.5	0.489 ± 0.062	0.393 ± 0.06	0.082 ± 0.01	0.996 ± 0.009	0.997 ± 0.006	0.972 ± 0.017	0.888 ± 0.069
	0.8	0.797 ± 0.034	0.637 ± 0.049	0.129 ± 0.01	1 ± 0	1 ± 0	1 ± 0.004	0.998 ± 0.014
40x30	0	0.104 ± 0.007	0.05 ± 0.006	0.013 ± 0.003	0.398 ± 0.061	0.599 ± 0.041	0.609 ± 0.063	0.413 ± 0.095
	0.3	0.289 ± 0.052	0.154 ± 0.037	0.051 ± 0.011	0.872 ± 0.077	0.915 ± 0.051	0.793 ± 0.073	0.69 ± 0.114
	0.5	0.513 ± 0.008	0.328 ± 0.052	0.092 ± 0.024	0.998 ± 0.008	0.998 ± 0.005	0.943 ± 0.022	0.915 ± 0.034
	0.8	0.8 ± 0.036	0.482 ± 0.052	0.123 ± 0.012	1 ± 0	1 ± 0	1 ± 0	1 ± 0
20x40	0	0.116 ± 0.018	0.079 ± 0.015	0.011 ± 0.003	0.193 ± 0.074	0.798 ± 0.018	0.798 ± 0.028	0.192 ± 0.114
	0.3	0.276 ± 0.052	0.217 ± 0.042	0.015 ± 0.004	0.831 ± 0.103	0.958 ± 0.026	0.856 ± 0.03	0.422 ± 0.122
	0.5	0.496 ± 0.058	0.401 ± 0.055	0.016 ± 0.004	0.992 ± 0.021	0.998 ± 0.005	0.937 ± 0.025	0.748 ± 0.101
	0.8	0.788 ± 0.048	0.633 ± 0.066	0.018 ± 0.006	1 ± 0	1 ± 0	1 ± 0	1 ± 0

Se observa en la tabla 6.1 que las correlaciones medias en el clúster estimado, utilizando solo los genes y arrays no recortados, se distribuyen en torno a las teóricas. Encontramos también proporciones de acierto superiores al 83% en la identificación de genes, cuando la correlación entre genes que co-expresan es de 0.3. Para correlaciones superiores a 0.5 estas

proporciones de acierto se sitúan superiores a 0.97. El tamaño del clúster, en los rangos estudiados, no parece tener demasiada influencia.

6.2.1.2. $K_G x 1$ clústers de co-expresión con $K_G > 1$

En este caso, se generan $K_G (=3)$ clústers en una matriz de $n_G (=150)$ genes y $n_A (=50)$ arrays. Para obtener los niveles de expresión en cada uno de los clústers de genes se siguió el mismo procedimiento que en el caso anterior. Las celdas de la tabla no incluidas en los clústers de genes o en los arrays correspondientes a cada uno de ellos, fueron generadas de forma independiente. Se generaron $n_{sim} (=50)$ réplicas para cada una de las situaciones.

Tabla 6.2. Simulaciones de $K_G x 1$ *clústers* de co-expresión con $K_G > 1$

	Simulación 1	Simulación 2	Simulación 3	Simulación 4
# genes, n_G	150	150	150	150
# condiciones, n_A	50	50	50	50
# biclústers, K_G	3	3	3	3
# genes, I	$\{40, 40, 40\}$	$\{40, 40, 40\}$	$\{40, 40, 40\}$	$\{40, 40, 40\}$
# arrays, J	$\{40, 40, 40\}$	$\{40, 40, 40\}$	$\{40, 40, 40\}$	$\{40, 40, 40\}$
correlación, ρ	$\{0.3, 0.3, 0.3\}$	$\{0.5, 0.5, 0.5\}$	$\{0.8, 0.8, 0.8\}$	$\{0.8, 0.3, 0.5\}$
# simulaciones, *nsim*	50	50	50	50
recorte genes, α_G	0.2	0.2	0.2	0.2
recorte arrays, α_A	0.2	0.2	0.2	0.2

Se consideran 4 patrones de simulación distintos en los que se modifican las correlaciones entre genes. En la tabla 6.2 se resumen las características de cada uno de ellos junto con los parámetros utilizados para la aplicación del método.

Tabla 6.3. $K_G x 1$ clústers de co-expresión con $K_G = 3$. Resultados variando la correlación entre genes del mismo grupo.

	# biclústers	$S(M, M^{opt})$	$S(M^{opt}, M)$
Simulación 1	2.98 ± 0.14	0.56 ± 0.2	0.61 ± 0.16
Simulación 2	3 ± 0	0.95 ± 0.03	0.95 ± 0.03
Simulación 3	3 ± 0	1 ± 0	1 ± 0
Simulación 4	3 ± 0	0.91 ± 0.03	0.91 ± 0.03

Para cada conjunto de simulaciones, en la tabla 6.3, se muestran medias y desviaciones típicas de las 50 simulaciones para el número de clústers encontrados y las dos variantes del GSM: $S\left(M, M^{opt}\right)$ para evaluar la relevancia de la solución y $S\left(M^{opt}, M\right)$ para la cobertura de la misma. Sólo en la simulación 1, caracterizada por tener todos los clústers con correlaciones bajas, se obtienen proporciones de acierto por debajo del 60%. Únicamente en este caso, en ocasiones, no se identifican los tres clústers. En el resto de escenarios los valores de los GSM son mayores a 0.9, y siempre aparecen los tres clústers de genes.

6.2.1.3. $K_G x K_A$ biclústers con $K_G > 1$ y $K_A > 1$

Se generan $K(=9)$ biclústers en una matriz de expresión de $n_G(=200)$ genes y $n_A(=96)$ arrays. El número de clústers de genes es $K_G(=3)$, todos ellos de tamaño $I(=50)$. En la dimensión de arrays, se generan $K_A(=3)$ clústers, de tamaños $J=\{24,\ 48,\ 24\}$, con valor de expresión medio 0, en el grupo más numeroso y expresión diferencial de ±4 en las otras dos clases. Todos los niveles de expresión se generan utilizando la distribución normal. Se prueban 3 patrones de simulación (tabla 6.4), variando el nivel de recorte en la dimensión array y la media y desviación típica por gen: en los dos primeros escenarios se consideran media 0 y varianza 1 para todos los genes, y en el tercero, se determinan de manera aleatoria. En cada caso se simulan $n_{sim}(=50)$ conjuntos de datos.

Tabla 6.4. Simulaciones de $K_G x K_A$ biclústers con $K_G > 1$ y $K_A > 1$

	Simulación 1	Simulación 2	Simulación 3
# genes, n_G	200	200	200
# arrays, n_A	96	96	96
# biclústers, K	9	9	9
# genes, I	$\{50,\ 50,\ 50\}$	$\{50,\ 50,\ 50\}$	$\{50,\ 50,\ 50\}$
# arrays, J_I	$\{24,\ 48,\ 24\}$	$\{23,\ 47,\ 23\}$	$\{23,\ 47,\ 23\}$
expresión biclúster, ∂_k	$[-4\ \ 0\ \ 4]$ $[-4\ \ 0\ \ -8]$ $[\ 4\ \ 0\ \ 8]$	$[-4\ \ 0\ \ 4]$ $[-4\ \ 0\ \ -8]$ $[\ 4\ \ 0\ \ 8]$	$[-4\ \ 0\ \ 4]$ $[-4\ \ 0\ \ -8]$ $[\ 4\ \ 0\ \ 8]$
media por gen, μ_g	0	0	$N(2,1)$
varianza por gen, σ_g^2	1	1	$\exp(2)$
# simulaciones, *nsim*	50	50	50
recorte genes, α_G	0.25	0.25	0.25
recorte arrays, α_A	0	0.03	0.03

En la tabla 6.5 se muestran los resultados. En los dos primeros escenarios para una mayoría de situaciones simuladas se encuentra la solución óptima. La simulación 3 es más compleja debido a la necesidad de estandarización y requiere muchos más comienzos aleatorios para llegar a la solución óptima. En este caso, el funcionamiento del método medido con las dos variantes del GSM, está por encima del 85%.

Tabla 6.5. $K_G x K_A$ biclústers con $K_G = K_A = 3$. Resultados variando el nivel de recorte y la media y desviación típica por gen.

	Simulación 1	Simulación 2	Simulación 3
# comienzos	20.86 ± 13.84	27.6 ± 26.57	268.36 ± 142.11
# biclústers	9 ± 0	9 ± 0	8.42 ± 0.57
$\hat{\pi}_k$	0.33 ± 0 0.33 ± 0 0.33 ± 0	0.33 ± 0 0.33 ± 0 0.33 ± 0	0.36 ± 0.05 0.31 ± 0.03 0.33 ± 0.04
$\hat{\pi}_{k1}$	0.25 ± 0 0.5 ± 0 0.25 ± 0	0.25 ± 0 0.5 ± 0.01 0.25 ± 0	0.29 ± 0.16 0.5 ± 0.16 0.27 ± 0.14
$\hat{\pi}_{k2}$	0.25 ± 0 0.5 ± 0 0.25 ± 0	0.25 ± 0 0.5 ± 0.01 0.25 ± 0	0.25 ± 0.05 0.49 ± 0.07 0.26 ± 0.08
$\hat{\pi}_{k3}$	0.25 ± 0 0.5 ± 0 0.25 ± 0	0.25 ± 0 0.5 ± 0.01 0.25 ± 0	0.27 ± 0.17 0.43 ± 0.2 0.24 ± 0.16
$\hat{\mu}_{k1}$	-4.05 ± 0.04 0 ± 0.01 4.07 ± 0.05	-4.07 ± 0.05 0 ± 0.01 4.06 ± 0.05	-1.53 ± 0.24 2.49 ± 0.26 -5.33 ± 1.1
$\hat{\mu}_{k2}$	-4.06 ± 0.04 0 ± 0.01 -8.11 ± 0.06	-4.06 ± 0.04 0 ± 0.01 -8.14 ± 0.06	-1.65 ± 0.91 1.84 ± 0.42 5.4 ± 1.43
$\hat{\mu}_{k3}$	4.06 ± 0.04 0 ± 0.01 8.11 ± 0.04	4.06 ± 0.05 0 ± 0.01 8.12 ± 0.07	5.54 ± 1.47 2.43 ± 1.75 8.75 ± 2.52
$\hat{\sigma}_k^2$	1.05 ± 0.01 1.11 ± 0.04 1.1 ± 0.02	1.07 ± 0.03 1.16 ± 0.04 1.15 ± 0.04	1.81 ± 0.32 1.41 ± 0.16 1.6 ± 0.2
$S(M, M^{opt})$	1 ± 0	1 ± 0	0.887 ± 0.075
$S(M^{opt}, M)$	1 ± 0	1 ± 0	0.887 ± 0.075

6.2.2. Aplicación al *dataset* de Tejidos Humanos

En este conjunto de datos, cuyas características se resumen en la sección B.1 del apéndice B, se miden 20172 genes en 32 tejidos humanos sanos con 3 réplicas biológicas en cada uno de ellos.

6.2.2.1. Búsqueda de clústers de co-expresión

Estamos interesados en identificar grupos de genes altamente correlados en sub-conjuntos de tejidos. Aplicamos el método $K_G x 1$ para buscar clústers de co-expresión (sección 6.1.1), a un conjunto de 5955 genes informativos, caracterizados por tener un nivel de expresión y

una variabilidad relevantes, en el sentido definido por el filtro etiquetado con L_t en la sección 4.3.2.1.

Aplicamos el procedimiento fijando $K_G = 96$ grupos de genes, un nivel de recorte de $\alpha_G = 0.3$ en la dimensión de genes y 16 grupos en la dimensión array, con recortes variando en el conjunto $\alpha_A = \{0,\ 0.1, 0.2, 0.3, 0.4, 0.5\}$. El número de comienzos aleatorios escogido en la aplicación del método fue 150.

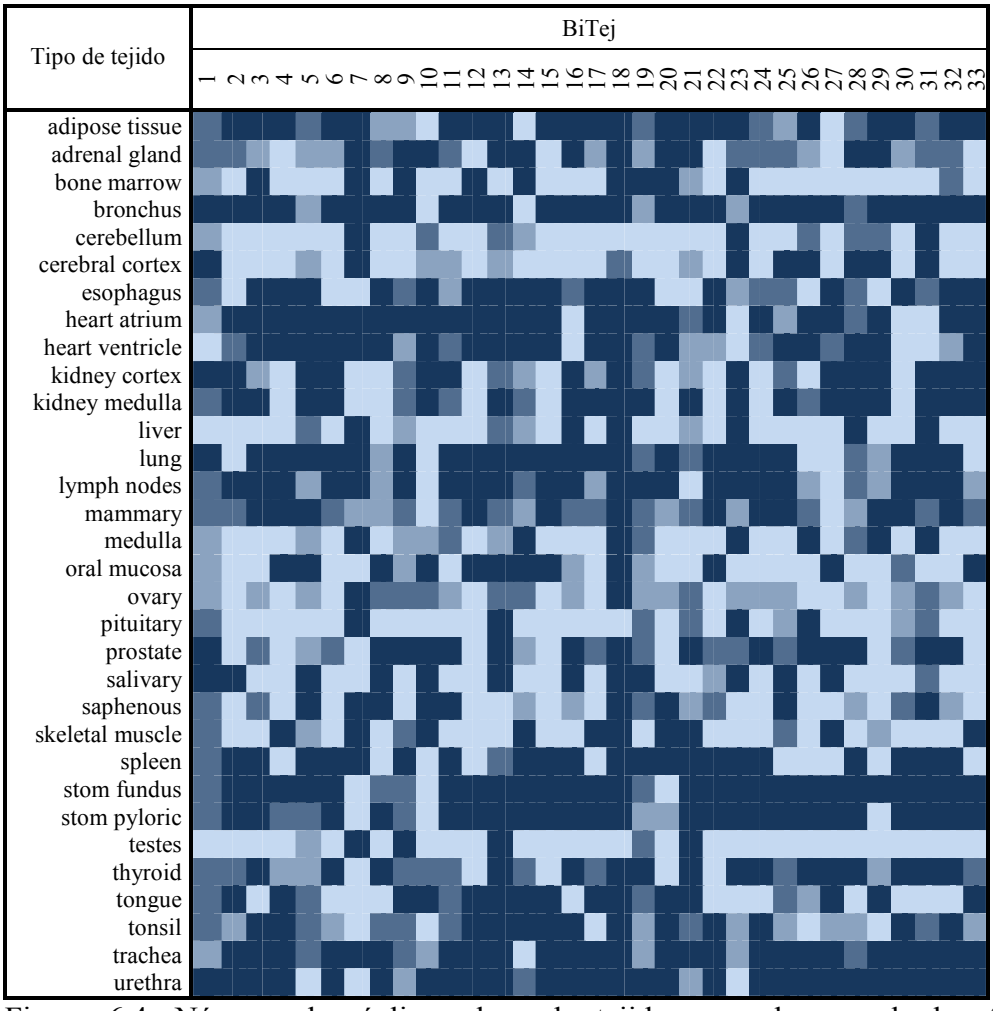


Figura 6.4. Número de réplicas de cada tejido en cada uno de los 33 clústers encontrados en el *dataset* de Tejidos Humanos con el método $K_G x1$ clústers de co-expresión. La intensidad de azul es mayor cuanto mayor es el número de réplicas de cada tipo de tejido incluidas en el clúster correspondiente.

Obtenemos los 33 clústers que se resumen en la tabla 6.6 y en la figura 6.4. Todas las correlaciones medias están entre 0.472 y 0.927 y el tamaño de los clústers se mueve en un amplio rango de posibles valores, desde 2 a 477 genes. El nivel de recorte en la dimensión de arrays suele ser alto, siendo el más frecuente $\alpha_A = 0.5$, y como mínimo, salvo algún

grupo que involucra a muy pocos genes, $\alpha_A = 0.3$. En este sentido destaca el clúster BiTej18, con 117 genes y un nivel de recorte $\alpha_A = 0.1$. En general, los tipos de tejido caracterizados por tener pocos genes específicamente expresados, como por ejemplo estómago, tráquea y uretra, aparecen en muchos de los clústers encontrados. Sin embargo, aquellos tejidos más específicos, como cerebelo, testículo o glándula pituitaria, no están muy representados en los clústers.

Tabla 6.6. *Dataset* de Tejidos Humanos. Información de los 33 clústers de co-expresión. Se señalan en negrita los clústers más relevantes.

Biclúster	# genes	# tejidos	$\overline{\rho}(Y_{IJ})$ media ± DT
BiTej1	3	58	0.521 ± 0.044
BiTej2	**307**	**48**	**0.743 ± 0.1**
BiTej3	305	58	0.611 ± 0.113
BiTej4	150	48	0.576 ± 0.139
BiTej5	2	58	0.73
BiTej6	218	48	0.58 ± 0.122
BiTej7	**64**	**58**	**0.795 ± 0.109**
BiTej8	98	48	0.508 ± 0.144
BiTej9	2	58	0.856
BiTej10	87	48	0.598 ± 0.103
BiTej11	**64**	**58**	**0.798 ± 0.066**
BiTej12	310	48	0.578 ± 0.142
BiTej13	2	77	0.927
BiTej14	2	58	0.742
BiTej15	227	48	0.601 ± 0.13
BiTej16	**45**	**58**	**0.8 ± 0.069**
BiTej17	**477**	**48**	**0.616 ± 0.121**
BiTej18	**117**	**86**	**0.668 ± 0.116**
BiTej19	3	58	0.828 ± 0.078
BiTej20	114	48	0.6 ± 0.126
BiTej21	2	58	0.826
BiTej22	228	48	0.472 ± 0.16
BiTej23	4	58	0.584 ± 0.112
BiTej24	**260**	**48**	**0.706 ± 0.11**
BiTej25	4	58	0.547 ± 0.127
BiTej26	**328**	**48**	**0.643 ± 0.121**
BiTej27	63	48	0.663 ± 0.113
BiTej28	37	58	0.695 ± 0.081
BiTej29	224	48	0.58 ± 0.122
BiTej30	**295**	**48**	**0.651 ± 0.117**
BiTej31	13	67	0.714 ± 0.059
BiTej32	45	58	0.588 ± 0.108
BiTej33	67	48	0.651 ± 0.099

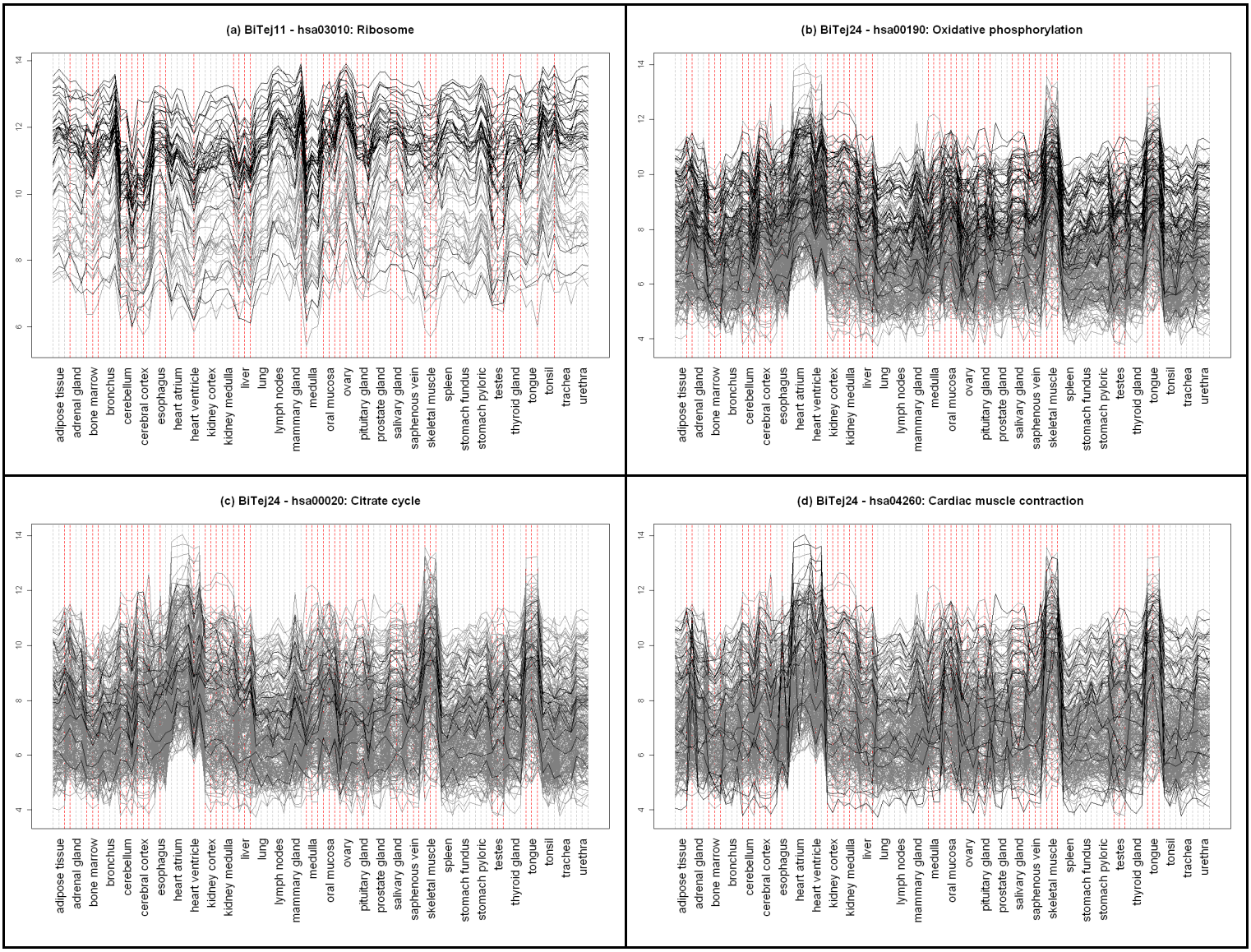


Figura 6.5. *Dataset* de Tejidos Humanos. (a) Perfil de expresión del clúster BiTej11 y su relación con la categoría KEGG hsa03010: *Ribosome*; y del clúster BiTej24 y su relación con las categorías (b) hsa00190: *Oxidative phosphorylation*, (c) hsa00020: *Citrate cycle* y (d) hsa04260: *Cardiac muscle contraction*. En rojo se marcan los arrays recortados. En negro se representan los genes que pertenecen a la KEGG correspondiente.

En la tabla 6.7 se muestra el análisis de enriquecimiento funcional de cada uno de los clústers obtenidos a partir de las categorías KEGG. El BiTej11 cubre el 59% de los genes anotados en la KEGG hsa03010: *Ribosome*, y además tiene un coeficiente de correlación medio próximo a 0.80. Se representa el perfil de expresión de este clúster en la figura 6.5 (a). Otras KEGG muy representadas son hsa00190: *Oxidative phosphorylation* (18%) y hsa00020: *Citrate cycle* (5%), hsa04260: *Cardiac muscle contraction* (7%) en el BiTej24.

Los perfiles de expresión de estos *clusters* y su relación con las categorías KEGG se representan en la figura 6.5 (b)-(d).

Tabla 6.7. *Dataset* de Tejidos Humanos. Relación entre las categorías KEGG sobre-representadas en los clúster*s* encontrados método $K_G x1$ clúster*s* de co-expresión.

Biclúster	KEGG	# genes KEGG	# genes biclúster	% genes biclúster	FDR p-valor
BiTej2 307 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.743$	hsa05340: Primary immunodeficiency	35	12	4.11	4.21E-07
	hsa04662: B cell receptor signaling pathway	75	14	4.79	2.97E-05
	hsa04650: Natural killer cell mediated cytotoxicity	133	17	5.82	1.48E-04
	hsa04612: Antigen processing and presentation	83	11	3.77	0.0378
BiTej11 64 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.798$	hsa03010: Ribosome	87	35	59.32	1.95E-56
BiTej16 45 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.8$	hsa04270: Vascular smooth muscle contraction	112	8	17.78	1.45E-04
	hsa04510: Focal adhesion	201	8	17.78	0.0076
BiTej17 477 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.616$	hsa00510: N-Glycan biosynthesis	46	11	2.44	1.54E-04
	hsa04142: Lysosome	117	13	2.88	0.0355
BiTej22 228 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.472$	hsa00071: Fatty acid metabolism	40	8	3.57	0.0025
BiTej24 260 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.706$	hsa00190: Oxidative phosphorylation	130	46	18.11	1.88E-43
	hsa00020: Citrate cycle (TCA cycle)	31	13	5.12	3.40E-10
	hsa04260: Cardiac muscle contraction	78	17	6.69	2.15E-09
	hsa00620: Pyruvate metabolism	40	10	3.94	8.16E-05
	hsa00010: Glycolysis / Gluconeogenesis	60	10	3.94	0.0031
BiTej26 328 genes $\bar{\rho}(\mathrm{Y}_{IJ}) = 0.643$	hsa04512: ECM-receptor interaction	84	12	3.77	0.0014
	hsa04510: Focal adhesion	201	17	5.35	0.0045

Hay tres clústers relevantes, BiTej7, BiTej18 y BiTej30, no relacionados con categorías KEGG. Sin embargo, esto no implica que no sean coherentes desde el punto de vista biológico. De hecho, en BiTej18 y BiTej30, se encuentran sobre-representadas, respectivamente, las categorías GO: 0007268 ~ *synaptic transmission* y GO: 0031424 ~ *keratinization*, de la ontología GO-BP. Por su parte, BiTej7, no parece estar relacionado con procesos biológicos, pero si con componentes celulares, estando significativamente representada la categoría GO: 0005576 ~ *Extracellular region*, de la ontología GO-CC.

6.2.2.2. Búsqueda de $K_G x K_A$ biclústers

Para la aplicación del método de los $K_G x K_A$ biclústers, consideramos los conjuntos de genes que aparecieron con algún par selectivamente expresado respecto de su núcleo de expresión. Estos listados de genes, uno por cada tipo de tejido, se obtienen a partir del

contraste propuesto en el capítulo 4, y se describen en la tabla 4.12. Realizamos una aplicación del método de los $K_G x K_A$ biclústers independiente para cada tejido, fijando siempre $K_G = 6$ clústers en la dimensión de genes, $K_A = 3$ clústers en la de arrays y niveles de recorte $\alpha_G = 0.3$ y $\alpha_A = 0$, en cada una de las dimensiones respectivamente. Se consideran 1000 comienzos aleatorios en todos los casos. En la tabla 6.8 aparecen los resultados obtenidos en estas aplicaciones del procedimiento.

El 37.5% de los tejidos se agrupan en dos clústers de genes. Como máximo, el método encuentra 4 grupos de genes en tejido adiposo, lengua y amígdala. En el otro extremo, los genes de pulmón, ovario, saliva, vena safena, estómago, glándula tiroidea y uretra, se clasifican en un único grupo. El 46.48% de los biclústers tienen dos grupos de arrays. El 31% de los clústers de genes, tienen tres grupos de arrays, habitualmente en dos niveles de expresión diferentes, pero en el mismo sentido, sobre-expresados. Salvo algún caso aislado, y como cabía esperar, las tres réplicas del mismo tejido suelen agruparse juntas, en los grupos de arrays más pequeños y diferencialmente expresados. En la tabla 6.8 resumimos los grupos encontrados, en al aplicación correspondiente a cada tipo de tejido, en aquellos casos en los que se ha encontrado más de un grupo de genes. Se muestra el número de genes, el número de arrays, entre paréntesis el número de réplicas del correspondiente tejido, la media y la desviación típica de la expresión en cada uno de los biclústers.

Tabla 6.8. *Dataset* de Tejidos Humanos. Información de las particiones encontradas con el método de los $K_G x K_A$ *biclústers* considerando los genes selectivamente expresados en cada tipo de tejido.

		Biclúster			
		BiTejKxK1	BiTejKxK2	BiTejKxK3	BiTejKxK4
Adipose tissue 165 genes	**# genes** / # arrays (# rép.)	**100** / 4 (3), 0 (0), 92 (0)	**4** / 4 (3), 91 (0) , 1 (0)	**2** / 18 (0), 74 (0), 4 (3)	**9** / 0 (0), 92 (0), 4 (3)
	$\hat{\mu}_{kh}$	2.97,0,0.31	4.07,0.25,1.76	2.1,1.42,4.37	0,0.3,4.96
	$\hat{\sigma}_{kh}$	1.32 ,0 ,1.07	1.17 ,0.47 ,0.93	0.36 ,0.35 ,0.48	0 ,1.58 ,0.99
Adrenal gland 122 genes	**# genes** / # arrays (# rép.)	**2** / 15 (3), 0 (0), 81 (0)	**45** / 3 (3), 89 (0), 4(0)	**38** / 4 (3), 0 (0), 92 (0)	
	$\hat{\mu}_{kh}$	3.41,0,1.28	3.59,0.5,1.39	2.97,0,0.25	
	$\hat{\sigma}_{kh}$	0.69 ,0 ,0.74	1.4 ,0.81 ,1.36	1.12 ,0 ,1.34	
Bone marrow 532 genes	**# genes** / # arrays (# rép.)	**14** / 15 (3), 0 (0), 81 (0)	**77** / 78 (0), 12 (3), 6 (0)	**281** / 0 (0), 0 (0), 96 (3)	
	$\hat{\mu}_{kh}$	3.71,0,0.96	1.1,3.72,2.57	0,0,0	
	$\hat{\sigma}_{kh}$	0.8 ,0 ,0.86	0.54 ,1.14 ,0.86	0 ,0 ,1.14	
Bronchus 153 genes	**# genes** / # arrays (# rép.)	**43** / 0 (0), 96 (3), 0 (0)	**16** / 90 (0), 0 (0), 6 (3)		
	$\hat{\mu}_{kh}$	0,0,0	0.47,0,3.27		
	$\hat{\sigma}_{kh}$	0 ,1.87 ,0	1.08 ,0 ,1.69		

Tabla 6.8. Continuación I

		Biclúster			
		BiTejKxK1	BiTejKxK2	BiTejKxK3	BiTejKxK4
Cerebellum 451 genes	**# genes** / # arrays (# rép.)	**105** / 7 (3), 0 (0), 89 (0)	**164** / 9 (3), 87 (0), 0 (0)	**47** / 83 (0), 12 (3), 1 (0)	
	$\hat{\mu}_{kh}$	2.44,0,0.49	3.7,0.51,0	0.86,2.78,1.23	
	$\hat{\sigma}_{kh}$	1.87 ,0 ,1.27	1.8 ,0.68 ,0	0.44 ,1.33 ,0.47	
cerebral cortex 527 genes	**# genes** / # arrays (# rép.)	**184** / 89 (0), 7 (3), 0 (0)	**185** / 6 (0), 87 (0), 3 (3)		
	$\hat{\mu}_{kh}$	0.49,3.3,0	2.95,0.55,4.11		
	$\hat{\sigma}_{kh}$	1.17 ,1.41 ,0	1.2 ,0.67 ,1.09		
esophagus 296	**# genes** / # arrays (# rép.)	**32** / 7 (3), 5 (0), 84 (0)	**175** / 90 (0), 0 (0), 6 (3)		
	$\hat{\mu}_{kh}$	6.06,3.56,0.66	0.46,0,4.19		
	$\hat{\sigma}_{kh}$	1.5 ,1.43 ,0.42	0.97 ,0 ,1.36		
heart atrium 198 genes	**# genes** / # arrays (# rép.)	**80** / 96 (3), 0 (0), 0 (0)	**19** / 81 (0), 12 (3), 3 (0)		
	$\hat{\mu}_{kh}$	0,0,0	0.78,5.72,2.87		
	$\hat{\sigma}_{kh}$	1.5 ,0 ,0	0.44 ,1.59 ,1.27		
heart ventricle 187 genes	**# genes** / # arrays (# rép.)	**7** / 12 (3), 75 (0), 9 (0)	**66** / 0 (0), 0 (0), 96 (3)		
	$\hat{\mu}_{kh}$	4.25,1.54,2.35	0,0,0		
	$\hat{\sigma}_{kh}$	0.73 ,0.62 ,0.64	0 ,0 ,1.57		
kidney cortex 362 genes	**# genes** / # arrays (# rép.)	**40** / 90 (0), 6 (3), 0 (0)	**213** / 0 (0), 92 (0), 4 (3)		
	$\hat{\mu}_{kh}$	0.37,4.92,0	0,0.27,3.17		
	$\hat{\sigma}_{kh}$	1.39 ,2.1 ,0	0 ,0.84 ,1.37		
kidney medulla 196 genes	**# genes** / # arrays (# rép.)	**114** / 90 (0), 0 (0), 6 (3)	**1** / 12 (0), 77 (0), 7 (3)	**22** / 2 (1), 94 (2), 0 (0)	
	$\hat{\mu}_{kh}$	0.37,0,3.34	8.22,1,4.46	3.15,0.15,0	
	$\hat{\sigma}_{kh}$	0.86 ,0 ,1.26	1.06 ,0.47 ,0.78	0.65 ,1.56 ,0	
liver 509 genes	**# genes** / # arrays (# rép.)	**284** / 3 (3), 93 (0), 0 (0)	**72** / 93 (0), 0 (0), 3 (3)		
	$\hat{\mu}_{kh}$	3.72,0.22,0	0.17,0,5.71		
	$\hat{\sigma}_{kh}$	1.24 ,1.05 ,0	0.66 ,0 ,1.1		
lymph nodes 277 genes	**# genes** / # arrays (# rép.)	**102** / 8 (0), 9 (3), 79 (0)	**3** / 12 (3), 84 (0), 0 (0)	**89** / 15 (1), 78 (0), 3 (2)	
	$\hat{\mu}_{kh}$	2.63,3.88,1	3.95,1.08,0	3.03,1.17,4.06	
	$\hat{\sigma}_{kh}$	0.97 ,1.09 ,0.52	0.66 ,1.41 ,0	1.05 ,0.77 ,0.93	
mammary gland 23 genes	**# genes** / # arrays (# rép.)	**2** / 36 (3), 60 (0), 0 (0)	**1** / 27 (0), 29 (0), 40 (3)	**4** / 0 (0), 0 (0), 96 (3)	
	$\hat{\mu}_{kh}$	-0.34,3.35,0	3.39,1.95,5.1	0,0,0	
	$\hat{\sigma}_{kh}$	0.7 ,0.59 ,0	0.4 ,0.35 ,0.61	0 ,0 ,2.13	
medulla 378 genes	**# genes** / # arrays (# rép.)	**147** / 5 (3), 4 (0), 87 (0)	**51** / 0 (0), 96 (3), 0 (0)		
	$\hat{\mu}_{kh}$	3.58,2.31,0.67	0,0,0		
	$\hat{\sigma}_{kh}$	0.97 ,1.59 ,0.98	0 ,1.43 ,0		
oral mucosa 384 genes	**# genes** / # arrays (# rép.)	**112** / 1 (1), 10 (2), 85 (0)	**72** / 0 (0), 0 (0), 96 (3)		
	$\hat{\mu}_{kh}$	5.71,4.26,0.77	0,0,0		
	$\hat{\sigma}_{kh}$	2.17 ,2.06 ,0.71	0 ,0 ,1.56		
pituitary gland 260 genes	**# genes** / # arrays (# rép.)	**24** / 0 (0), 0 (0), 96 (3)	**121** / 10 (3), 0 (0), 86 (0)	**4** / 0 (0), 93 (0), 3 (3)	
	$\hat{\mu}_{kh}$	0,0,0	4.05,0,0.72	0,0.17,4.06	
	$\hat{\sigma}_{kh}$	0 ,0 ,1.34	1.75 ,0 ,0.82	0 ,0.43 ,1.09	
prostate gland 213 genes	**# genes** / # arrays (# rép.)	**8** / 0 (0), 93 (0), 3 (3)	**141** / 90 (0), 3 (0), 3 (3)		
	$\hat{\mu}_{kh}$	0,0.22,5.23	0.4,1.13,3.21		
	$\hat{\sigma}_{kh}$	0 ,1.9 ,0.66	0.96 ,0.88 ,1.25		

Tabla 6.8. Continuación II

		Biclúster			
		BiTejKxK1	BiTejKxK2	BiTejKxK3	BiTejKxK4
skeletal muscle 438 genes	**# genes** / # arrays (# rép.)	**159** / 8 (3), 88 (0), 0 (0)	**131** / 0 (0), 90 (0), 6 (3)	**16** / 90 (0), 0 (0), 6 (3)	
	$\hat{\mu}_{kh}$	2.93,0.61,0	0,0.42,3.66	0.41,0,4.28	
	$\hat{\sigma}_{kh}$	1.04 ,0.7 ,0	0 ,1.29 ,1.26	0.46 ,0 ,1.45	
spleen 241 genes	**# genes** / # arrays (# rép.)	**79** / 3 (3), 79 (0), 14 (0)	**81** / 77 (0), 10 (0), 9 (3)	**9** / 10 (3), 0 (0), 86 (0)	
	$\hat{\mu}_{kh}$	3.91,1.1,2.77	1.12,2.71,4.17	3.64,0,0.77	
	$\hat{\sigma}_{kh}$	0.86 ,0.83 ,1.11	0.53 ,1.04 ,1.17	1.3 ,0 ,1.44	
testes 845 genes	**# genes** / # arrays (# rép.)	**222** / 3 (3), 88 (0), 5 (0)	**369** / 0 (0), 93 (0), 3 (3)		
	$\hat{\mu}_{kh}$	5.38,0.37,0.11	0,0.2,4.02		
	$\hat{\sigma}_{kh}$	1.37 ,0.26 ,0.19	0 ,0.71 ,1.73		
tongue 359 genes	**# genes** / # arrays (# rép.)	**49** / 84 (0), 0 (0), 12 (3)	**19** / 4 (1), 92 (2), 0 (0)	**113** / 0 (0), 84 (0), 12 (3)	**70** / 4 (3), 90 (0), 2 (0)
	$\hat{\mu}_{kh}$	0.98,0,2.95	3.05,0.26,0	0,0.86,4.41	5.44,0.39,6.46
	$\hat{\sigma}_{kh}$	0.57 ,0 ,1.01	1.27 ,1.33 ,0	0 ,1.06 ,1.98	1.69 ,0.9 ,1.85
tonsil 276 genes	**# genes** / # arrays (# rép.)	**90** / 17 (3), 0 (0), 79 (0)	**23** / 12 (0), 8 (3), 76 (0)	**13** / 87 (0), 9 (3), 0 (0)	**67** / 7 (0), 6 (3), 83 (0)
	$\hat{\mu}_{kh}$	3.39,0,1.14	2.93,4.41,1.1	0.8,5.46,0	2.73,3.7,0.79
	$\hat{\sigma}_{kh}$	1.3 ,0 ,0.9	1.01 ,0.89 ,0.52	1.93 ,2.08 ,0	0.82 ,0.84 ,0.5
trachea 174 genes	**# genes** / # arrays (# rép.)	**90** / 90 (0), 6 (3), 0	**6** / 0 (0), 0 (0), 96 (3)		
	$\hat{\mu}_{kh}$	0.44,3.17,0	0,0,0		
	$\hat{\sigma}_{kh}$	1.32 ,1.2 ,0	0 ,0 ,2.16		

En esta aplicación, en general, los grupos de arrays obtenidos involucran a tipos de tejido relacionados. Como ejemplo, en la figura 6.6 se muestran los perfiles de expresión de los tres clústers de genes asociados al tejido cerebelo. En los tres casos, el cerebelo aparece sobre-expresado junto con las tres réplicas de corteza cerebral. En BiTejKxK1 (figura 6.6 (a)) estos dos tipos de tejido se sobre-expresan, junto con una réplica de medula, aproximadamente a 2 unidades respecto de grupo mayoritario, formado por el resto de arrays. En BiTejKxK2 (figura 6.6 (b)), el conjunto formado por cerebelo, corteza cerebral y médula tiene un grado de sobre-expresión mayor, aproximadamente de 3 unidades. Por último en BiTejKxK3 (figura 6.6 (c)) al grupo de cerebelo, corteza cerebral y medula, se une la glándula pituitaria, sobre-expresado 2 unidades respecto del grupo de arrays mayoritario, que incluye el resto de tipos de tejidos salvo una de las réplicas de ovario, que se agrupa por separado.

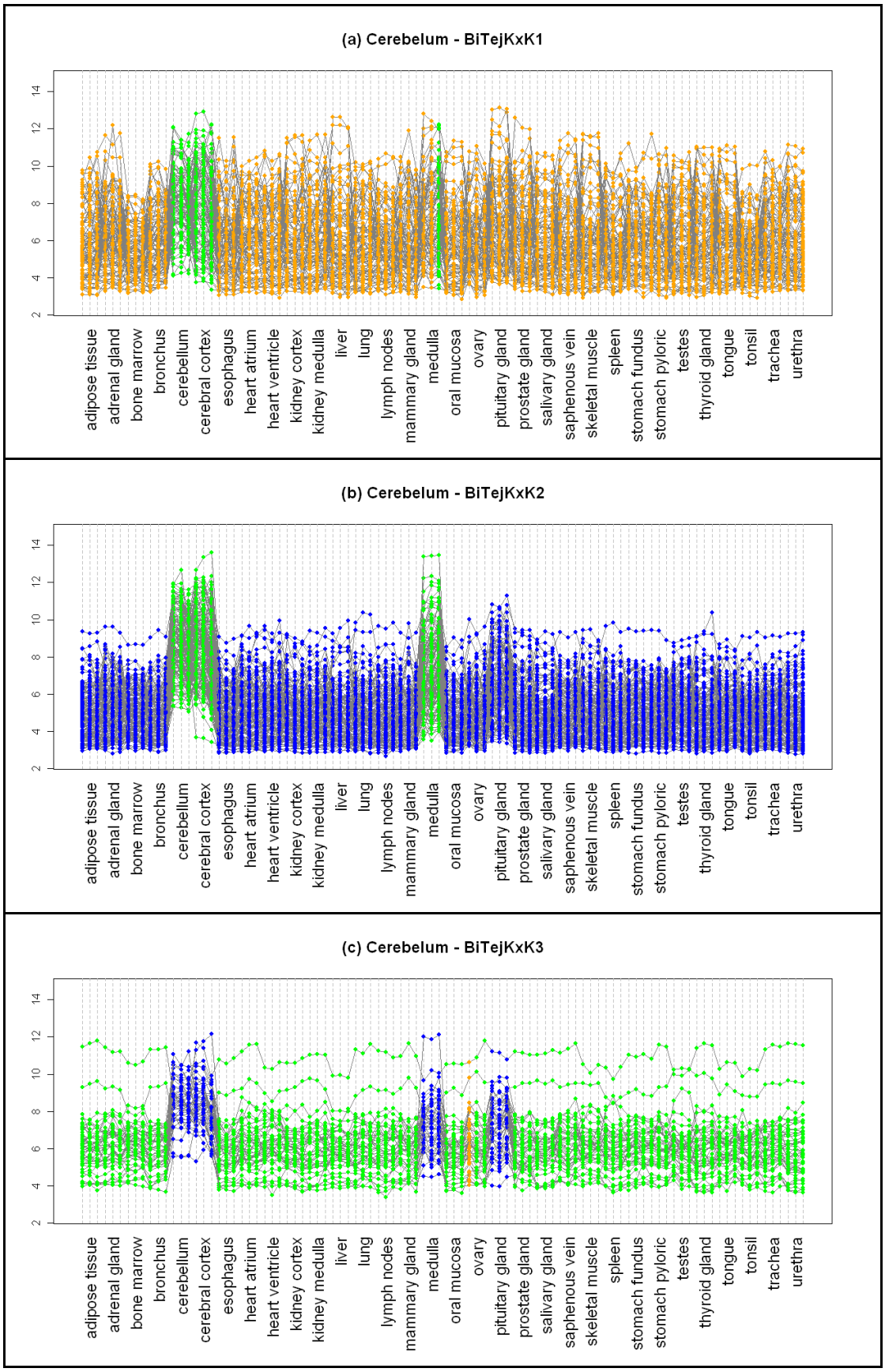


Figura 6.6. *Dataset* de Tejidos Humanos. Perfil de expresión de los biclústers asociados al tejido *Cerebelum*. En distintos colores se señalan los grupos de la dimensión arrays.

Esta aproximación permite también identificar grupos asociados a tipos de tejido muy específico. Un ejemplo de esta clase sería el hígado, que se representa en la figura 6.7. En este listado de 509 genes, se encuentran dos clúster de genes, ambos con dos clústers de arrays que agrupan a las tres réplicas de hígado, pero a distintos niveles de expresión. En el primer grupo (figura 6.7 (a)), la sobre-expresión media de estas tres réplicas se estima en 3.5 unidades por encima de la expresión mayoritaria. En el segundo caso (figura 6.7 (b)), el hígado está más sobre-expresado, aproximadamente 5.5 unidades más que el resto de tejidos.

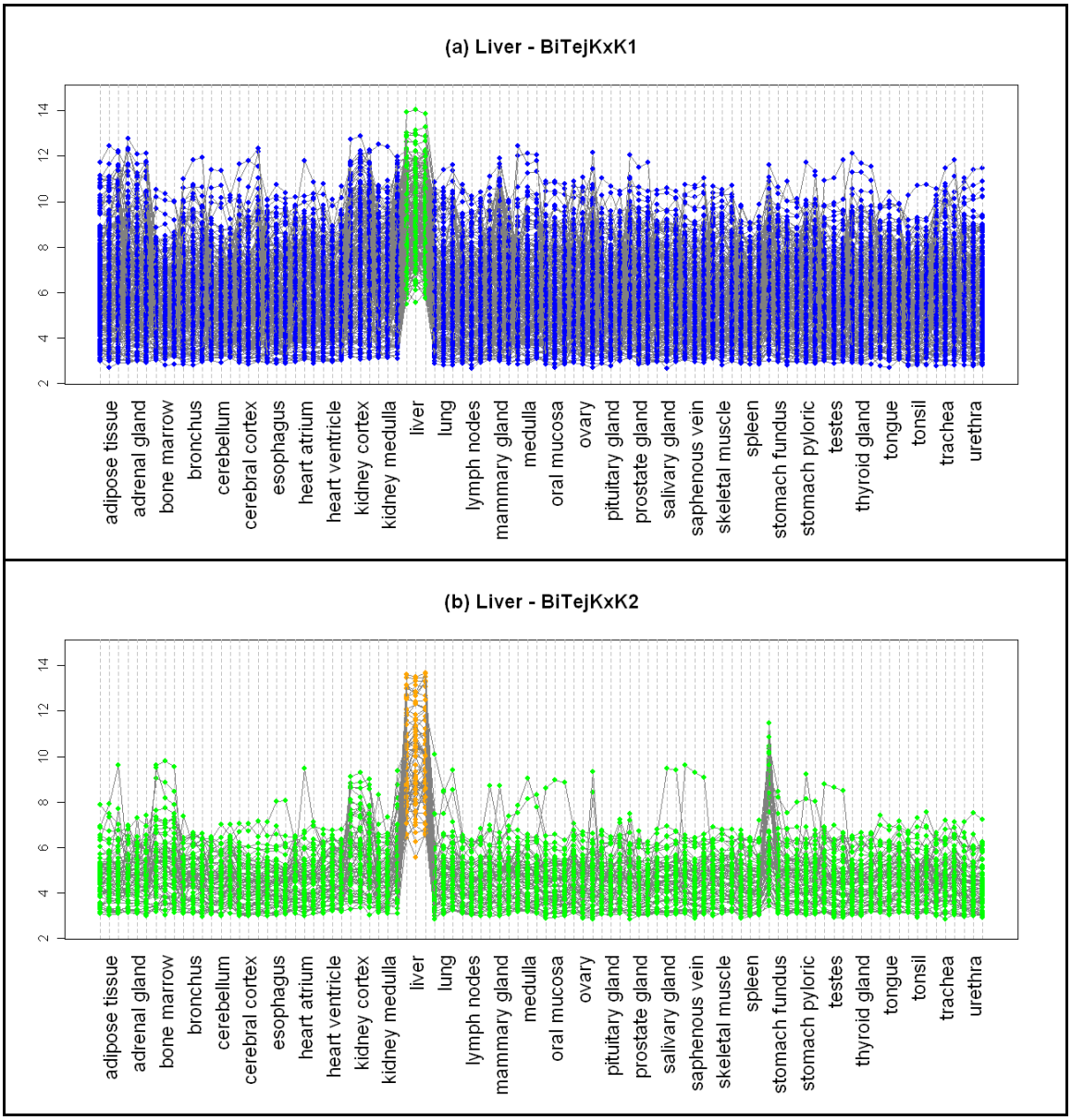


Figura 6.7. *Dataset* de Tejidos Humanos. Perfil de expresión de los biclústers asociados al tejido *Liver*. En distintos colores se señalan los grupos de la dimensión arrays.

6.2.3. Aplicación al *dataset* de Cáncer de Mama I

En este *dataset*, propuesto en **[Hatzis et al, 2011]** y cuyas características se resumen en la sección B.3 del apéndice B, se miden 12576 genes en 310 muestras de cáncer de mama, divididas en dos grupos: 113 que se corresponden con pacientes sensibles a un determinado tratamiento de quimioterapia y 197 no sensibles a ese mismo tratamiento.

6.2.3.1. Búsqueda de clústers de co-expresión

Una posible aplicación del método que busca clústers de co-expresión está relacionada con la selección de sondas representativas del conjunto de sondas disponibles para un mismo gen. Como ya se ha comentado en el capítulo 2, una de las características de los microarrays de *Affymetrix*, es que cada conjunto de sondas asignado a un gen debe resumirse en un único valor. Un problema al que se enfrentan muchos de los métodos de pre-procesado, es que, con bastante frecuencia, aparecen conjuntos de sondas con una o varias sondas que responden de manera distinta al resto. Como ejemplo, en la figura 6.8 se representa el conjunto de sondas 211633_x_at, asociado al gen $IGHG1$, formado por 11 sondas, 3 de las cuales están poco correladas con el resto. En esta sección, proponemos utilizar la búsqueda de clústers de co-expresión para encontrar el mejor conjunto de sondas en un *probe-set* en el sentido de co-expresión en una mayoría de arrays.

A continuación se muestran los resultados obtenidos de la aplicación del método de los $K_G x1$ clústers de co-expresión, a cada uno de 654 conjuntos de sondas en el *dataset* de Cáncer de Mama I. Estos aparecen al considerar 95 genes caracterizados por tener asignados 6 o más conjuntos de sondas. La distribución del número de conjuntos de sondas por gen se encuentra en la tabla 6.9. Este conjunto va a permitir evaluar la correlación entre las sondas de un mismo conjunto de sondas y entre sondas del mismo gen en distintos conjuntos de sondas.

Tabla 6.9. *Dataset* de Cáncer de Mama I. Número de conjuntos de sondas y genes analizados según el número de conjuntos de sondas por gen.

	probe-sets por gen								Total
	6	7	8	9	10	11	12	13	
# probe-sets	348	112	64	54	40	11	12	13	654
# genes	58	16	8	6	4	1	1	1	95

Como medida de evaluación, vamos a utilizar el coeficiente de correlación medio. Aplicamos el método a cada conjunto de sondas de manera independiente fijando los niveles de recorte en, $\alpha_G = 0.3$ para la dimensión sonda y $\alpha_A = 0.2$ en la dimensión array. En todos los casos se consideran 150 comienzos aleatorios.

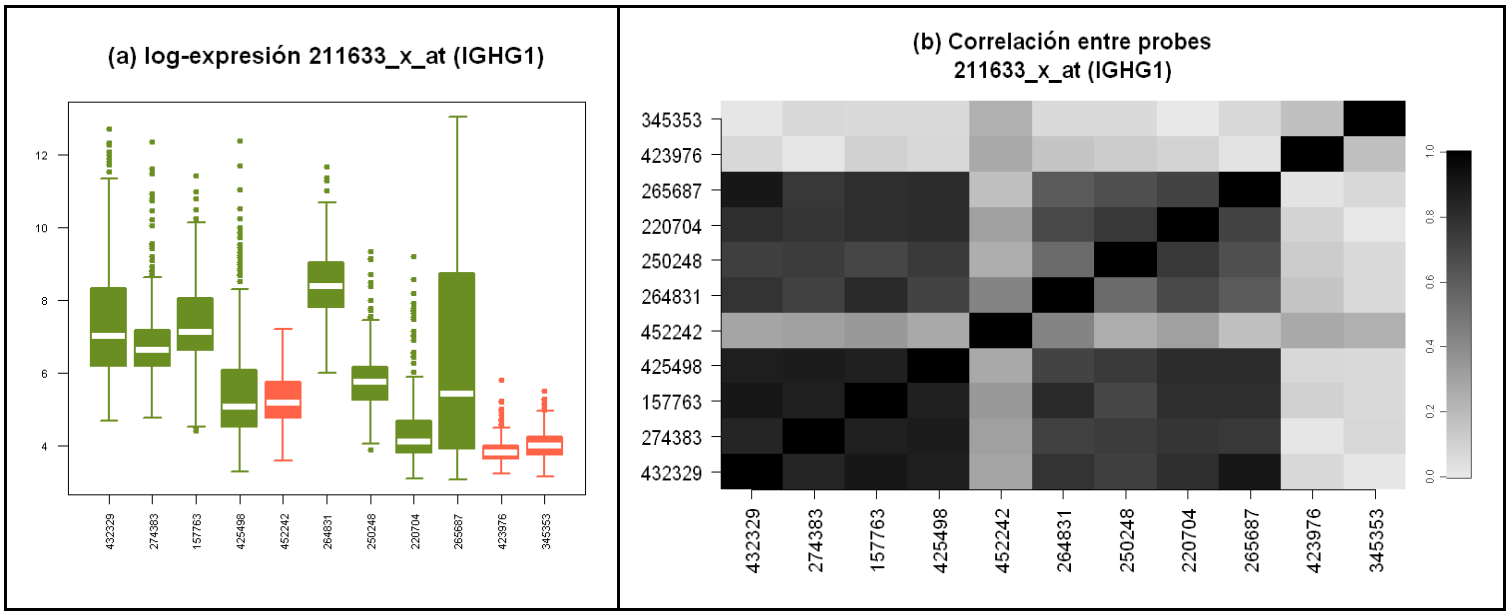


Figura 6.8. *Dataset* de Cáncer de Mama I. (a) Niveles de expresión del conjunto de sondas 211633_x_at (IGHG1) en cada una de sus sondas. Cada caja representa la distribución de una sonda. En verde están marcadas las *probes* seleccionadas a partir de un recorte en la dimensión probe de $\alpha_G = 0.3$ y de $\alpha_A = 0.2$ en la dimensión array. En rojo están marcadas las *probes* eliminadas. (b) Matriz de correlación entre las *probes* del conjunto 211633_x_at (IGHG1).

Tabla 6.10. *Dataset* de Cáncer de Mama I. Distribución del coeficiente de correlación medio *intra-probeset*, $\bar{\rho}_{PS}(.)$, e intra-gen, $\bar{\rho}_{GEN}(.)$.

	Percentiles								
	0%	10%	25%	50%	70%	75%	90%	95%	100%
$\bar{\rho}_{(PS)}(\mathbf{Y}^*_{GA})$	-0.0482	-0.0253	-0.0149	-0.0006	0.0136	0.0183	0.0410	0.0580	0.1692
$\bar{\rho}_{PS}(\mathbf{Y}_{GA})$	-0.0086	0.0816	0.1419	0.2741	0.4337	0.4849	0.674	0.7422	0.9294
$\bar{\rho}_{PS}(\mathbf{Y}_{GJ})$	0.0196	0.1119	0.1787	0.3214	0.4881	0.5579	0.7418	0.8022	0.9415
$\bar{\rho}_{PS}(\mathbf{Y}_{IJ})$	0.0565	0.2024	0.3016	0.4976	0.7045	0.7435	0.8547	0.9003	0.9747
$\bar{\rho}_{(GEN)}(\mathbf{Y}^*_{GA})$	-0.0057	-0.0012	0.0010	0.0038	0.0065	0.0074	0.0117	0.0148	0.0259
$\bar{\rho}_{GEN}(\mathbf{Y}_{GA})$	0.012	0.07	0.1032	0.163	0.2489	0.2889	0.4671	0.539	0.7413
$\bar{\rho}_{GEN}(\mathbf{Y}_{IA})$	0.0135	0.08	0.1217	0.2073	0.3359	0.3685	0.5843	0.6607	0.8508

En la tabla 6.10 mostramos la distribución de los coeficientes de correlación medios intra-*probeset* a partir de (i) todas las sondas y todos los arrays, $\bar{\rho}_{PS}(\mathbf{Y}_{GA})$, (ii) todas las sondas y los arrays seleccionados, $\bar{\rho}_{PS}(\mathbf{Y}_{GJ})$, y (iii) las sondas y arrays seleccionados, $\bar{\rho}_{PS}(\mathbf{Y}_{IJ})$. También se ofrece $\bar{\rho}_{(PS)}(\mathbf{Y}^*_{GA})$ calculado de la misma forma, pero a partir de 1000

permutaciones aleatorias de la asignación a *probe-set*. En general, las correlaciones medias entres las sondas seleccionadas por el procedimiento son superiores a las obtenidas cuando se considera *probe-set* completo.

En esta tabla se muestra, también, la distribución de los coeficientes de correlación medios dentro de cada gen teniendo en cuenta (iv) todas las sondas, $\overline{\rho}_{GEN}(\mathbf{Y}_{GA})$, y (v) sólo las seleccionadas en cada conjunto de sondas, $\overline{\rho}_{GEN}(\mathbf{Y}_{IA})$. Igual que el caso de *probe-sets*, se calcula $\overline{\rho}_{(GEN)}(\mathbf{Y}^{*}_{GA})$, permutando aleatoriamente la asignación de *probe-set* a gen. Aunque el cambio es menor que el observado a nivel de *probe-set*, los coeficientes de correlación medios son mayores cuando sólo se tienen en cuenta las sondas seleccionadas. Estas distribuciones se representan en la figura 6.9.

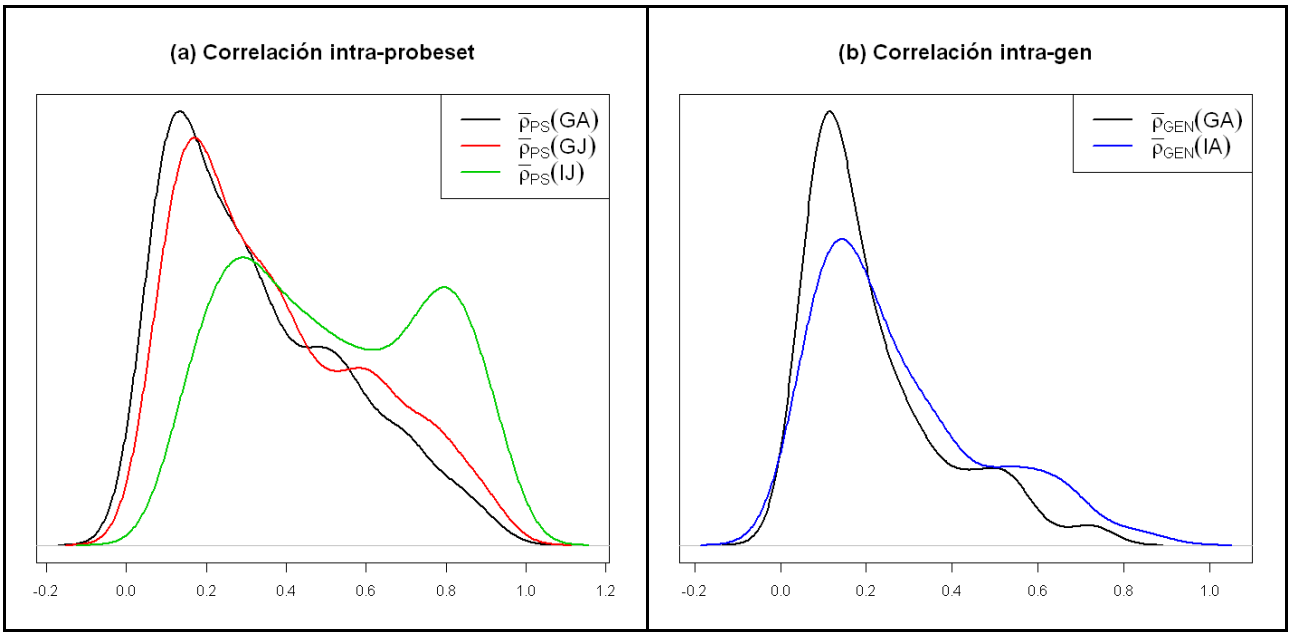


Figura 6.9. *Dataset* de Cáncer de Mama I. Distribución del coeficiente de correlación medio (a) entre sondas de cada una de los conjuntos de sondas de la muestra de 654 y (b) entre *probes* asignadas a cada uno de los 95 genes seleccionados.

6.2.3.2. Búsqueda de patrones de *outliers*

En la sección 5.3.3 ilustramos como el hospital de procedencia de las muestras, recogido en la variable *source*, estaba relacionado con la aparición de expresión diferencial en un conjunto de genes. En aquella ocasión, se propuso evaluar la relación entre las variables disponibles y los patrones de contaminación identificados en los genes. El método descrito en la sección 6.1.3 permite obtener grupos de genes cuyos niveles de expresión aparecen como atípicos en el mismo conjunto de arrays, permitiendo relacionar estos grupos de genes con variables clínicas conocidas. En esta sección se aplica esta metodología al *dataset* de Cáncer de Mama I.

Para cada gen g, se definen los arrays atípicos por sobre-expresión según la expresión (6.33) fijando un nivel de recorte de $\alpha = 0.45$ para determinar el *core* de cada gen y una constante $f = 1$, correspondiente a considerar como atípico lo que dista una desviación típica de la media. No somos demasiados restrictivos en la definición de atípico al tomar como umbral $f = 1$ para, de esta forma, intentar identificar mejor el patrón asociado a la sobre-expresión. El método de los $K_G x 2$ clústers de *outliers* se aplica al conjunto de 6629 genes (52.7% del total de genes) que, según esta definición, cuentan con más de 80 arrays *outlier*. Los parámetros del método se fijan en $K_G = 30$ grupos de genes y niveles de recorte $\alpha_G = 0.2$, para la dimensión de genes y $\alpha_A = 0.05$ para la dimensión de arrays. Se consideran 1000 comienzos aleatorios.

La solución obtenida, considerando sólo los grupos con más de 10 genes, se resume en la tabla 6.11. Para cada clúster de genes, denotado por k, se muestra: (i) el número de genes, (ii) el peso de cada clúster de arrays, $\hat{\pi}_{kj}$, donde $j = 2$ en el grupo de arrays clasificados como *outlier*, (iii) la consistencia de cada biclúster, medida como el porcentaje de pares gen - array *outlier* en cada grupo, $\hat{o}_{kj}$, (iv) el coeficiente de correlación medio en el clúster de genes considerando todos los arrays, denotado por $\overline{\rho}(\mathbf{Y}_{IA})$, y (v) los percentiles 20 y 80 de las diferencias de medianas de expresión en cada clúster de arrays, en valor absoluto. Además, se incluye la relación del patrón de sobre-expresión correspondiente a cada clústers de genes obtenido con las variables *source* y ER. Esta relación se evalúa en términos de sensibilidad y especificidad a la hora de clasificar como atípicas las muestras de una determinada categoría.

Destaca que los distintos patrones de contaminación parecen no responder a grupos de genes muy correlados. Esto significaría que esta partición no podría encontrarse buscando agrupar genes correlados. Por otra parte, el porcentaje de pares gen - array clasificados como *outlier* en los biclúster que intentan recoger a los arrays atípicos (con $j = 2$), es alto, variando entre el 66% y 81%. A su vez, en los grupos que intentan identificar a los arrays no atípicos (con $j = 1$) estos porcentajes son inferiores al 25%. Con este tipo de porcentajes y con la diferencia de medianas individualizados para cada grupo de genes, intentamos medir la *calidad* de cada biclúster identificado.

En cuanto a la diferencia de medianas, destacan los clústers 12 y 29, en los que esta diferencia se sitúa en valores superiores a 1.6 unidades. Para estos dos grupos de genes, los valores de sensibilidad y especificidad en la identificación de *outliers* se encuentran, ambos, por encima del 75%. Clústers de genes asociados a biclústers con la *calidad* mostrada por estos dos grupos, el 12 y el 29, con más de 100 genes cada uno y más del 30% de arrays implicados en el grupo etiquetado como atípico, merecen un estudio pormenorizado. En este

caso, los grupos de arrays que caracterizan estos dos grupos de genes, aparecen asociados con la variable *source* y la categoría ER+, con valores de sensibilidad y especificidad, en los dos casos, superiores al 85%. De esta forma, estos patrones de comportamiento atípico identificados, encuentran explicación biológica en su relación con estas variables, pero si no hubiera sido así, la calidad de los patrones encontrados hubiese sido suficiente como para suponer la existencia de alguna variable *hidden* que los justificase.

Tabla 6.11. *Dataset* de Cáncer de Mama I. Grupos con más de 10 genes obtenidos a partir del método $K_G x 2$ clústers de *outliers* aplicado a los 6629 genes con más arrays *outlier*, y su relación con las variables *source* y ER.

k	# gen	Peso de los clúster de arrays		% de outliers		$\bar{\rho}(\mathbf{Y}_{IA})$	Diferencia medianas de expresión		Source: ISPY		ER-		ER+	
		$\hat{\pi}_{k1}$	$\hat{\pi}_{k2}$	$\hat{o}_{k1}$	$\hat{o}_{k2}$		P_{20}	P_{80}	Sens	Esp	Sens	Esp	Sens	Esp
1	60	90.2	9.8	22.64	68.1	0.152	0.194	0.5091	0.375	0.9911	0.0153	0.8364	0.1636	0.9847
2	160	90.48	9.52	22.69	68.66	0.1577	0.1499	0.3465	0.0244	0.8774	0.0866	0.8982	0.1018	0.9134
4	294	71.09	28.91	19.2	69.9	0.3119	0.336	1.0285	0.3377	0.7281	0.6639	0.9657	0.0343	0.3361
5	577	75.51	24.49	14.94	75.46	0.4462	0.211	0.3736	0.2208	0.7465	0.2283	0.7425	0.2575	0.7717
9	327	78.23	21.77	17.17	73.55	0.3269	0.1721	0.365	0.4795	0.8688	0.1969	0.7665	0.2335	0.8031
10	249	70.85	29.15	16.19	79.49	0.4923	0.2733	0.7271	0.0122	0.6009	0.2619	0.6864	0.3136	0.7381
11	371	85.03	14.97	21.92	69.88	0.2267	0.2692	0.6643	0.0122	0.7972	0.0752	0.7888	0.2112	0.9248
12	110	50.51	49.49	14.36	75.01	0.4179	0.5163	1.6821	0.5	0.507	0.0455	0.1411	0.8589	0.9545
13	190	76.27	23.73	17.66	73.19	0.3797	0.2271	0.4227	0.0122	0.6761	0.1938	0.7289	0.2711	0.8062
15	271	78.57	21.43	15.75	76.18	0.4229	0.2433	0.4249	0.1316	0.7569	0.1953	0.7711	0.2289	0.8047
18	96	88.78	11.22	22.53	66.04	0.1589	0.2515	0.5878	0.0513	0.8657	0	0.7963	0.2037	1
19	149	74.49	25.51	16.2	72.98	0.4663	0.2221	0.4032	0.3247	0.7696	0.2302	0.7262	0.2738	0.7698
20	425	80.27	19.73	20.17	67.47	0.3059	0.3474	0.7654	0.1566	0.7867	0.2581	0.8471	0.1529	0.7419
21	133	68.81	31.19	16.65	73.33	0.465	0.4112	0.975	0.2963	0.6822	0.5039	0.8333	0.1667	0.4961
22	280	70.07	29.93	16.19	76.16	0.4887	0.3893	1.0622	0.1375	0.6402	0.3889	0.7679	0.2321	0.6111
23	395	76.53	23.47	14.02	78.08	0.4986	0.2238	0.4315	0.3553	0.8073	0.2031	0.741	0.259	0.7969
24	424	83.73	16.27	21.45	71.03	0.2605	0.3196	0.7875	0	0.7757	0.0698	0.7651	0.2349	0.9302
25	315	72.11	27.89	18.27	71.66	0.3592	0.4349	0.9302	0.0864	0.6479	0.0458	0.5337	0.4663	0.9542
28	113	76.19	23.81	17.73	75.83	0.3559	0.2885	0.6171	0.5342	0.8597	0.216	0.7456	0.2544	0.784
29	141	66.1	33.9	17.17	81.04	0.4651	0.5667	1.7196	0.9103	0.8664	0.3701	0.6845	0.3155	0.6299
30	191	75.85	24.15	19.08	69.77	0.3111	0.3149	0.8545	0.2237	0.7523	0.0075	0.5625	0.4375	0.9925

Además de las relaciones entre los clústers de genes y las variables clínico-patológicas, señaladas en la tabla 6.11, los clústers 4 y 12 se asocian, con valores de sensibilidad y especificidad superiores al 65%, con las siguientes variables,

- Clúster 4 asociado con ER- y además con PAM50 basal y DLDA30 pCR
- Clúster12 asociado con ER+ y además con, PR+, grado 2, PAM50 LumA y DLDA30 RD

Hemos estudiado también el grado de asociación entre los arrays clasificados como *outliers* para los diferentes grupos de genes. En la tabla 6.12 se muestra el porcentaje de concordancia en cada grupo con más de 10 genes. Los grupos que muestran más solapamiento en la dimensión de arrays son, los clústers 5, 9, 15 y 23.

Tabla 6.12. *Dataset* de Cáncer de Mama I. Concordancia entre arrays clasificados como *outliers* en cada clúster de genes. Los colores indican grupos con porcentaje de concordancia superior al 60%.

k	# arrays outlier	k																				
		1	2	4	5	9	10	11	12	13	15	18	19	20	21	22	23	24	25	28	29	30
1	29	100	3.45	0	24.14	58.62	0	0	89.66	0	17.24	3.45	24.14	3.45	6.9	10.34	58.62	0	3.45	72.41	93.1	41.38
2	28	3.57	100	3.57	100	82.14	96.43	10.71	46.43	78.57	82.14	3.57	60.71	0	3.57	60.71	78.57	0	0	57.14	25	0
4	85	0	1.18	100	10.59	7.06	15.29	8.24	0	10.59	5.88	0	11.76	40	74.12	27.06	5.88	5.88	0	8.24	34.12	0
5	72	9.72	38.89	12.5	100	62.5	70.83	5.56	40.28	55.56	77.78	2.78	58.33	1.39	11.11	44.44	68.06	0	0	43.06	31.94	1.39
9	64	26.56	35.94	9.38	70.31	100	39.06	1.56	48.44	28.12	57.81	1.56	57.81	0	7.81	37.5	82.81	0	0	65.62	60.94	6.25
10	86	0	31.4	15.12	59.3	29.07	100	9.3	39.53	65.12	58.14	4.65	39.53	1.16	12.79	54.65	41.86	3.49	9.3	23.26	13.95	5.81
11	44	0	6.82	15.91	9.09	2.27	18.18	100	72.73	47.73	0	36.36	43.18	52.27	34.09	9.09	0	68.18	72.73	2.27	6.82	47.73
12	146	17.81	8.9	0	19.86	21.23	23.29	21.92	100	23.29	15.75	21.23	25.34	12.33	10.96	11.64	22.6	26.03	52.05	22.6	30.82	47.95
13	70	0	31.43	12.86	57.14	25.71	80	30	48.57	100	47.14	11.43	65.71	5.71	11.43	50	35.71	18.57	21.43	21.43	10	10
15	63	7.94	36.51	7.94	88.89	58.73	79.37	0	36.51	52.38	100	0	50.79	0	6.35	57.14	73.02	0	0	39.68	25.4	1.59
18	33	3.03	3.03	0	6.06	3.03	12.12	48.48	93.94	24.24	0	100	18.18	18.18	39.39	0	3.03	30.3	84.85	0	6.06	48.48
19	75	9.33	22.67	13.33	56	49.33	45.33	25.33	49.33	61.33	42.67	8	100	10.67	9.33	30.67	41.33	22.67	21.33	38.67	37.33	14.67
20	58	1.72	0	58.62	1.72	0	1.72	39.66	31.03	6.9	0	10.34	13.79	100	75.86	18.97	0	44.83	34.48	3.45	24.14	27.59
21	92	2.17	1.09	68.48	8.7	5.43	11.96	16.3	17.39	8.7	4.35	14.13	7.61	47.83	100	23.91	6.52	16.3	18.48	2.17	21.74	9.78
22	88	3.41	19.32	26.14	36.36	27.27	53.41	4.55	19.32	39.77	40.91	0	26.14	12.5	25	100	39.77	7.95	2.27	32.95	29.55	1.14
23	69	24.64	31.88	7.25	71.01	76.81	52.17	0	47.83	36.23	66.67	1.45	44.93	0	8.7	50.72	100	0	0	53.62	46.38	8.7
24	48	0	0	10.42	0	0	6.25	62.5	79.17	27.08	0	20.83	35.42	54.17	31.25	14.58	0	100	81.25	4.17	12.5	68.75
25	82	1.22	0	0	0	0	9.76	39.02	92.68	18.29	0	34.15	19.51	24.39	20.73	2.44	0	47.56	100	1.22	8.54	64.63
28	70	30	22.86	10	44.29	60	28.57	1.43	47.14	21.43	35.71	0	41.43	2.86	2.86	41.43	52.86	2.86	1.43	100	84.29	11.43
29	100	27	7	29	23	39	12	3	45	7	16	2	28	14	20	26	32	6	7	59	100	18
30	71	16.9	0	0	1.41	5.63	7.04	29.58	98.59	9.86	1.41	22.54	15.49	22.54	12.68	1.41	8.45	46.48	74.65	11.27	25.35	100

En la tabla 6.13 se muestra el resultado del enriquecimiento funcional de los clústers con más de 10 genes y con alguna categoría GO-BP y/o KEGG sobre-representada (pvalor FDR ≤ 0.01). El clúster 4 y el clúster 12, asociados con la variable ER, no tienen ninguna categoría significativamente sobre-representada en estas ontologías. Sin embargo, aparecen muy relacionados con la componente celular GO:0031974 ~ *membrane-enclosed lumen*.

Otros clústers de cierta *calidad* en el sentido antes mencionado, bastante relevantes por su tamaño y sin categorías enriquecidas, son el 13, 15 y 19. En todos ellos, aparecen sobre-representadas categorías GO de otras ontologías. En Componentes Celulares, destacan GO: 0044459~*plasma membrane part* y GO:004421~*extracellular region part*, muy relacionadas con el clúster 15, y GO:0031226~*intrinsic to plasma membrane*, con el clúster 19. En la ontología Función Molecular, aparecen enriquecidas GO:0043565~*sequence specific DNA binding*, en el clúster 13 y GO:0022838~*substrate specific cannel activity*, en el 19.

Tabla 6.13. *Dataset* de Cáncer de Mama I. Análisis de enriquecimiento funcional de los genes agrupados en los clústers con más de 10 genes. Se muestran las categorías con un p-valor ≤ 0.01 tras la corrección FDR.

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
5	540	GO:0007267~cell-cell signaling	63	hsa04080:Neuroactive ligand-receptor interaction	38
		GO:0007268~synaptic transmission	35		
		GO:0019226~transmission of nerve impulse	38		
		GO:0044057~regulation of system process	33		
		GO:0003013~circulatory system process	25		
		GO:0008015~blood circulation	25		
		GO:0050877~neurological system process	75		
		GO:0007610~behavior	40		
		GO:0008016~regulation of heart contraction	15		
		GO:0035270~endocrine system development	14		
		GO:0006811~ion transport	52		
		GO:0030182~neuron differentiation	36		
		GO:0042127~regulation of cell proliferation	51		
		GO:0045165~cell fate commitment	18		
		GO:0007626~locomotory behavior	26		
		GO:0008284~positive regulation of cell proliferation	33		
		GO:0008217~regulation of blood pressure	15		
		GO:0019932~second-messenger-mediated signaling	23		
		GO:0048878~chemical homeostasis	37		
9	312	GO:0007610~behavior	30		
		GO:0007611~learning or memory	15		
		GO:0007268~synaptic transmission	22		
		GO:0007267~cell-cell signaling	31		
		GO:0019226~transmission of nerve impulse	23		
		GO:0007612~learning	10		
10	238	GO:0007610~behavior	25		
		GO:0006955~immune response	29		
		GO:0015669~gas transport	6		
20	419	GO:0006396~RNA processing	60	hsa03040:Spliceosome	25
		GO:0016071~mRNA metabolic process	42		
		GO:0006397~mRNA processing	39		
		GO:0008380~RNA splicing	36		
		GO:0022613~ribonucleoprotein complex biogenesis	25		
		GO:0044265~cellular macromolecule catabolic process	50		
		GO:0009057~macromolecule catabolic process	51		

Tabla 6.13. Continuación I

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
20	419	GO:0042254~ribosome biogenesis	18		
		GO:0000375~RNA splicing, via transesterification reactions	20		
		GO:0000377~RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	20		
		GO:0000398~nuclear mRNA splicing, via spliceosome	20		
		GO:0043933~macromolecular complex subunit organization	46		
		GO:0006403~RNA localization	16		
		GO:0065003~macromolecular complex assembly	43		
		GO:0019941~modification-dependent protein catabolic process	39		
		GO:0043632~modification-dependent macromolecule catabolic process	39		
		GO:0051603~proteolysis involved in cellular protein catabolic process	40		
		GO:0044257~cellular protein catabolic process	40		
		GO:0051236~establishment of RNA localization	15		
		GO:0050658~RNA transport	15		
		GO:0050657~nucleic acid transport	15		
		GO:0006511~ubiquitin-dependent protein catabolic process	23		
		GO:0030163~protein catabolic process	40		
		GO:0015931~nucleobase, nucleoside, nucleotide and nucleic acid transport	15		
		GO:0034660~ncRNA metabolic process	21		
		GO:0070271~protein complex biogenesis	33		
		GO:0006461~protein complex assembly	33		
21	132	GO:0007049~cell cycle	58	hsa04110:Cell cycle	17
		GO:0022403~cell cycle phase	46	hsa03030:DNA replication	9
		GO:0000279~M phase	42		
		GO:0022402~cell cycle process	48		
		GO:0000278~mitotic cell cycle	41		
		GO:0007067~mitosis	33		
		GO:0000280~nuclear division	33		
		GO:0000087~M phase of mitotic cell cycle	33		
		GO:0048285~organelle fission	33		
		GO:0051301~cell division	29		
		GO:0006260~DNA replication	22		
		GO:0006259~DNA metabolic process	30		
		GO:0000070~mitotic sister chromatid segregation	12		
		GO:0000819~sister chromatid segregation	12		
		GO:0007059~chromosome segregation	13		
		GO:0051726~regulation of cell cycle	21		
		GO:0000075~cell cycle checkpoint	13		
		GO:0051276~chromosome organization	24		
		GO:0000226~microtubule cytoskeleton organization	14		
		GO:0006261~DNA-dependent DNA replication	10		

Tabla 6.13. Continuación II

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
21	132	GO:0007051~spindle organization	9		
		GO:0007017~microtubule-based process	16		
		GO:0006974~response to DNA damage stimulus	18		
		GO:0010564~regulation of cell cycle process	11		
		GO:0007346~regulation of mitotic cell cycle	12		
		GO:0006302~double-strand break repair	8		
		GO:0007126~meiosis	9		
		GO:0051327~M phase of meiotic cell cycle	9		
		GO:0030261~chromosome condensation	6		
		GO:0051321~meiotic cell cycle	9		
		GO:0007076~mitotic chromosome condensation	5		
		GO:0031570~DNA integrity checkpoint	7		
		GO:0006281~DNA repair	13		
		GO:0033554~cellular response to stress	18		
22	271	GO:0006955~immune response	76	hsa04650:Natural killer cell mediated cytotoxicity	21
		GO:0045321~leukocyte activation	45	hsa04640:Hematopoietic cell lineage	16
		GO:0001775~cell activation	46	hsa05340:Primary immunodeficiency	10
		GO:0046649~lymphocyte activation	40	hsa04660:T cell receptor signaling pathway	15
		GO:0002684~positive regulation of immune system process	38	hsa04514:Cell adhesion molecules (CAMs)	16
		GO:0042110~T cell activation	30	hsa04060:Cytokine-cytokine receptor interaction	21
		GO:0006952~defense response	47		
		GO:0050865~regulation of cell activation	27		
		GO:0051249~regulation of lymphocyte activation	25		
		GO:0050867~positive regulation of cell activation	22		
		GO:0002694~regulation of leukocyte activation	25		
		GO:0050863~regulation of T cell activation	21		
		GO:0002696~positive regulation of leukocyte activation	20		
		GO:0048584~positive regulation of response to stimulus	27		
		GO:0051251~positive regulation of lymphocyte activation	19		
		GO:0030098~lymphocyte differentiation	19		
		GO:0050778~positive regulation of immune response	21		
		GO:0030097~hemopoiesis	25		
		GO:0050870~positive regulation of T cell activation	16		
		GO:0045619~regulation of lymphocyte differentiation	15		
		GO:0030217~T cell differentiation	15		
		GO:0002521~leukocyte differentiation	19		
		GO:0045621~positive regulation of lymphocyte differentiation	12		
		GO:0048534~hemopoietic or lymphoid organ development	25		
		GO:0002520~immune system development	25		
		GO:0045058~T cell selection	9		
		GO:0045580~regulation of T cell differentiation	12		
		GO:0002757~immune response-activating signal transduction	12		

Tabla 6.13. Continuación III

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
22	271	GO:0002252~immune effector process	17		
		GO:0002764~immune response-regulating signal transduction	12		
		GO:0045582~positive regulation of T cell differentiation	10		
		GO:0031349~positive regulation of defense response	13		
		GO:0006968~cellular defense response	12		
		GO:0006954~inflammatory response	24		
		GO:0042098~T cell proliferation	9		
		GO:0002429~immune response-activating cell surface receptor signaling pathway	10		
		GO:0042981~regulation of apoptosis	38		
		GO:0045088~regulation of innate immune response	11		
		GO:0009611~response to wounding	30		
		GO:0043067~regulation of programmed cell death	38		
		GO:0010941~regulation of cell death	38		
		GO:0046651~lymphocyte proliferation	10		
		GO:0002768~immune response-regulating cell surface receptor signaling pathway	10		
		GO:0045059~positive thymic T cell selection	6		
		GO:0045061~thymic T cell selection	7		
		GO:0032943~mononuclear cell proliferation	10		
		GO:0070661~leukocyte proliferation	10		
		GO:0045089~positive regulation of innate immune response	10		
		GO:0006935~chemotaxis	16		
		GO:0042330~taxis	16		
		GO:0001910~regulation of leukocyte mediated cytotoxicity	8		
		GO:0043368~positive T cell selection	6		
		GO:0002443~leukocyte mediated immunity	12		
		GO:0007166~cell surface receptor linked signal transduction	61		
		GO:0031341~regulation of cell killing	8		
		GO:0001817~regulation of cytokine production	16		
		GO:0002253~activation of immune response	12		
		GO:0007155~cell adhesion	32		
		GO:0022610~biological adhesion	32		
		GO:0007204~elevation of cytosolic calcium ion concentration	12		
		GO:0002449~lymphocyte mediated immunity	10		
		GO:0046635~positive regulation of alpha-beta T cell activation	7		
		GO:0007626~locomotory behavior	18		
		GO:0007242~intracellular signaling cascade	44		
		GO:0050900~leukocyte migration	9		
		GO:0051480~cytosolic calcium ion homeostasis	12		
		GO:0033077~T cell differentiation in the thymus	7		
		GO:0042113~B cell activation	10		
		GO:0045060~negative thymic T cell selection	5		

Tabla 6.13. Continuación IV

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
22	271	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	10		
		GO:0002250~adaptive immune response	10		
24	417	GO:0008380~RNA splicing	26		
		GO:0016071~mRNA metabolic process	30		
		GO:0006397~mRNA processing	26		
28	113	GO:0007155~cell adhesion	23		
		GO:0022610~biological adhesion	23		
		GO:0001944~vasculature development	12		
		GO:0001568~blood vessel development	11		
29	137	GO:0007155~cell adhesion	35	hsa04510:Focal adhesion	19
		GO:0022610~biological adhesion	35	hsa04512:ECM-receptor interaction	11
		GO:0001568~blood vessel development	23	hsa04610:Complement and coagulation cascades	8
		GO:0001944~vasculature development	23		
		GO:0030198~extracellular matrix organization	14		
		GO:0043062~extracellular structure organization	15		
		GO:0032101~regulation of response to external stimulus	13		
		GO:0031589~cell-substrate adhesion	11		
		GO:0016477~cell migration	16		
		GO:0048514~blood vessel morphogenesis	14		
		GO:0048870~cell motility	16		
		GO:0051674~localization of cell	16		
		GO:0001525~angiogenesis	12		
		GO:0001501~skeletal system development	16		
		GO:0040012~regulation of locomotion	13		
		GO:0051270~regulation of cell motion	13		
		GO:0006928~cell motion	19		
		GO:0030334~regulation of cell migration	12		
		GO:0007160~cell-matrix adhesion	9		
		GO:0040017~positive regulation of locomotion	9		
		GO:0051272~positive regulation of cell motion	9		
		GO:0042127~regulation of cell proliferation	22		

6.2.3.2.1. Influencia de los parámetros y comparación con reglas *boxplot*

En esta sección evaluamos el funcionamiento de la metodología para diferentes elecciones de parámetros y realizamos comparaciones con las soluciones que se obtienen cuando el *input* de este análisis proviene de la utilización de reglas tipo *boxplot*. En este análisis hemos modificado los niveles de recorte, tanto en la dimensión de genes como en la de arrays, y la definición de *outlier,* con diferentes elecciones de la constante f que mide la separación mínima de éstos al núcleo de expresión de cada gen.

En la tabla 6.14 se resumen los resultados obtenidos en cada uno de los escenarios considerados. En todos los casos, antes de aplicar el método, se seleccionan los genes cuyo número de muestras clasificadas como *outlier* supera la mediana de este valor (concretamente el percentil 52.7). En todos los casos se establece un recorte de α_G en la dimensión de genes y de α_A en la dimensión de arrays, con $K_G = 30$ grupos de genes y 1000 comienzos aleatorios.

Tabla 6.14. *Dataset* de Cáncer de Mama I. Comparación de soluciones modificando parámetros del método.

Definición outlier	Recorte	# clúster de genes	# genes / clúster		% de acuerdo en la clasificación de genes (media ± DT, ponderada por # genes)	$\hat{o}_{k2}$ (media ± DT, ponderada por # genes)	Relación con source y ER			
			mediana	rango			categoría	Cl: # genes	Sens	Esp
$f = 1$ $X_o = 80$ 6629 genes	$\alpha_G = 0$ $\alpha_a = 0$	26	182	62 - 968	65.88 ± 16.55	70.84 ± 3.46	Source: ISPY	10: 118	0.8795	0.8062
								17: 94	0.7831	0.9648
							ER+	4: 136	0.8352	0.9552
							ER-	6: 62	0.7910	0.8125
								27: 289	0.6940	0.9545
	$\alpha_G = 0.2$ $\alpha_a = 0$	26	140.5	26 - 807	65.87 ± 21.87	72.12 ± 2.89	Source: ISPY	13: 176	0.8313	0.8546
							ER+	18: 128	0.7670	0.9627
							ER-	11: 148	0.7313	0.9659
	$\alpha_G = 0$ $\alpha_a = 0.05$	29	183	22 - 832	63.9 ± 20.9	72.26 ± 4.06	Source: ISPY	23: 185	0.9103	0.8796
							ER+	16: 133	0.8580	0.9545
								21: 22	0.6647	0.9200
							ER-	8: 206	0.8049	0.9708
	$\alpha_G = 0.2$ $\alpha_a = 0.05$	21	249	60 - 577	100 ± 0	73.15 ± 3.84	Source: ISPY	29: 141	0.9103	0.8664
							ER+	12: 110	0.8589	0.9545
							ER-	4: 294	0.6639	0.9657
$f = 2$ $X_o = 36$ 6035 genes	$\alpha_G = 0$ $\alpha_a = 0$	2	3017.5	1256 - 4779	NC	71.93 ± NA	-	-	-	-
	$\alpha_G = 0.2$ $\alpha_a = 0$	2	2414	1161 - 3667	NC	70.94 ± NA	-	-	-	-
	$\alpha_G = 0$ $\alpha_a = 0.05$	23	91	19 - 2342	39.15 ± 3.27	74.31 ± 3.96	Source: ISPY	15: 96	0.7162	0.9276
							ER+	12: 63	0.6988	0.9627
	$\alpha_G = 0.2$ $\alpha_a = 0.05$	13	195	54 - 2021	40.83 ± 2.86	75.65 ± 4.3	Source: ISPY	25: 106	0.5143	0.9378
Boxplot rule $X_o = 53$ 6882 genes	$\alpha_G = 0$ $\alpha_a = 0$	2	3441	1662 - 5220	NC	69.45 ± NA	-	-	-	-
	$\alpha_G = 0.2$ $\alpha_a = 0$	2	2851	1995 - 3707	NC	68.17 ± NA	-	-	-	-
	$\alpha_G = 0$ $\alpha_a = 0.05$	20	168	55 - 2158	32 ± 3.04	74.96 ± 3.99	-	-	-	-
	$\alpha_G = 0.2$ $\alpha_a = 0.05$	18	128.5	29 - 1149	30.06 ± 3.58	74.47 ± 4.34	-	-	-	-

NC = soluciones no comparables; X_o = número mínimo de arrays *outliers* por gen

De cada solución se muestra el número de clústers de genes junto a la mediana y el rango del tamaño de éstos; la media del porcentaje de muestras *outlier* en cada grupo, denotado por $\hat{o}_{k2}$; y el porcentaje medio de acuerdo respecto de la clasificación de genes obtenida en la solución descrita anteriormente. Este porcentaje de acuerdo se calcula utilizando los grupos de genes que muestran mayor grado de coincidencia en las dos soluciones comparadas. Tanto este porcentaje, como $\hat{o}_{k2}$, puede verse muy afectados por el tamaño de los grupos, por lo que el resultado aparece ponderado por el número de genes de cada clúster. Otro aspecto que ilustramos en la tabla es la aparición o no de biclústers en la solución relacionados con las variables clínicas *source* y ER.

Cuando tomamos como *input* los resultados del recorte imparcial, el grado de dependencia de los niveles de recorte en el resultado obtenido es bajo Se observa que, al aumentar estos niveles de recorte, especialmente en la dimensión de arrays, aumenta la calidad de los clústers. En todos los escenarios se identifican clústers de genes asociados con las variables *source* y ER. Sin embargo, cuando se utiliza una definición de *outlier* más estricta, aumentando el valor de la constante f en (6.33), añadir recorte en el nivel de arrays es fundamental para encontrar grupos relevantes. En los dos casos en los que se fija $\alpha_A = 0$, la solución únicamente tiene dos grupos de genes muy grandes, uno de ellos sin grupo de arrays *outlier*. Cuando sólo se utiliza el recorte en la dimensión de arrays se establecen grupos relacionados con *source* y ER+, pero no con ER-. Cuando además se considera recorte en la dimensión de genes, dejan de obtenerse los grupos relacionados con estas variables, aunque se obtiene un clúster con valores de sensibilidad y especificidad para la categoría ISPY superiores al 50%.

Este procedimiento de identificación de patrones atípicidad se podría utilizar partiendo de definiciones de outlier obtenidas utilizando la regla *boxplot*.
La definición equivalente a (6.33) sería,

$$y_{gj} \geq q_3 + f_{boxplot} \cdot IQR \qquad (6.43)$$

donde q_3 es el percentil 75, IQR el rango inter-cuartílico y $f_{boxplot} = 0.25$, que aproximadamente se corresponde con la constante $f = 1$ en el caso de una normal estándar. En la tabla anterior evaluamos como afecta este cambio en el input a la solución obtenida. Al igual que ocurría cuando en (6.33) se fijaba en $f = 2$, añadir recorte en el nivel de arrays es fundamental para encontrar grupos relevantes. En los casos sin recorte en esta dimensión, la solución únicamente tiene dos grupos de genes. Cabe destacar que en este caso, con ninguno de los escenarios propuestos, se obtienen grupos relacionados con las variables *source* y ER.

6.2.4. Aplicación al *dataset* de Cáncer de Mama II

En este *dataset*, propuesto en **[Miller et al, 2005]** y cuyas características se resumen en la sección B.4 del apéndice B, se miden 12576 genes en 251 muestras de cáncer de mama utilizando la misma plataforma que en el *dataset* de Cáncer de Mama I.

6.2.4.1. Búsqueda de patrones de *outliers*

El objetivo es comprobar el funcionamiento del método de los $K_G x2$ clústers de *outliers*, descrito en 6.1.3, en otro conjunto de datos relacionado también con el cáncer de mama. En este *dataset* se dispone de una batería de variables clínicas similar a las medidas en el conjunto Cáncer de Mama I, por lo que va a ser posible evaluar la eficacia del método para encontrar clústers asociados con algunas de estas características.

Tabla 6.15. *Dataset* de Cáncer de Mama II. Grupos con más de 10 genes obtenidos con el método $K_G x2$ clúster de *outlier* en 6401 genes.

k	# gen	Peso de los clúster de arrays		% de outliers		$\overline{\rho}(\mathrm{Y}_{IA})$	Diferencia medianas de expresión	
		$\hat{\pi}_{k1}$	$\hat{\pi}_{k2}$	$\hat{o}_{k1}$	$\hat{o}_{k2}$		P_{20}	P_{80}
1	123	78.66	21.34	17.29	72.96	0.3852	0.3113	0.8526
2	41	65.97	34.03	11.31	82.32	0.666	0.7503	1.3610
4	69	77.31	22.69	15.8	78.21	0.4455	0.4571	0.8069
5	19	83.06	16.94	18.29	72.68	0.2925	0.1932	0.3036
6	191	86.13	13.87	20.41	67.27	0.2064	0.2551	0.5552
8	69	79.92	20.08	16.28	74.64	0.3874	0.3015	0.8339
9	103	80.67	19.33	16.86	73.58	0.4039	0.1630	0.2760
12	95	84.87	15.13	19.71	67.51	0.2403	0.2376	0.3943
14	151	84.87	15.13	19.46	68.25	0.2983	0.2084	0.4780
15	45	83.33	16.67	18.3	73.28	0.3067	0.3306	0.6379
16	163	83.26	16.74	19.3	68.28	0.2577	0.2872	0.5195
17	192	89.5	10.5	21.14	68.19	0.1786	0.1417	0.2878
18	60	80.83	19.17	18.01	72.28	0.3293	0.2231	0.6550
19	217	81.93	18.07	18.97	67.67	0.3011	0.2766	0.5059
22	186	78.66	21.34	18.65	71.74	0.3395	0.2692	0.8257
23	726	76.47	23.53	15.87	72.54	0.4322	0.1658	0.2901
24	318	70.59	29.41	16.48	74.37	0.4399	0.3564	0.7974
25	248	87.82	12.18	21.35	68.6	0.2027	0.2791	0.6068
27	638	80.25	19.75	16.99	71.79	0.3786	0.1622	0.2857
28	324	74.79	25.21	15.72	79.55	0.5176	0.3151	0.9917
29	331	91.6	8.4	22.08	73.22	0.1794	0.2582	0.7746
30	221	80.67	19.33	17.17	72.72	0.3728	0.1534	0.2708

Se aplica el método en el sub-conjunto de 6401 genes (50.9% del total de genes) con más de 60 arrays clasificados como *oulier* por arriba, definidos de la misma forma que para el *set* anterior: según la expresión (6.33) con $\alpha = 0.45$ y $f = 1$. También los parámetros del

método, se fijan en los mismos valores que en el primer conjunto: recorte $\alpha_G = 0.2$ en la dimensión de genes y $\alpha_A = 0.05$ en la dimensión de arrays, con $K_G = 30$ grupos de genes y 1000 comienzos aleatorios. En este caso se obtienen los clústers resumidos en la tabla 6.15.

El porcentaje de muestras clasificadas como *outlier* en los biclústers relacionados con los arrays atípicos (con $j = 2$), es alto, variando entre el 67% y 82%, mientras que en los grupos relacionados con los arrays no-*outlier*, ($j = 1$), estos porcentajes son inferiores al 22%. Por lo tanto, en términos de *calidad*, estos clústers son similares a los obtenidos en el *dataset* de Cáncer de Mama I. En cuanto a la diferencia de medianas, destaca el clúster 2 como el grupo de genes con un percentil 20 para las diferencias en medianas de expresión, entre las muestras clasificadas como *outliers* y las clasificadas como no-*outliers, de 0.75*.

En la tabla 6.16 se muestra la relación de la partición encontrada con las variables clínico-patológicas. Se encuentran dos clústers relevantes, con una sensibilidad y especificidad superiores al 65%,

- Clúster 24 asociado con la variable p53-, p53 DLDA 1, DLDA error 1, Elston G3, ER- y PgR+.
- Clúster 2 asociado con DLDA error 1.

Tabla 6.16. *Dataset* de Cáncer de Mama II. Relación de las variables clínico-patológicas con los clústers de más de 10 genes obtenidos con el método $K_G x 2$ clúster de *outlier* en 6401 genes.

k	Variable	Categoría	# arrays	Sens	Esp
24	p53	p53-	193	0.7115	0.8226
		p53+	58	0.1774	0.2885
	p53 DLDA	0	179	0.0824	0.1765
		1	72	0.8235	0.9176
	DLDA error	0	213	0.2304	0.3235
		1	38	0.6765	0.7696
	Elston	G?	2	0	0.7034
		G1	67	0.0154	0.6012
		G2	128	0.2049	0.6121
		G3	54	0.898	0.8624
	ER	ER-	34	0.7879	0.7854
		ER?	4	0	0.7009
		ER+	213	0.2189	0.2973
	PgR	PgR-	61	0.1713	0.3158
		PgR+	190	0.6842	0.8287
2	DLDA error	0	213	0.278	0.2727
		1	38	0.7273	0.722

En la tabla 6.17 se muestra el resultado del enriquecimiento funcional de los clústers 2 y 24, junto con otros clústers relevantes, tanto por su calidad, como por las diferencias en los niveles de expresión. Al igual que ocurría en el caso anterior, hay algún clúster bastante claro que no muestra enriquecimiento con las categorías GO-BP y/o KEGG. Llama la atención el clúster 4 formado por 69 genes con un bloque de 57 arrays clasificados como *outlier*, $\hat{o}_{k2} = 78.21$ y el percentil 20 de la diferencia de medianas de expresión en 0.4571. La categoría más sobre-representada en este clúster es GO:0044237~*celluar metabolic process*, con un p-valor de 0.038 tras la corrección FDR y 40 genes anotados en ella. A pesar de que no resulta significativo al nivel fijado, 0.01, sí podría indicar cierta coherencia biológica entre los genes que forman este grupo.

Tabla 6.17. *Dataset* de Cáncer de Mama II. Análisis de enriquecimiento funcional de los genes agrupados en *clústers* relevantes. Se muestran las categorías GO-BP y KEGG sobre-representadas con un p-valor ≤ 0.01 tras la corrección FDR.

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
2	39	GO:0009615~response to virus	13		
		GO:0006955~immune response	14		
		GO:0006952~defense response	11		
8	67	GO:0048584~positive regulation of response to stimulus	11		
		GO:0050778~positive regulation of immune response	9		
		GO:0045321~leukocyte activation	10		
		GO:0001817~regulation of cytokine production	9		
		GO:0002757~immune response-activating signal transduction	6		
		GO:0001775~cell activation	10		
		GO:0002764~immune response-regulating signal transduction	6		
		GO:0002684~positive regulation of immune system process	9		
22	183	GO:0010033~response to organic substance	29		
		GO:0001568~blood vessel development	16		
		GO:0001944~vasculature development	16		
		GO:0002237~response to molecule of bacterial origin	10		
		GO:0009719~response to endogenous stimulus	19		
		GO:0048514~blood vessel morphogenesis	14		
		GO:0009725~response to hormone stimulus	18		
		GO:0001525~angiogenesis	12		
		GO:0042493~response to drug	14		
		GO:0032496~response to lipopolysaccharide	9		
24	313	GO:0007049~cell cycle	99	hsa04110:Cell cycle	28
		GO:0000278~mitotic cell cycle	73	hsa03030:DNA replication	13
		GO:0022402~cell cycle process	85	hsa03050:Proteasome	10
		GO:0022403~cell cycle phase	75	hsa04115:p53 signaling pathway	11
		GO:0000279~M phase	66		

Tabla 6.17. Continuación I

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
24	313	GO:0000087~M phase of mitotic cell cycle	51		
		GO:0007067~mitosis	50		
		GO:0000280~nuclear division	50		
		GO:0048285~organelle fission	50		
		GO:0051301~cell division	49		
		GO:0006259~DNA metabolic process	56		
		GO:0006260~DNA replication	37		
		GO:0051726~regulation of cell cycle	39		
		GO:0007017~microtubule-based process	30		
		GO:0006261~DNA-dependent DNA replication	16		
		GO:0007059~chromosome segregation	18		
		GO:0010564~regulation of cell cycle process	20		
		GO:0007051~spindle organization	14		
		GO:0006974~response to DNA damage stimulus	33		
		GO:0051439~regulation of ubiquitin-protein ligase activity during mitotic cell cycle	16		
		GO:0051276~chromosome organization	37		
		GO:0000226~microtubule cytoskeleton organization	21		
		GO:0007346~regulation of mitotic cell cycle	21		
		GO:0000075~cell cycle checkpoint	17		
		GO:0051438~regulation of ubiquitin-protein ligase activity	16		
		GO:0051340~regulation of ligase activity	16		
		GO:0006281~DNA repair	27		
		GO:0031396~regulation of protein ubiquitination	17		
		GO:0031145~anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	14		
		GO:0051437~positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	14		
		GO:0051443~positive regulation of ubiquitin-protein ligase activity	14		
		GO:0000070~mitotic sister chromatid segregation	11		
		GO:0051351~positive regulation of ligase activity	14		
		GO:0000819~sister chromatid segregation	11		
		GO:0031397~negative regulation of protein ubiquitination	14		
		GO:0051436~negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	13		
		GO:0051352~negative regulation of ligase activity	13		
		GO:0051444~negative regulation of ubiquitin-protein ligase activity	13		
		GO:0051327~M phase of meiotic cell cycle	15		
		GO:0007126~meiosis	15		
		GO:0031398~positive regulation of protein ubiquitination	14		
		GO:0051321~meiotic cell cycle	15		
		GO:0031400~negative regulation of protein modification process	16		
		GO:0051329~interphase of mitotic cell cycle	15		

Tabla 6.17. Continuación II

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
24	313	GO:0051325~interphase	15		
		GO:0033554~cellular response to stress	34		
		GO:0010498~proteasomal protein catabolic process	14		
		GO:0043161~proteasomal ubiquitin-dependent protein catabolic process	14		
		GO:0032269~negative regulation of cellular protein metabolic process	18		
		GO:0008283~cell proliferation	28		
		GO:0051248~negative regulation of protein metabolic process	18		
		GO:0032270~positive regulation of cellular protein metabolic process	20		
		GO:0051247~positive regulation of protein metabolic process	20		
		GO:0031570~DNA integrity checkpoint	10		
		GO:0034621~cellular macromolecular complex subunit organization	24		
		GO:0031401~positive regulation of protein modification process	17		
28	314	GO:0006955~immune response	96	hsa04060:Cytokine-cytokine receptor interaction	34
		GO:0046649~lymphocyte activation	43	hsa05340:Primary immunodeficiency	14
		GO:0002684~positive regulation of immune system process	45	hsa04650:Natural killer cell mediated cytotoxicity	22
		GO:0045321~leukocyte activation	45	hsa04640:Hematopoietic cell lineage	18
		GO:0001775~cell activation	47	hsa04660:T cell receptor signaling pathway	19
		GO:0042110~T cell activation	32	hsa04062:Chemokine signaling pathway	22
		GO:0051249~regulation of lymphocyte activation	32	hsa05330:Allograft rejection	10
		GO:0006952~defense response	57	hsa04514:Cell adhesion molecules (CAMs)	16
		GO:0050865~regulation of cell activation	33		
		GO:0002694~regulation of leukocyte activation	32		
		GO:0050867~positive regulation of cell activation	26		
		GO:0050863~regulation of T cell activation	26		
		GO:0002696~positive regulation of leukocyte activation	25		
		GO:0051251~positive regulation of lymphocyte activation	24		
		GO:0050778~positive regulation of immune response	26		
		GO:0050870~positive regulation of T cell activation	19		
		GO:0030098~lymphocyte differentiation	20		
		GO:0048584~positive regulation of response to stimulus	28		
		GO:0002521~leukocyte differentiation	21		
		GO:0050671~positive regulation of lymphocyte proliferation	15		
		GO:0070665~positive regulation of leukocyte proliferation	15		
		GO:0032946~positive regulation of mononuclear cell proliferation	15		
		GO:0050670~regulation of lymphocyte proliferation	17		
		GO:0070663~regulation of leukocyte proliferation	17		
		GO:0032944~regulation of mononuclear cell proliferation	17		
		GO:0006954~inflammatory response	30		
		GO:0002768~immune response-regulating cell surface receptor signaling pathway	13		

Tabla 6.17. Continuación III

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
28	314	GO:0050871~positive regulation of B cell activation	12		
		GO:0030217~T cell differentiation	15		
		GO:0002253~activation of immune response	17		
		GO:0030097~hemopoiesis	25		
		GO:0048534~hemopoietic or lymphoid organ development	26		
		GO:0009611~response to wounding	37		
		GO:0002697~regulation of immune effector process	17		
		GO:0002429~immune response-activating cell surface receptor signaling pathway	12		
		GO:0050864~regulation of B cell activation	13		
		GO:0002520~immune system development	26		
		GO:0050851~antigen receptor-mediated signaling pathway	11		
		GO:0002764~immune response-regulating signal transduction	13		
		GO:0006968~cellular defense response	13		
		GO:0002757~immune response-activating signal transduction	12		
		GO:0046651~lymphocyte proliferation	11		
		GO:0043067~regulation of programmed cell death	43		
		GO:0070661~leukocyte proliferation	11		
		GO:0032943~mononuclear cell proliferation	11		
		GO:0010941~regulation of cell death	43		
		GO:0045621~positive regulation of lymphocyte differentiation	10		
		GO:0042981~regulation of apoptosis	42		
		GO:0046635~positive regulation of alpha-beta T cell activation	9		
		GO:0042098~T cell proliferation	9		
		GO:0045619~regulation of lymphocyte differentiation	12		
		GO:0045058~T cell selection	8		
		GO:0045580~regulation of T cell differentiation	11		
		GO:0002252~immune effector process	16		
		GO:0050848~regulation of calcium-mediated signaling	8		
		GO:0042330~taxis	17		
		GO:0006935~chemotaxis	17		
		GO:0001817~regulation of cytokine production	18		
		GO:0045582~positive regulation of T cell differentiation	9		
		GO:0002703~regulation of leukocyte mediated immunity	11		
		GO:0002250~adaptive immune response	12		
		GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	12		
		GO:0042129~regulation of T cell proliferation	11		
		GO:0046634~regulation of alpha-beta T cell activation	9		
		GO:0002637~regulation of immunoglobulin production	8		
		GO:0050850~positive regulation of calcium-mediated signaling	7		

Tabla 6.17. Continuación IV

k	# genes anotados	GO biological process		KEGG pathway	
		Categoría	# genes	Categoría	# genes
28	314	GO:0019882~antigen processing and presentation	12		
		GO:0042102~positive regulation of T cell proliferation	9		
		GO:0002285~lymphocyte activation during immune response	7		
		GO:0002706~regulation of lymphocyte mediated immunity	10		
		GO:0002700~regulation of production of molecular mediator of immune response	9		
		GO:0002822~regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	10		
		GO:0002819~regulation of adaptive immune response	10		
		GO:0030890~positive regulation of B cell proliferation	7		
		GO:0046638~positive regulation of alpha-beta T cell differentiation	7		
		GO:0007626~locomotory behavior	20		
		GO:0019835~cytolysis	7		
		GO:0045087~innate immune response	14		
		GO:0050852~T cell receptor signaling pathway	7		
		GO:0007166~cell surface receptor linked signal transduction	65		
		GO:0002263~cell activation during immune response	8		
		GO:0002366~leukocyte activation during immune response	8		

Capítulo 7

Conclusiones

Esta Tesis realiza una prospectiva en la aplicación de procedimientos de recorte imparcial al análisis de datos en matrices de expresión génica. Las siguientes líneas incluyen las conclusiones de la investigación realizada.

- El estimador smart basado en el MCD es útil para identificar el núcleo de expresión típica de un gen. Las propiedades de este estimador aportadas en esta tesis facilitan la obtención de un algoritmo eficiente, que reduce considerablemente el coste computacional del disponible actualmente, permitiendo la aplicación intensiva requerida en su aplicación al análisis de datos de expresión génica.

- Las adaptaciones propuestas de los procedimientos de recorte imparcial a muestras provenientes de múltiples tejidos humanos sanos son útiles para identificar genes constitutivos o *housekeeping*, genes selectivos y genes específicos.

- Los procedimientos de recorte imparcial son útiles para detectar expresión diferencial en problemas de dos muestras con expresión heterogénea, respondiendo tanto al caso de genes con expresión predominantemente diferencial como al caso de genes cuya expresión diferencial se caracteriza por un grupo minoritario de muestras.

- La identificación a través de reglas basadas en el recorte imparcial de grupos de genes que muestran patrones comunes en cuanto a sus comportamientos atípicos, permite relacionar estos patrones de expresión diferencial con las características clínico-patológicas disponibles e incluso la posible identificación de variables ocultas.

- Los procedimientos de recorte imparcial pueden ser adaptados para su aplicación en la búsqueda de agrupaciones de genes y de agrupaciones conjuntas de genes y arrays basadas en co-expresión.

Apéndice A

Demostración teorema 1

Conservando la notación del capítulo 3, $\varphi_{\Theta}(A)$ y $p_{\Theta}(A)$ son, respectivamente, las funciones de densidad y de probabilidad asignadas al conjunto A para una distribución normal con parámetros Θ, y π la proporción de observaciones contaminantes. Inicialmente enunciamos y demostramos la proposición 1, necesaria para la demostración del teorema 1.

Proposición 1. El estimador smart coincide con el estimador truncado del modelo normal restringido a $p_{\Theta}(A) \geq p_n(A)$ o con el censurado del modelo normal restringido a $p_{\Theta}(A) \leq p_n(A)$. Cuando el smart verifica $p_{\Theta}(A) = p_n(A)$ ambos coinciden.

Demostración. **[Cuesta Albertos et al, 2008]** demostraron que el estimador smart coincide con el estimador truncado restringido en $p_{\Theta}(A) \geq p_n(A)$ o, en caso contrario, coincide con el estimador censurado y la estimación de π es igual a 0. Se puede probar que cuando el smart no coincide con el truncado restringido en $p_{\Theta}(A) \geq p_n(A)$, el censurado tiene que verificar necesariamente $p_{\Theta}(A) < p_n(A)$. Vamos a suponer que esto no es cierto, es decir que $p_{\Theta}(A) \geq p_n(A)$, y llegaremos a una contradicción.

La función objetivo del smart tiene una representación como,

$$\begin{aligned}
&\sum_{i=1}^{n} \log\left(\varphi_{\theta}(x)\right) + p_n\left(A^c\right)\log\left(p_{\theta}\left(A^c\right)\right) \\
&+ p_n(A)\log\left((1-\pi)\,p_{\theta}(A)\right) + p_n\left(A^c\right)\log\left(1-(1-\pi)\,p_{\theta}(A)\right) \\
&- p_n(A)\log\left(p_{\theta}(A)\right) - p_n\left(A^c\right)\log\left(p_{\theta}\left(A^c\right)\right)
\end{aligned}$$

sujeto a $\pi > 0$.

Si el estimador censurado, que alcanza el máximo de la función anterior cuando aplicamos la restricción $\pi = 0$, verifica $p_{\Theta}(A) \geq p_n(A)$, entonces es posible conseguir un valor estrictamente mejor en la función objetivo anterior que el correspondiente a $(\theta_c, 0)$, el dado

por $\left(\theta_c, \pi^*\right)$ con $\pi^* = \frac{p_\theta(A) - p_n(A)}{p_\theta(A)}$, que también verifica la restricción $\pi > 0$, ya que hemos supuesto que el estimador censurado, θ_c, verificaba la restricción $p_\Theta(A) \geq p_n(A)$. Con lo que llegamos a contradicción con que $\left(\theta_c, 0\right)$ es óptimo.

Si el smart verifica $p_\Theta(A) = p_n(A)$, coincide con el truncado restringido a $p_\Theta(A) \geq p_n(A)$ y, por tanto, con el truncado restringido a $p_\Theta(A) = p_n(A)$. Bajo esta última restricción, el truncado y el censurado coinciden, ya que sus funciones objetivo admiten una representación equivalente.

Teorema 1 La estimación de localización y forma del smart-MCD coincide con la correspondiente estimación del MCD.

Demostración Supongamos que la estimación smart verifica $p_\Theta(A) > p_n(A)$, entonces, por **[Cuesta Albertos et al, 2008]**, el smart coincide con el truncado restringido a $p_\Theta(A) > p_n(A)$, que por alcanzar la solución en este abierto, el smart será un cero de la derivada de la log-verosimilitud truncada, por lo que verificará, adicionalmente a $p_\Theta(A) \leq p_n(A)$ las dos ecuaciones siguientes

$$E_\theta\left(x/A\right) = p\left(x/A\right) \tag{1}$$

$$Var_\theta\left(x/A\right) = Var\left(x/A\right) \tag{2}$$

donde $p\left(x/A\right)$ y $Var\left(x/A\right)$ son respectivamente, la media y la matriz de covarianzas correspondientes a las observaciones en el conjunto A, pero por estar basado el conjunto A en la solución del MCD, $p\left(x/A\right)$ y $Var\left(x/A\right)$ coinciden, respectivamente, con la media y matriz de covarianzas del MCD. Vamos a suponer, sin pérdida de generalidad (se puede llegar a esta representación con una estandarización), que ambos resúmenes son 0 e $\lambda_0 I$ y que el conjunto A es $\left(x'x \leq 1\right)$.

Por verificarse (2) λ_{0^P} es el determinante de la varianza condicionada en la bola unidad de una distribución $N\left(\mu, \Sigma\right)$. El determinante de una varianza condicionada para una

distribución normal está acotado superiormente por el correspondiente a una distribución uniforme en esa bola, u. Por tanto, λ_{0^P} pertenece al intervalo $(0,u)$.

Una distribución $N(0,\lambda I)$ verifica que su media condicionada al conjunto A es 0 y el determinante de su varianza condicionada en el conjunto A es $g(\lambda)I$, para g función continua, estrictamente creciente, con $g(0)=0$ y asíntota horizontal $y=u$. Por ello tiene que existir λ^* con $g(\lambda^*)=\lambda_{0^P}$. Se puede ver que el par $(0,\lambda^* I)$ verifica las ecuaciones (1) y (2) y, por la unicidad probada en **[Cuesta Albertos et al, 2008]**, $(0,\lambda^* I)$ coincide con el estimador smart. Con lo que probamos que cuando el smart verifica $p_\Theta(A) > p_n(A)$, entonces su localización y su matriz de forma coinciden con las correspondientes al estimador MCD.

Supongamos que la estimación smart verifica $p_\Theta(A) \le p_n(A)$, entonces, por la proposición 1 coincide con el estimador censurado restringido a $p_\Theta(A) \le p_n(A)$, que a su vez, por **[Cuesta Albertos et al, 2008]** coincide con el estimador censurado sin restricciones. Por ello, el estimador censurado será un cero de la derivada de la log-verosimilitud censurada y por tanto, verificará adicionalmente a $p_\Theta(A) \le p_n(A)$,

$$p_n(A)\,p\left(x/_A\right)+p_n\left(A^c\right)E_\theta\left(x/_{A^c}\right)=E_\theta(x) \tag{3}$$

$$p_n(A)\,Var\left(x/_A\right)+p_n\left(A^c\right)Var_\theta\left(x/_{A^c}\right)=Var_\theta(x) \tag{4}$$

Como antes, vamos a suponer sin pérdida de generalidad que $p\left(x/_A\right)$ y $Var\left(x/_A\right)$ son, respectivamente 0 e $\lambda_0 I$ y que el conjunto A es $(x'x \le 1)$.

Vamos a demostrar que existe una solución de (3) y (4) de la forma $(0,\lambda I)$. Es fácil comprobar que $(0,\lambda I)$ verifica (3). Para comprobar que (4) se verifica, derivado de que la matriz de forma lo verifica y de aplicar la descomposición de la varianza $Var_\theta(x)=E_\theta\left(Var_\theta\left(x/_A\right)\right)+Var_\theta\left(E_\theta\left(x/_A\right)\right)$, bastaría con comprobar que se verifica,

$$p_\theta(A)\left|Var_\theta\left(x/_A\right)\right|+\left(p_\theta\left(A^c\right)-p_n\left(A^c\right)\right)\left|Var_\theta\left(x/_{A^c}\right)\right|=p_n(A)\left|Var\left(x/_A\right)\right| \tag{5}$$

Por coincidir el smart con el censurado, el truncado no tiene un máximo local en $p_\Theta(A) > p_n(A)$, y por tanto, para parámetros θ de la forma $(0, \lambda I)$ en el conjunto $p_\Theta(A) > p_n(A)$, tiene que darse

$$\left|Var_\theta\left(x/_A\right)\right| < \left|Var\left(x/_A\right)\right| \tag{6}$$

porque si no existiría $(0, \lambda^* I)$ verificando $p_\Theta(A) > p_n(A)$ junto con (1) y (2) y tendríamos el máximo local, ya que $\left|Var_\theta\left(x/_A\right)\right|$ es continua y creciente como función de λ. Entonces, por (6) y por la continuidad de $\left|Var_\theta\left(x/_A\right)\right|$ tenemos $\inf_{\{\theta / p_\theta(A) \le p_n(A)\}} \left|Var_\theta\left(x/_A\right)\right| \le \left|Var\left(x/_A\right)\right|$, y derivado de ello que

$$\begin{aligned} &\inf_{\{\theta / p_\theta(A) \le p_n(A)\}} p_\theta(A)\left|Var_\theta\left(x/_A\right)\right| + \left(p_\theta\left(A^c\right) - p_n\left(A^c\right)\right)\left|Var_\theta\left(x/_{A^c}\right)\right| \\ &\le p_n(A)\left|Var\left(x/_A\right)\right| \end{aligned} \tag{7}$$

Por otro lado tenemos que,

$$\sup_{\{\theta / p_\theta(A) \le p_n(A)\}} p_\theta(A)\left|Var_\theta\left(x/_A\right)\right| + \left(p_\theta\left(A^c\right) - p_n\left(A^c\right)\right)\left|Var_\theta\left(x/_{A^c}\right)\right| = \infty \tag{8}$$

ya que $\lim_{p_\theta(A) \to 0} p_\theta(A)\left|Var_\theta\left(x/_A\right)\right| + \left(p_\theta\left(A^c\right) - p_n\left(A^c\right)\right)\left|Var_\theta\left(x/_{A^c}\right)\right| = \infty$.

Por (7) y (8) y por la continuidad de la función que aparece en el lado izquierdo de ambas expresiones tiene que existir θ^* del tipo $\theta^* = (0, \lambda^* I)$, verificando

$$p_{\theta^*}(A)\left|Var_{\theta^*}\left(x/_A\right)\right| + \left(p_{\theta^*}\left(A^c\right) - p_n\left(A^c\right)\right)\left|Var_{\theta^*}\left(x/_{A^c}\right)\right| = p_n(A)\left|Var\left(x/_A\right)\right|$$

Con lo que tenemos que (5) se verifica.

Como $\theta^* = (0, \lambda^* I)$ verifica las ecuaciones (3) y (4), por la unicidad probada en **[Cuesta Albertos et al, 2008]**, $(0, \lambda^* I)$ coincide con el estimador smart. Con lo que probamos que cuando el smart verifica $p_\Theta(A) \le p_n(A)$, entonces su localización y su matriz de forma coinciden con las correspondientes al estimador MCD.

Las pruebas aportadas de que el smart restringido a $p_{\Theta}(A) \leq p_n(A)$ y a $p_{\Theta}(A) > p_n(A)$ tiene localización y matriz de forma equivalente a la del MCD demuestran la equivalencia de los dos estimadores en cuanto a estos dos aspectos y, por tanto, completan la prueba.

Apéndice B

Conjuntos de datos utilizados

A continuación se hace una breve descripción de los conjuntos de datos reales que se han utilizado a lo largo de toda la memoria.

B.1. *Dataset* de Tejidos Humanos

Se analiza un conjunto de muestras de mRNA humano procedente de 32 tejidos, glándulas y órganos de individuos sanos. En la tabla B.1.1 se proporciona el listado de los tejidos representados en este conjunto. Para cada tipo de tejido se dispone de 3 réplicas biológicas, por lo que en total, el número de muestras analizadas es de 96. Las muestras fueron procesadas utilizando el microarray *GeneChip® Human Genome U133 Plus 2.0*, que incluye 54675 *probe-sets* y que, según la anotación de *Affymetrix*, se corresponden con 14500 genes humanos bien caracterizados. La matriz de expresión definitiva cuenta con 20172 genes anotados en Ensembl, obtenidos a partir del re-mapeo proporcionado por GATExplorer **[Risueño et al, 2010]**. El método de pre-procesado de los datos utilizado es RMA **[Irizarry et al, 2003]**.

Tabla B.1.1. Tejidos representados en el *dataset* Tejidos Humanos

Adipose tissue, Adrenal gland
Bone marrow, Bronchus
Cerebellum, Cerebral cortex
Esophagus
Heart atrium, Heart ventricle
Kidney cortex, Kidney medulla
Liver, Lung, Lymph nodes
Mammary gland, Medulla
Oral mucosa , Ovary
Pituitary gland, Prostate gland
Salivary gland, Saphenous vein, Skeletal muscle, Spleen, Stomach fundus, Stomach pyloric
Testes, Thyroid gland, Tongue, Tonsil, Trachea
Urethra

B.2. *Dataset* de Cáncer de Pulmón

Se analiza un conjunto de muestras de mRNA humano procedente de individuos con cáncer de pulmón de dos tipos, cuyas características se describen en **[Sanchez-Palencia et al, 2011]**. Se dispone de un total de 91 muestras, 45 de ellas procedentes de tejido sano y 46 de células tumorales: 14 adenocarcinomas y 32 squamous-cell carcinomas. Las muestras fueron procesadas utilizando el microarray *GeneChip® Human Genome U133 Plus 2.0* y pueden obtenerse en el repositorio público GEO **[Edgar et al, 2002]**, con el código de acceso GSE18842. La matriz de expresión definitiva cuenta con 20172 genes anotados en Ensembl, obtenidos a partir del re-mapeo proporcionado por GATExplorer **[Risueño et al, 2010]**. El método de pre-procesado de los datos utilizado es RMA **[Irizarry et al, 2003]**.

B.3. *Dataset* de Cáncer de Mama I

Se analiza un conjunto de muestras de mRNA humano procedente de individuos con cáncer de mama, cuyas características se describen en **[Hatzis et al, 2011]**. Se dispone de un total de 310 muestras en dos grupos, 113 pacientes sensibles a cierto tratamiento y 197 no sensibles a dicho tratamiento. Las muestras fueron procesadas utilizando el microarray *GeneChip® Human Genome U133A* y pueden obtenerse en el repositorio público GEO **[Edgar et al, 2002]**, con el código de acceso GSE25055. La matriz de expresión definitiva cuenta con 12576 genes anotados en Ensembl, obtenidos a partir del re-mapeo proporcionado por GATExplorer **[Risueño et al, 2010]**. El método de pre-procesado de los datos utilizado es RMA **[Irizarry et al, 2003]**.

Se recogen, además, algunas características clínico-patológicas de cada uno de los arrays. Las más relevantes son:

- *Source*, procedencia de las muestras que toma dos posibles valores: ISPY, procedentes del consorcio I-SPY-1 (*Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis*) y MDACC, procedentes del departamento de patología del M.D. Anderson *Cancer Center*, Houston, Texas.
- Edad en el momento del diagnóstico.
- ER, receptor de estrógeno que toma dos posibles valores: positivo o negativo.
- PR, receptor de progesterona que toma dos posibles valores: positivo o negativo.
- HERB2, receptor de HERB2 con dos posibles valores: positivo o negativo.
- Grado, que clasifican a las muestras en 4 clases.
- Respuesta patológica: RD en los casos con enfermedad residual y pCR para la respuesta completa.
- Tiempo de supervivencia.
- PAM50, en la que se recoge la clasificación en cinco grupos: luminal A, luminal B, basal, Her2-enriquecido y normal, a partir de la expresión de 50 genes.
- DLDA30, en la que se recoge la predicción de respuesta patológica completa pCR a partir de la expresión de 30 genes.

B.4. *Dataset* de Cáncer de Mama II

Se analiza un conjunto de muestras de mRNA humano procedente de individuos con cáncer de mama, cuyas características se describen en **[Miller et al, 2005]**. Se dispone de un total de 502 muestras que representan el 65% de los cáncer de mama de la provincia de Uppsala (Suecia), entre enero de 1987 y diciembre de 1989. Las muestras fueron procesadas a utilizando dos plataformas: el microarray *GeneChip® Human Genome U133A* y el *GeneChip® Human Genome U133B*; que pueden obtenerse en el repositorio público GEO **[Edgar et al, 2002]**, con el código de acceso GSE3494. En este trabajo se utiliza la serie de 251 muestras procesadas con HG-U133A. La matriz de expresión definitiva cuenta con 12576 genes anotados en Ensembl, obtenidos a partir del re-mapeo proporcionado por GATExplorer **[Risueño et al, 2010]**. El método de pre-procesado de los datos utilizado es RMA **[Irizarry et al, 2003]**.

Las características clínico-patológicas recogidas para cada una de las muestras son:

- P53 *status*, que toma dos posibles valores: positivo o negativo. Respecto de esta característica se utiliza un análisis discriminante para hacer la clasificación de pacientes obteniendo dos variables: p53 DLDA y DLDA error.
- Grado histológico según la escala de Elston, que divide a las muestras en tres grupos.
- ER, receptor de estrógeno que toma dos posibles valores: positivo o negativo.
- PR, receptor de progesterona que toma dos posibles valores: positivo o negativo.
- Edad en el momento del diagnóstico.
- Tamaño del tumor.
- Nodo linfático, que toma dos posibles valores: positivo o negativo
- Tiempo de supervivencia.

Referencias

[Adams et al, 2000] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al. The genome sequence of Drosophila melanogaster. *Science*, 287: 2185 – 2195; 2000.

[Affymetrix, 1999] Affymetrix microarrays suite user guide. Affymetrix, Santa Clara CA, 1999.

[Affymetrix, 2002] Statistical algorithms description document. Affymetrix, Santa Clara CA, 2002.

[Allison et al, 2006] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews. Genetics*, 7: 55 – 65; 2006.

[Alvarez-Esteban et al, 2012] Alvarez-Esteban PC, del Barrio E, Cuesta-Albertos JA, Matrán C. Searching for a common pooling pattern among several samples. Preprint (http://personales.unican.es/cuestaj/Similarity_ksamples.pdf).

[Alwine et al, 1977] Alwine JC, D J Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *PNAS*, 74: 5350–5354; 1977.

[Anderson et al, 2000] Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64: 912–913, 2000.

[Arfin et al, 2000] Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW. Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *Journal of Biological Chemistry*, 275: 29672 – 29684; 2000.

[Ashburner et al, 2000] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25: 25-29; 2000.

[Ayadi et al, 2012] Ayadi W, Elloumi M, Hao JK.·Pattern-driven neighborhood search for biclustering of microarray data. *BMC Bioinformatics,* 13: S11; 2012.

[Baldi y Long, 2001] Baldi P, Long AD. A Bayesian framework for the analysis of microarrays expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17: 509-519; 2001.

[Bamps y Hope, 2008] Bamps S, Hope IA. Large-scale gene expression pattern analysis, in situ, in Caenorhabditis elegans. *Briefings in Functional Genomics and Proteomics*, 7: 175–183; 2008.

[Benjamini y Hochberg, 1995] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57: 289-300; 1995.

[Bergmann et al, 2003] Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 67: 031902; 2003.

[Blattner et al, 1997] Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V et al. The complete genome sequence of Escherichia coli K-12. *Science*, 277: 1453 – 1462; 1997.

[Bolsover et al, 1997] Bolsover SR, Hyams JS, Jones S, Shepard EA, White HA. From genes to cells. New York: Wiley; 1997.

[Bolstad et al, 2003] Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19: 185-193; 2003.

[Bourgon et al, 2010] Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *PNAS*, 107: 9546-9551; 2010.

[Brazma et al, 2001a] Brazma A, Parkinson H, Schlitt T, Shojatalab M. A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. 2001. *http://www.ebi.ac.uk/microarray/biology_intro.html*

[Brazma et al, 2001b] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29: 365 – 371; 2001

[Breiman, 1996] Breiman L. Bagging Predictors. *Machine Learning*, 24: 123-140; 1996.

[Broberg, 2003] Broberg P. Statistical methods for ranking differentially expressed genes. *Genome Biology*, 4: R41; 2003.

[Brown et al, 2001] Brown CS, Goodwin PC, Sorger PK. Image metrics in the statistical analysis of DNA microarray data. *PNAS*, 97: 262 – 267; 2001.

[Busygin et al, 2008] Busygin S, Prokopyev O, Pardalos P. Biclustering in Data Mining. *Computer and Operations Research*, 35: 2964 – 2987; 2008.

[Butte et al, 2001] Butte AJ, Dzau VJ, Glueck SB. Further defining housekeeping, or «maintenance» genes Focus on «a compedium of gene expression in normal human tissues. *Physiological Genomics*, 7: 95-96; 2001.

[Chang et al, 2011] Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL, Hsu IC. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE*, 6: e22859; 2011.

[Chen et al, 2005] Chen D, Liu Z, Ma X, Hua D. Selecting genes by test statistics. *Journal of Biomedicine and Biotechnology*, 2: 132 – 138; 2005.

[Chen et al, 2007] Chen Z, McGee M, Liu Q, Scheuermann RH. A distribution free summarization method for Affymetrix GeneChip arrays. *Bioinformatics*, 23: 321-327; 2007.

[Cheng y Church, 2000] Cheng Y, Church GM. Biclustering of expression data. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*: 93-103; 2000.

[Choe et al, 2005] Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6: R16; 2005.

[Chu et al, 2011] Chu G, Li J, Narasimhan B, Tibshirani R, Tusher V. SAM "Significance Analysis of Micorarrays" Users guide and technical document. *http://www-stat.stanford.edu/~tibs/SAM/*

[Clancy 2008] Clancy S. RNA splicing: introns, exons and spliceosome. *Nature Education*, 1; 2008.

[Cuesta-Albertos et al, 1997] Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed k-Means: An Attempt to Robustify Quantizers. *The Annals of Statistics*, 25: 553-576; 1997.

[Cuesta-Albertos et al, 2008] Cuesta-Albertos JA, Matrán C, Mayo-Iscar A. Trimming and likelihood: robust location and dispersion estimation in the elliptical model. *The Annals of Statistics*, 36: 2284-2318; 2008.

[Cui et al, 2005] Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6: 59 – 75; 2005.

[Cui y Churchill, 2003] Cui X, Churchill GA. Statistical tests for differential expression in cDNA experiments. *Genome Biology*, 4: 210; 2003.

[Dempster et al, 1977] Dempster AP, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B*, 39: 1-38; 1977.

[DeRisi et al, 1997] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expresión on a genomic scale. *Science*, 278: 680 – 686; 1997.

[Díaz-Uriarte, 2005] Díaz-Uriarte R. Supervised Methods with Genomic Data: a Review and Cautionary View. Data analysis and visualization in genomics and proteomics. New York: Wiley, pp. 193-214; 2005.

[Divina et al, 2012] Divina F, Pontes B, Giráldez R, Aguilar-Ruiz S. An effective measure for assessing the quality of biclusters. *Computers in Biology and Medicine*, 42: 254 – 256; 2012.

[Dong et al, 2011] Dong B, Zhang P, Chen X, Liu L, Wang Y, He S, Chen R. Predicting housekeeping genes based on Fourier analysis. *PLoS ONE*, 6: e21012; 2011.

[Draghici, 2002] Draghici S. Statistical intelligence: effective analysis of high-density microarray data. *Drug Discovery Today*, 7: S55-S63; 2002.

[Dudoit et al, 2002a] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors suing gene expression data. *Journal of the American Statistical Association*, 97: 77-87; 2002.

[Dudoit et al, 2002b] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarrays experiments. Technical report 110, Division of Biostatistics, Univ. California Berkeley; 2002.

[Dudoit et al, 2003] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18: 71-103; 2003.

[Edgar et al, 2002] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30: 207-210; 2002.

[Efron et al, 2001] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96: 1151 – 1160; 2001.

[Eisen et al, 1998] Eisen MB, Spellman PT, Brownand PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95: 14863-14868; 1998.

[Eisenberg y Levanon, 2003] Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends in Genetics*, 19: 362 – 365; 2003.

[Fechner 1860] Fechner GT. Elemente der Psychophysik. Leipzig: Breitkopf & Härtel; 1860.

[Farcomeni, 2009] Farcomeni A. Robust double clustering: a method based on alternating concentration steps. *Journal of Classification*, 26: 77 – 101; 2009.

[Fowlkes et al, 2008] Fowlkes CC, Hendriks CL, Keränen SV, Weber GH, Rübel O et al. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, 133: 364–374; 2008.

[Freitas et al, 2011] Freitas A, Afreixo V, Pinheiro M, Oliveira JL, Moura G, Santos M. Improving the performance of the Iterative Signature Algorithm for the identification of relevant patterns. *Statistical Analysis and Data Mining*, 4: 71 – 83 ;2011.

[Freund y Schapire, 1996] Freund Y, Schapire R. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the 13th International Conference*: 148-156; 1996.

[Fritz et al, 2012] Fritz H, García-Escudero LA, Mayo-Iscar A. A fast algorithm for robust constrained clustering. Preprint.

[Gallegos y Ritter, 2009] Gallegos MT, Ritter G. Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification*, 3: 135–167; 2009.

[García-Escudero et al, 2008] García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36: 1324 – 1345; 2008.

[Garrett y Grisham, 2002] Garrett RH, Grisham CM. Principles of biochemistry – with human focus. Pacific Grove, CA: Brooks Cole; 2002.

[Geschwind y Gregg, 2002] Geschwind DH, Gregg JP. Microarrays for the neurosciences: an essential guide. Cambridge, MA: MIT Press; 2002.

[Gleiss et al, 2011] Gleiss A, Sanchez-Cabo F, Perco P, Tong D, Heinxe G. Adaptative trimmed t-statistics for identifying predominantly high expression in a microarray experiment. *Statistics in Medicine*, 30: 52-61; 2011.

[Gordaliza, 1991] Gordaliza A. Best approximations to random variables based on trimming procedures. *Journal of Approximation Theory*, 64: 162 – 180; 1991.

[Greer et al, 2010] Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L. Housekeeping genes; expression levels may change with density of cultured cells. *Journal of Immunological Methods,* 355: 76–79; 2010.

[Gu y Liu, 2008] Gu J, Liu JS. Bayesian biclustering of gene expression data. *BMC Genomics*, 9: S4; 2008.

[Guo y Pan, 2004] Guo X, Pan W. Using weighted permutation scores to detect differential gene expression with microarray data. *Journal of Bioinformatics and Computational Biology*, 3: 989-1006; 2004.

[Hackstadt y Hess, 2009] Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10: 11; 2009.

[Hardiman, 2002] Hardiman G. Microarray technologies – an overview. *Pharmacogenomics*, 3: 293-297; 2002.

[Hartigan, 1972] Hartigan JA. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67: 123-129; 1972.

[Hatfield et al, 2003] Hatfield GW, Hung SP, Baldi P. Differential analysis of DNA microarray gene expression data. *Molecular Microbiology*, 47: 871 – 877; 2003.

[Hatzis et al, 2011] Hatzis C, Pusztai L, Valero C, Booser DJ, Esserman L et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*, 305: 1873-1881; 2011.

[Hsiao et al, 2001] Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV et al. A compendium of gene expression in normal human tissues. *Physiological Genomics*, 7: 97-104; 2001.

[Huang da et al, 2009] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene list using DAVID bioinformatics resources. *Nature Protocols*, 4: 44-57; 2009.

[Hubbard et al, 2002] Hubbard T, Barker D, Birney E, Cameron G, Chen Y et al. The Ensembl genome database project. *Nucleic Acids Research*, 30: 38-41; 2002.

[Hubbard, 2006] Hubbard R. Why we don't really know what statistical significance means: a mayor educational failure. *Journal of Marketing Education*, 28: 114–120; 2006.

[Hubbell et al, 2002] Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics*, 18: 1585 – 1592; 2002.

[International Human Genome Consortium, 2001] International Human Genome Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409: 860 – 921; 2001.

[Irizarry et al, 2003] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4: 249-264; 2003.

[Jain et al, 2002] Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. *Genome Research*, 12: 325 – 332; 2002.

[Jiang et al, 2001] Jiang CH, Tsien JZ, Schultz PG, Hu Y. The effects of aging on gene expression in the hypothalamus and cortex of mice. *PNAS*, 98: 1930–1934; 2001.

[Jiang et al, 2008] Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey MJ, Luo ZW. Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, 9: 284; 2008.

[Jiao y Zhang, 2008] Jiao S, Zhang S. On correcting the overestimation of the permutation-based false discovery rate estimator. *Bioinformatics,* 24: 1655 – 1661; 2008.

[Johnson y Li, 2007] Johnson WE, Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8: 118 – 127; 2007.

[Jung et al, 2005] Jung SH, Bang H, Young S. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6: 157-169; 2005.

[Kanehisa et al, 2006] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleid Acids Research*, 34: D354-357; 2006.

[Khatri y Drâghici, 2005] Khatri P, Drâghici S. Ontological analysis of gene expression data: current tools, limitations, and problems. *Bioinformatics*, 18: 3587-3595; 2005.

[Kouadjo et al, 2007] Kouadjo KE, Nishida Y, Cadrin-Girard JF, Yoshioka M, St-Amand J. Housekeeping and tissue-specific genes in mouse tissues. *BMC Genomics*, 8: 127; 2007.

[LaPointe et al, 2012] LaPointe LC, Pedersen SK, Dunne R, Brown GS, Pimlott L et al. Discovery and Validation of Molecular Biomarkers for Colorectal Adenomas and Cancer with Application to Blood Testing. *PLoS ONE*, 7: e29059; 2012

[Lee et al, 2000] Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridations. *PNAS*, 97: 9834 – 9839; 2000.

[Lian, 2008] Lian H. MOST: detecting cancer differential gene expression. *Biostatistics*, 9: 411-418; 2008.

[Liang et al, 2006] Liang S, Li Y, Be X, Howes S, Liu W. Detecting and profiling tissue-selective genes. *Physiological Genomics*, 26: 158 – 162; 2006.

[Lin et al, 2010] Lin WJ, Hsueh HM, Chen JJ. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 11: 48; 2010.

[Long et al, 2001] Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *Journal of Biological Chemistry*, 276: 19937-19944; 2001.

[Lyoins-Weiker et al, 2004] Lyoins-Weiker J, Patel S, Becich MJ, Godfrey TE. Test for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics*, 5: 110; 2004.

[MacDonald y Ghosh, 2006] MacDonald JW, Ghosh D. COPA-cancer outlier profile analysis. *Bioinformatics*, 22: 2950-2951; 2006.

[Madeira y Oliveira, 2004] Madeira SC, Oliveira AL. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1: 24-45; 2004.

[Magglott et al, 2005] Maglott D, Ostell K, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33: D54-58; 2005.

[McCarthy y Smith, 2009] McCarthy D, Smyth G. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25: 765-771; 2009.

[Miller et al, 2005] Miller LD, Smeds J, George J, Vega VB, Vergara L et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS*, 102: 13550 - 13555; 2005.

[Misra et al, 2002] Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Research*, 12: 1112-1120; 2002.

[Mpindi et al, 2011] Mpindi JP, Sara H, Haapa-Paananen S, Kilpinen S, Pisto T et al. GTI: A novel algorithm for identifying outlier gene expression profiles from integrated microarray datasets. *PLoS ONE*, 6: e17259; 2011.

[Morgan y Sonquistz, 1963] Morgan J, Sonquistz J. Problems in the analysis of survey data and a proposal. *Journal of American Statistical Association*, 58: 415-434; 1963.

[Mosteller y Tukey, 1977] Mosteller F, Tukey JW. Data Analysis and Regression: A Second Course in Statistics. Reading, Mass: Addison-Wesley, pp. 203-209;1977.

[Naef et al, 2001] Naef F, Lim DA, Patil N, Magnasco MO. From features to expression: High density oligonucleotide array analysis revisited. Technical Report 1: 1–9; 2001.

[Pan, 2002] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18: 546 – 554; 2002.

[Pan, 2003] Pan W. On the use of permutation in and the performance of a class of non-parametric methods to detect differential gene expression. *Bioinformatics*, 19: 1333-1340; 2003.

[Pavlidis, 2003] Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. *Methods*, 31: 282-289; 2003.

[Pounds y Cheng, 2005] Pounds S, Cheng C. Sample size determination for the false discovery rate. *Bioinformatics*, 21: 4263-4267; 2005.

[Prelic et al, 2006] Prelic A, Bleuer S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Henning L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22: 1122–1129; 2006.

[Prieto et al, 2008] Prieto C, Risueño A, Fontanillo C, de las Rivas J. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS ONE*, 3: e3911; 2008.

[R Development Core Team, 2011] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[Risueño et al, 2010] Risueño A, Fontanillo C, Dinger ME, de las Rivas J. GATExplorer: Genomic And Transcriptomic Explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. *BMC Bioinformatics*, 11: 221; 2010.

[Robinson et al, 1994] Robinson PA, Marley JJ, High AS, Hume WJ. Differential expression of protease inhibitor and small protein-rich protein genes between normal human oral tissue and odontogenic keratocysts*. Archives of Oral Biology*, 39: 251 – 259; 1994.

[Rousseeuw, 1985] Rousseeuw PJ. Multivariate Estimation with High Breakdown Point. In Mathematical Statistics and Applications, Vol. B (eds. W. Grossmann et al.) pp 283 - 297. Dor-drecht: Reidel Publishing Co; 1985.

[Rousseeuw et al, 2011] Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Maechler M. robustbase: Basic Robust Statistics. R package version 0.7-6; 2011.

[Rousseeuw y Van Driessen, 1999] Rousseeuw PJ, Van Driessen K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41: 212 – 223; 1999.

[Sanchez-Palencia et al, 2011] Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Farez-Vidal ME. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer*, 129: 355-364; 2011.

[Schena et al, 1995] Schena M, Shalon D, Davis R, Brown P. Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science*, 270: 467-470; 1995.

[Schena et al, 1996] Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarry-based expression monitoring of 1000 genes. *PNAS*, 93: 10614 – 10619; 1996.

[Shi et al, 2006] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24: 1151 - 1161; 2006.

[Shyamsundar et al, 2005] Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, Sethuraman A, van de Rijn M, Botstein D, Brown PO, Pollack JR. A DNA microarray survey of gene expression in normal human tissues. *Genome Biology*, 6: R22; 2005.

[Siu et al, 2001] Siu IM, Lal A, Riggins GJ. A database for regional gene expression in the human brain. *Gene Expression Patterns*, 1: 33-38; 2001.

[Slamon et al, 1987] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235: 177–182; 1987.

[Smyth, 2004] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3: A3; 2004.

[Smyth et al, 2005] Gordon K Smyth GK, Joëlle Michaud J, Hamish S Scott HS. Use of within array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21: 2067-2075; 2005.

[Southern, 2001] Southern EM. DNA microarrays. History and overview. *Methods in Molecular Biology*, 170: 1-15; 2001.

[Spencer et al, 2011] Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL et al. A spatial and temporal map of C. elegans gene expression. *Genome Research*, 21:325-341; 2011.

[Stekel, 2003] Stekel D. Microarray bioinformatics. New York: Cambridge University Press; 2003.

[Stevens 1957] Stevens SS. On the psychophysical law. *Psychological Review*, 64: 153-181; 1957.

[Storey, 2002] Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society series B*, 64: 479 – 498; 2002.

[Su et al, 2002] Su AI, Cooke MP, Ching KA, Hakak Y, Walter JR et al. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99: 4465-4470; 2002.

[Su et al, 2004] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101: 6062–6067; 2004.

[Tamayo et al, 1999] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS*, 96: 2907-2912; 1999.

[Tanaka et al, 2000] Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X et al. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15000 mouse developmental cDNA microarray. *PNAS*, 97: 9127 – 9132; 2000.

[Tanay et al, 2002] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18: 136-44; 2002.

[Tang et al, 2001] Tang C, Zhang L, Ramanathan M. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*: 41–48; 2001.

[Tavazoie et al, 1999] Tavazoie S, Hughes J, Campbell M, R. Cho R, Church G. Systematic determination of genetic network architecture. *Nature Genetics*, 22: 281-285; 1999.

[The UniProt Consortium, 2012] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40: D71-D75; 2012.

[Tibshirani, 2006] Tibshirani R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7:106; 2006.

[Tibshirani et al, 2010] Tibshirani R, Chu G, Hastie T, Narasimhan B. samr: SAM: Significance Analysis of Microarrays. R package version 1.28; 2010.

[Tibshirani y Hastie, 2007] Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. *Biostatistics*, 8: 2-8; 2007.

[Tomancak et al, 2007] Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 8: R145; 2007.

[Tomlins et al, 2005] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R et al. Recurrent fusión of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, 310: 644 – 648; 2005.

[Tu et al, 2006] Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7: 31; 2006.

[Tukey, 1977] Tukey JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley; 1977.

[Tusher et al, 2001] Tusher V, Tibshirani R, Chu C. Significance analysis of microarrays applied to ionizing radiation response. *PNAS*, 98: 5116 - 5121; 2001.

[van Iterson et al, 2010] van Iterson MV, Boer JM, Menezes RX. Filtering, FDR and power. *BMC Bioinformatics*, 11: 450; 2010.

[Van Mechelen et al, 2004] Van Mechelen I., Bock H.H., De Boeck P. Two mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13: 363-394; 2004.

[Venet, 2003] Venet D. MatArray: a Matlab toolbox for microarray data. *Bioinformatics*, 19: 659-660; 2003.

[Venter et al, 2001] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. The sequence of the human genome. *Science*, 291: 1304–1351; 2001.

[Wang et al, 2010] Wang L, Srivastava AK, Schwartz CE. Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics*, 11: S15; 2010.

[Watson et al, 1987] Watson JD, Hopkings NH, Roberts JW, Steitz JA, Weiner AM. The functioning of higher eukaryotic genes. Molecular Biology of the Gene. San Francisco: Benjamin-Cummings, pp. 704; 1987.

[Watson y Crick, 1953] Watson JD, Crick FHC. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature*, 171: 737-738; 1953.

[Weber 1834] Weber EH, De pulsu, resorptione, auditu et tactu. Annotationes anatomicae et physiologicae. Leipzig: C.F. Köhler; 1834.

[Won Lee et al, 2005] Won Lee J, Bok Lee J, Park M, Song S. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, 48: 869-885; 2005.

[Wood et al, 2002] Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R et al. The genome sequence of Schizosaccharomyces pombe. *Nature*, 415: 871 – 880; 2002.

[Wu, 2007] Wu B. Cancer outlier differential gene expression detection. *Biostatistics*, 8: 566-575; 2007.

[Xie et al, 2005] Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21: 4280-4288; 2005.

[Yu et al, 2006] Yu X, Lin J, Zack DJ, Qian J. Computational analysis of tissue-specific comninatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Research*, 34: 4925-4936; 2006.

[Zhang, 2007] Zhang S. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics*, 8: 230; 2007.

[Zhang et al, 2010] Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics*, 3: 1; 2010.

[Zhang y Cao, 2009] Zhang S, Cao J. A close examination of double filtering with fold change and t test in microarray analysis. *BMC Bioinformatics*, 10: 402; 2009.

[Zhu et al, 2008] Zhu J, He F, Song S, Wang J, Yu J. How many human genes can be defined as housekeeping with current expression data?. *BMC Genomics*, 9: 172; 2008.

[Zhu et al, 2010] Zhu Y, Natoli R, Valter K, Stone J. Differential gene expression in mouse retina related to regional differences in vulnerability to hyper. *Molecular Vision*, 16: 740-755; 2010.