REVISIÓN

Análisis de datos de microarreglos de ADN. Parte II: Cuantificación y análisis de la expresión génica

✉ Jamilet Miranda, Ricardo Bringas

Centro de Ingeniería Genética y Biotecnología, CIGB
Ave. 31 e/ 158 y 190, Cubanacán, Playa, CP 10600, Ciudad de La Habana, Cuba
E-mail: jamilet.miranda@cigb.edu.cu

RESUMEN

Los experimentos de microarreglos de ADN constan de una primera etapa, en la que se define la pregunta biológica objeto de investigación y se selecciona el diseño experimental que mejor responda a los objetivos. La segunda etapa del experimento comienza una vez que las muestras se procesan e hibridan en los chips. Cuando se obtienen las imágenes fluorescentes, se inicia la etapa de procesamiento, en la que se analizan las imágenes para asignar los valores de expresión, y se aplican métodos estadístico-matemáticos que permitan cumplir los objetivos de la investigación. En este artículo se enfatiza en la metodología para la cuantificación y el análisis de los datos de expresión. Se describen los programas más utilizados para estos análisis y se proponen esquemas de trabajo para acometer algunos de los objetivos más frecuentes. Además, se comentan aplicaciones de esta tecnología en Oncología, donde ha habido avances en cuanto a: la clasificación de nuevos subtipos de cáncer, la identificación de nuevos blancos terapéuticos y la predicción de estadios de la enfermedad.

Palabras clave: Microarreglos de ADN, expresión de genes, cuantificación, análisis estadístico

Biotecnología Aplicada 2008;25:290-300

ABSTRACT

Analysis of DNA microarray data. Part II: Quantification and analysis of gene expression. The first step in a DNA microarray experiment is to define the biological question to be addressed and the selection of an appropriate experimental design. The experiment, however, does not end after the samples are processed and hybridized to the chips, since the fluorescent images that are its primary result have to undergo a data analysis process to assign expression values, using the mathematical-statistical methods suited to the goals of the study. Here a particular emphasis is made on the methodology used to quantify and analyze gene expression data. The most frequently used programs are briefly described and working plans are proposed for some of the most common experimental objectives. Additionally, we discuss and comment the applications of this technology in the field of Oncology, where it has enabled the discovery and classification of new cancer subtypes, and has helped to identify new therapeutic targets, as well as improving the prediction of disease stages.

Keywords: DNA Microarrays, gene expression, quantification, statistical analysis

*I*ntroducción

La secuenciación de genomas completos ha permitido el desarrollo de otras tecnologías de cobertura genómica que proveen información sobre diferentes elementos del funcionamiento de los organismos vivos, lo que ha llevado a un aumento considerable del volumen de datos biológicos. Una de ellas es la de los microarreglos de ADN. Las tareas de colección, manejo y análisis de estos datos de expresión de genes se han incrementado notablemente, y han dado lugar a grandes sistemas de información con estructuras adecuadas para estos propósitos. A su vez, el desarrollo de métodos estadísticos para el análisis de conjuntos de datos con un elevado número de variables y un limitado número de mediciones, ha ganado importancia con el desarrollo de la tecnología de los microarreglos.

En la interpretación de los resultados y principalmente en aquellos estudios encaminados a descifrar las bases moleculares de enfermedades u otros eventos biológicos, el análisis estadístico de las matrices de expresión debe complementarse con la prospección de fuentes de información y ontologías biológicas disponibles, tales como interacciones proteína-proteína, regulaciones de la transcripción y, en general, diferentes anotaciones funcionales a partir de experimentos o predicciones. Este análisis integral forma parte del enfoque de la Biología de Sistema [1]. Aunque existen ya aplicaciones que ayudan a la integración y análisis estadístico de esta información [2], aún es un reto el desarrollo de algoritmos estadístico-matemáticos que ayuden a la formulación de hipótesis biológicas, a partir de redes de genes/proteínas vinculadas a datos de expresión.

En un artículo anterior se trataron aspectos relacionados con la definición de la pregunta biológica, los objetivos y el diseño del experimento, de suma importancia para la obtención de datos que permitan responder a las preguntas formuladas. En este artículo se expone la metodología para el análisis de datos de microarreglos, válida para datos propios o datos que se obtienen de fuentes públicas.

*C*uantificación de la expresión génica

Una vez que se realiza la hibridación de las muestras en los chips, se realiza la lectura de cada uno de estos y se obtienen las imágenes correspondientes. El análisis de imágenes proporciona los datos de expresión de cada gen en cada muestra y genera la llamada matriz de expresión, que tiene como filas a los genes y como columnas a las diferentes muestras analizadas. Inicialmente, esta matriz posee los datos primarios y luego los datos normalizados y preprocesados. A esta primera etapa

1. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, *et al*. A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci USA.;2005:102:17302-7.

2. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004; 20: 578-80.

✉ Autor de correspondencia

se le denomina análisis inicial de los datos. Posteriormente se necesitan métodos estadísticos para una etapa de análisis superior de estas matrices de expresión. Uno de los retos en el trabajo con microarreglos es la colección, el manejo y el análisis de estos datos.

Análisis de imágenes para la cuantificación de la expresión

La lectura de cada chip genera imágenes en forma de matrices de puntos o píxeles, frecuentemente de 16 bits, o lo que es lo mismo, con un rango de valores de intensidad entre 0 y 65 535 (2^{16}-1). Los valores de intensidad de un gen se calculan a partir de un conjunto de valores de la señal correspondiente. Los pasos para determinar los valores de intensidad son:

Localización de la señal

Proceso generalmente automático, que establece la localización del rectángulo que contiene la señal, y asigna las coordenadas a cada gen.

Segmentación

Es el proceso de clasificación de los puntos que componen la imagen del chip como señal o fondo. Este es un paso muy importante, ya que la intensidad de la señal de cada gen depende en buena medida de las diferencias entre la intensidad de lo que se defina como señal y el fondo. Entre los métodos para esta identificación están los que asumen que la señal tiene forma circular. El método de segmentación por círculo fijo determina como señal un círculo de igual radio para todas las señales del chip, mientras que el método de círculo adaptable usa los valores de intensidad para estimar un radio separadamente para cada señal. Ambos métodos poseen limitaciones debido a la forma irregular de las señales. Por esta razón, en su lugar se han aplicado otros métodos que determinan los contornos de la señal de cada gen, de acuerdo con los valores de intensidad de cada uno de los píxeles que la componen [3, 4]. Estos métodos han permitido mejores resultados en la cuantificación de la expresión.

Cálculo de la intensidad (señal)

A partir de los valores contenidos dentro del área de la señal definida en el paso anterior, se calcula la intensidad para cada gen.

Corrección del fondo

Una vez leído el chip, los valores de intensidad pueden contener algún nivel de señal no específica. Cuantificando los puntos en las áreas adyacentes a las señales, se hace un estimado de estos valores. Hay diferentes métodos para la corrección del fondo, algunos calculan un único fondo para todo el chip y otros, llamados de corrección local, determinan un valor de fondo para cada señal o para grupos de señales cercanas. Debido a que dentro de un arreglo los niveles de fondo pueden variar, se recomiendan métodos de corrección locales [5, 6].

Criterios de exclusión de señales

En el proceso de determinación de los valores de intensidad, se deben detectar aquellas señales con inconsistencia en los valores de los pixeles que le dan origen y evaluar su posible eliminación. Las señales con alta variabilidad en los valores de los píxeles que la conforman deben ser eliminadas: este puede ser el caso de señales de intensidades bajas o similares al valor de fondo.

Yang y col. mostraron que en algunos casos la corrección del fondo puede reducir sustancialmente la precisión, ya que se incrementa la variabilidad de las señales con valores de intensidad bajos, mientras que los diferentes procedimientos de segmentación introducen poca variación en la precisión [7]. Por su parte, Wang y col. usaron una función de puntuación de calidad, y observaron que las señales con buena calidad tenían mediciones menos variables y viceversa, así demostraron que la variabilidad inherente en las mediciones de razones de intensidad está inversamente relacionada con la calidad de la señal [8].

Almacenamiento de los datos de expresión

Para facilitar el almacenamiento y el análisis de los datos de expresión han surgido varios LIMS (*Laboratory Information Managment Systems*) diseñados específicamente para datos de microarreglos (Tabla 1). Estos sistemas cumplen con un estándar internacional establecido por la Sociedad MGED (*Microarray Gene Expression Data*), conocido por las siglas MIAME (*Minimum Information About a Microarray Experiment*) [9], que posee tablas estructuradas para el almacenamiento de los datos de las muestras, las condiciones experimentales estudiadas y los valores de expresión. En los últimos años ha aumentado la tendencia a incorporar los análisis de datos como parte de estos sistemas. Uno de los más populares es el BASE (*BioArray Software Environment*, http://base.thep.lu.se) [10], que es un servidor de base de datos que contiene la información de los biomateriales, los datos primarios de expresión e imágenes, y además posee facilidades para la normalización, la visualización y el análisis de los datos de expresión.

Existen varios LIMS que sirven además como repositorios públicos de datos de microarreglos. Entre los más importantes por el volumen de datos de expresión y su actualización está la SMD (*Standford Microarray Database*, http://genome-www5.stanford.edu/) [11, 12] que permite el acceso público a más de 300 experimentos, el ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) [13, 14] del EMBL-EBI (*European Mo-*

3. Beucher S, Meyer F. The morphological approach to segmentation: the watershed Transformation. In: Dougherty E, editor. Mathematical morphology in image processing. New York: Marcel Dekker; 1993. p. 433-81.

4. Adams R, Bischof L. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell 1994;16:641-47.

5. Yang YH, Buckley MJ, Dudoit S, Speed TP. Analysis of Cdna microarray images. Brief. Bioinf 2001;2:341-9.

6. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. Genome Res 2002;12(2): 325-32.

7. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002;11:108-36.

8. Wang X, Ghosh S, Guo S. Quantitative quality control in microarray image processing and data adquisition. Nucleic Acids Res 2001;29:E75-5.

9. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29: 365-71.

10. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 2002;3:SOFTWARE0003.

11. Sherlock G, Hernández-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, *et al.* The Stanford Microarray Database. Nucleic Acids Res 2001;29:152-5.

12. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernández-Boussard T, *et al.* The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res 2005;33:D580-2.

13. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, *et al.* ArrayExpress- A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68-71.

Tabla 1. Sistemas más populares para el almacenamiento y manejo de datos de microarreglos

LIMS* disponibles	Instituto	Sistema gestor de base de datos	URL
BASE (BioArray Software Enviroment)	Oncology Department, Lunds University	MySQL, PostgreSQL	http://www.lu.se/
MaxdSQL (Manchester Array Express Database)	Microarray Bioinformatics Group, Manchester University	Oracle, MySQL, PostgreSQL	http://www.bioinf.man.ac.uk/microarray/maxd/
MADAM (MicroArray Data Manager)	The Institute of Genomic Research - TIGR	MySQL	http://www.tigr.org/software/tm4/madam.html
SMD (Stanford Microarray Database)	Stanford University	Oracle, LAD (The Longhorn Array Database) es una implementación de SMD en PostgreSQL	http://genome-www5.stanford.edu/MicroArray/SMD/download/
ArrayExpress	EMBL-EBI	Oracle	http://www.ebi.ac.uk/arrayexpress/

*Laboratory Information Management Systems.

lecular Biology Laboratory-European Bioinformatics Institute), con más de 1800 experimentos y el GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) [15, 16] del NCBI (*National Center for Biotechnology Information)*, con más de 4900 series de datos de expresión públicas. Además del análisis de cada experimento, la gran cantidad de información disponible públicamente atrae el interés de los investigadores para el estudio de mecanismos moleculares globales y específicos de enfermedades. Esto justifica la creciente aparición de estudios que realizan meta-análisis de datos de microarreglos [17-19].

Preanálisis: filtrado, normalización y preprocesamiento de los datos de expresión

Una vez realizado el análisis de imagen y calculadas las intensidades de cada señal, se hacen imprescindibles las transformaciones de estos datos primarios, debido a que las intensidades de cada señal, además de reflejar los niveles de ARNm, pueden contener sesgos asociados a la forma de impresión en el chip, a la eficiencia del marcaje de las muestras y a otras fuentes de variabilidad mencionadas antes. Algunas de estas transformaciones son el filtrado y normalización de los datos y la aplicación de preprocesamientos.

Filtrado de los datos

Como primer paso se recomienda el filtrado de los datos primarios para eliminar los valores que probablemente sean el producto de errores. Se utilizan criterios como el cálculo del coeficiente de variación (*CV*) para cada gen, que se calcula como la desviación estándar (*SD*) entre la media de un conjunto de razones de expresión de múltiples señales del mismo gen: $CV = SD / Media$, y se eliminan los genes con *CV* mayores que un determinado umbral. Otro criterio consiste en eliminar los valores de intensidad mayores que un umbral, que pudieran ser valores de señales sobresaturadas. También se sugiere realizar una inspección visual de las imágenes para detectar efectos en los arreglos, tales como el rayado, la neblina, el borde, la burbuja, y eliminar las intensidades correspondientes antes de proceder a la normalización [20].

Normalización

El proceso de normalización debe ser la primera transformación aplicada a los datos de expresión y un paso esencial antes de pasar a su análisis. Esta transformación se realiza con el objetivo de minimizar las variaciones sistemáticas en la cuantificación de los niveles de hibridación de las muestras de ARNm, de manera que las diferencias biológicas se puedan distinguir fácilmente [21], y hacer comparables los niveles de expresión entre los chips. La normalización por lo general se aplica en el interior de cada chip o entre múltiples chips, y para ello se deben seleccionar los métodos y las variables o regiones del arreglo (conjunto de genes) que servirán para la estandarización de los datos.

Conjunto de genes a usar para la normalización

En general los métodos de normalización suponen que:

1) La mayoría de los genes cambian su expresión y se normalizan basados en determinadas señales (genes de referencia), o que;

2) La mayoría de los genes del arreglo no cambian su expresión y se normalizan basados en las intensidades de todas las señales del chip [22, 23].

La variante uno se aplica generalmente cuando se utiliza un arreglo que contiene una selección de genes que se conoce asociado al problema biológico que se estudia, por ejemplo, genes relacionados con determinada enfermedad. Como referencia se usa un conjunto de genes que previamente se conoce que tienen niveles de expresión similares en todas las muestras analizadas. Este es el caso de genes que realizan funciones esenciales y que deben expresarse siempre a niveles similares (los llamados *housekeeping*). La otra posibilidad es incluir genes controles en el chip, que no se expresen en ninguna muestra, como pudieran ser genes de organismos lejanos evolutivamente; pero esta variante es menos recomendada que la primera. La variante dos es común cuando se utilizan chips que incluyen sondas de todos los genes de un organismo. Otras variantes de normalización han sido introducidas por Yang y col. [21], quienes proponen separar el chip por grupos de impresión y aplicar los métodos de normalización en cada grupo localmente. De esta forma se logra eliminar el llamado efecto punta que puede ocurrir entre diferentes grupos de impresión dentro del chip.

Métodos de normalización

Existen varios métodos de normalización:

- Global o lineal (aplicable a chips del tipo ADNc y Affymetrix): el factor de normalización es el mismo para todos los genes del chip.
- Dependiente de la intensidad (aplicable a chips del tipo ADNc y Affymetrix): el factor de normalización depende de la intensidad de cada señal.
- Dependiente de la localización (aplicable a chips del tipo ADNc): el factor de normalización depende de la localización de la señal en el chip.

Estos métodos, en general, han sido de mucha utilidad para normalizar datos de expresión de genes [21, 24, 25] y se pueden aplicar a los datos de expresión del interior de un arreglo o para la normalización entre pares de arreglos.

La aplicación de estos métodos difiere según el chip empleado. A modo de ejemplo mostramos el chip de ADNc. Para esto, definiremos a *R* como el conjunto de valores de intensidades de la muestra marcada en rojo y *G* como el conjunto de valores de la marcada en verde.

El método global o lineal asume que se parte de iguales cantidades de ARN en las muestras que se comparan, por lo que el número de moléculas en cada una es también el mismo y además asume que los elementos impresos en el chip son una muestra aleatoria de los genes de un organismo. Por lo tanto, aproximadamente el mismo número de moléculas marcadas de cada muestra debe hibridar en el arreglo, y así, la suma total de intensidades de todos los elementos del arreglo debe ser la misma en cada muestra. Basado en esto, se calcula un factor *k* de normalización sumando las intensidades en ambos canales:

$$k = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} G_i}$$

14. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, *et al.* ArrayExpress- A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005; 33:D553-5.

15. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207-10.

16. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, *et al.* NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res 2007;35:D760-5.

17. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004; 36:1090-8.

18. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large scale-meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA. 2004; 101:9309-14.

19. Kim RD, Park PJ. Improving identification of differentially expressed genes in microarray studies using information of public databases. Genome Biol 2004; 5:R70

20. Troyanskaya Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520-5.

21. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002;3(11).

22. Quackenbush J. Microarray data normalization and transformation. Nature Genetics Supp 2002;2:496-501.

23. Kroll TC, Wolfl S. Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. Nucleic Acids Res 2002;30:e50.

24. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 2002;12:111-39.

25. Steinhoff C, Vingron M. Normalization and quantification of diferential expression in gene expression microarrays. Brief Bioinform 2006;7:166-77.

así la razón de expresión normalizada T'_i para cada elemento del chip queda:

$$T'_i = \frac{R'_i}{G'_i} = \frac{1}{k}\frac{R_i}{G_i}$$

Esta transformación es equivalente a sustraer una constante del logaritmo de la razón de expresión:

$$\log_2 (T'_i) = \log_2 (T_i) - \log_2 (k)$$

Existen variantes de este método en dependencia de si se considera que la media o mediana de las intensidades es la misma en el interior de cada arreglo o a través de todos los arreglos; también es posible aplicarlo a solo un subconjunto de genes en lugar de a todos los genes del arreglo. Kroll y col. hicieron un análisis comparativo de los métodos básicos para la normalización de la expresión de los genes basados en un factor entre las intensidades [23]. Estudiaron diferentes variantes del factor de normalización, tales como la media de la expresión de conjuntos de genes de referencia, la suma, la media, la mediana, el cuantil o percentil de los logaritmos de todos los valores de expresión y la media sin tener en cuenta los valores más altos de intensidad. En este trabajo se concluye que usar como factor de normalización la media de los valores centrales después de eliminar el 5 y 10% de los valores más altos de intensidad, es un método simple y robusto para normalizar este tipo de datos. Un problema de estos métodos es que no tienen en cuenta los efectos intensidad y bloque, descritos en varios estudios [26-28].

Dudoit y col. [24] sugirieron el uso de un gráfico que se construye al representar en un eje de coordenadas los valores *logs* de los datos primarios:

$$M = \log_2 (R / G) \quad \text{v.s} \quad A = \frac{1}{2}\log_2 (R*G)$$

que resulta muy útil para identificar ruidos en las señales por diferencias en la eficiencia de marcaje. Este se conoce como gráfico-MA, por los nombres de las variables que se grafican.

Si se hibrida en un mismo chip una muestra marcada con ambos fluoróforos, se espera que los valores $\log_2 (R / G)$ sean 0; sin embargo, lo que se observa en la mayoría de los casos es una desviación de estos valores tanto para intensidades bajas como altas. Lowess (*LOcally Weighted rEgression and Smoothing Scatterplots*) [29] es un método de normalización que puede eliminar estos efectos específicos del marcaje, que depende de los valores de intensidad. Con *lowess* se estima la función, que ajusta los valores del gráfico-MA y solo estará afectada por los genes expresados diferencialmente, los cuales se comportan para estos propósitos como datos aberrantes (*outliers*). Al hacer esta normalización se realiza la transformación:

$$\log_2 R / G \longrightarrow \log_2 R / G - c\,(A) = \log_2 R / k\,(A)*G$$

donde $A = \log_2 \sqrt{(R*G)}$ y los valores R', G' normalizados por *lowess* quedan transformados de la siguiente manera:

$$G' = 2^{c\,(A)} * G \;, \quad R' = R$$

Este método puede aplicarse sobre todos los elementos del chip o puede estimarse una función en cada partición del chip.

En el trabajo de Yang IV y col. [21] se estudió este efecto dependiente de la intensidad: se realizaron hibridaciones en 30 líneas celulares y para cada una se mostraron los gráficos-MA y se calcularon las SD antes y después de realizar la corrección de los datos por *lowess*. Esa investigación demostró la conveniencia del uso del método *lowess* en las alteraciones dependientes de la intensidad.

Por otra parte, Yang YH y col. [7] hicieron una comparación más amplia en datos provenientes de microarreglos de ADNc, que abarcó diferentes métodos de normalización, tales como el global, el *lowess* basada en todos los genes del chip, el *lowess* en cada chip por bloques de impresión, el *lowess* en cada par de chips con marcaje reverso (aplicable solamente a microarreglos de ADNc para eliminar el sesgo por el fluoróforo) y un método de regularización de la varianza aplicado en cada chip, por bloques de impresión y entre múltiples chips. En este trabajo el método *lowess* aplicado en cada chip por grupos de impresión mostró los mejores resultados. Además se evidenció que esta normalización es necesaria, ya que a diferencia de la global, elimina sesgos que dependen de la intensidad y de la ubicación espacial.

La normalización en cada chip con los métodos global y *lowess* por grupos de impresión, se puede considerar un consenso. Además, en el doble marcaje se adiciona la normalización por marcaje reverso. Por otra parte, hay autores que trabajan en la búsqueda de métodos de normalización que ofrezcan mejores resultados y que partan de hipótesis biológicas diferentes. Fan y col. [30] introdujeron un método basado en un modelo semilineal, aplicado en el interior de cada chip para estimar los efectos de intensidad y bloque a partir de 100 réplicas de un mismo gen en el chip. Luego estos valores estimados de los efectos se eliminan de todas las intensidades, y finalmente se aplica una normalización global para corregir otros efectos de los chips. Este tipo de método es una alternativa para los casos en que no se cumplan las hipótesis biológicas del método *lowess*, de que la mayoría de los genes no varía su expresión o que el número de genes sobreexpresados y reprimidos es similar dentro de cada bloque de impresión.

Preprocesamientos a los datos de expresión

Una vez normalizados, los valores de expresión deben pasar por una etapa llamada preprocesamiento, que consiste en una serie de transformaciones adicionales a los datos, que persigue remediar parcialmente problemas que aún persistan en los valores de expresión de genes. Algunas de estas transformaciones son:

Tratamiento de las réplicas dentro de cada arreglo

Es recomendado analizar y filtrar las réplicas inconsistentes y posteriormente estimar el valor de expresión de cada gen como el promedio o mediana de los valores de intensidad de todas sus réplicas en el arreglo.

El llenado de los datos perdidos o en blanco

La falta de datos en las matrices de expresión puede ocurrir por diferentes razones. Entre ellas, las insuficientes resoluciones de los escáneres, problemas en las imágenes o en las propias placas utilizadas. Para tratar estos datos perdidos, los criterios más em-

26. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations, and assessment of gene effects. Nucleic Acids Res 2001;29:2549-57.

27. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003;18:71-103.

28. Ma S, Kosorok MR, Huang J, Xie H, Manzella L, Soares MB. Robust semiparametric cDNA microarray normalization and significance analysis. Biometrics 2006;62:555-61.

29. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829-36.

30. Fan J, Tam P, Woude GV, Ren Y. Normalization and analysis of cDNA microarrays using within array replications applied to neuroblastoma cell response to a cytoquine. Proc Natl Acad Sci USA 2004; 101:1135-40.

pleados son llenar con la mediana o la media de los datos de expresión del gen o la imputación por los k vecinos más cercanos (*kNN-Imputation*), este último es el más robusto [20]. El llenado es un paso imprescindible si se necesita aplicar métodos de agrupamiento para determinar grupos de genes y muestras coexpresados. Si por el contrario, el tipo de análisis que se va a realizar es una prueba *t*, este paso resulta innecesario; de hacerlo, podría introducir modificaciones en los resultados.

Filtrado de patrones planos

Un criterio es filtrar los genes que mantienen su expresión casi constante en todos los chips; así la desviación estándar de su perfil de expresión es pequeña y menor que determinado umbral. Para la selección de un umbral resulta útil la visualización del número de genes por rangos de valores de desviación estándar. Mantener estos genes al realizar análisis a otros niveles puede falsear los resultados, fundamentalmente de algoritmos de agrupamiento [31].

Análisis exploratorio de los datos (EDA)

Las técnicas exploratorias se usan con el objetivo de conocer más acerca de los datos disponibles, identificar relaciones entre las variables, sin necesidad de dar información previa de estas y, en ocasiones, realizar una reducción o selección de las variables que mejor explican el comportamiento del sistema. Una parte de estas técnicas se ubica dentro de la llamada *estadística descriptiva*, que emplea el análisis univariado y métodos basados en la visualización de la distribución de las variables y el cálculo de la media, mediana y desviación estándar de cada variable. En los estudios de microarreglos puede ser útil visualizar histogramas de la cantidad de genes por rangos de estas medidas calculadas. Por el alto número de variables, las *técnicas exploratorias multivariadas*, como el análisis de componentes principales [32, 33], resultan más informativas aún: se construyen componentes como combinaciones lineales ortogonales de las variables originales. Cada componente principal explica un porciento de la variabilidad del sistema en estudio, preferiblemente las dos primeras componentes deben explicar el 80% o más de esta variabilidad. En este proceso se obtiene un coeficiente que representa el peso de cada variable en cada componente. Las variables o genes más importantes serán aquellos que tienen mayores valores absolutos de sus coeficientes en la primera y segunda componente. Puede suceder que solo la primera componente explique un porciento alto de esta variabilidad, y en ese caso, para los análisis posteriores, se puede quedar solamente con las variables de mayores coeficientes en la primera componente principal. Para su mejor comprensión, ilustraremos el uso de esta técnica sobre los datos de expresión de los 50 genes con mayores diferencias entre los tumores y tejidos sanos de próstata del estudio de Lapointe y col. [34]. En la figura 1 se representan los valores de cada muestra en los ejes de coordenadas de la primera y segunda componentes principales. Ambas componentes explican un 80% de la variabilidad del sistema en estudio: 76.7 y 3.3 respectivamente, y puede observarse cómo dividen los

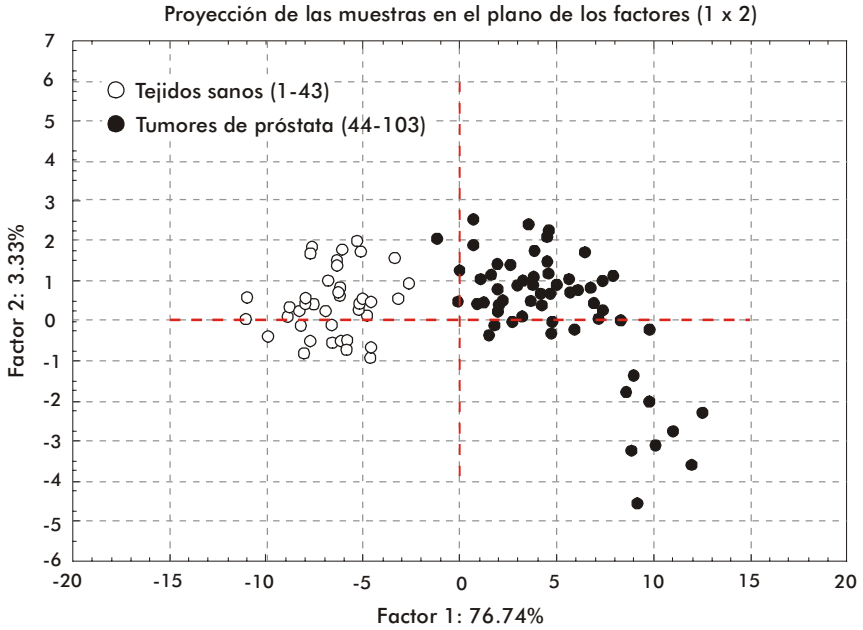

Figura 1. Proyección de los valores de las 103 muestras en el plano de la primera (eje x) y segunda (eje y) componente principal o factores. Estos componentes se obtienen al aplicar un análisis de componentes principales sobre los 50 genes que más cambiaban su expresión entre las clases formadas por muestras de tejidos sanos y tumores de próstata. Estas dos com-ponentes separan los dos grupos experimentales estudiados. El gráfico que se muestra es una salida del programa Stat 6.1 y en cada eje se muestra el porciento de variabilidad que explica cada factor (76.74, la primera componente, y 3.33, la segunda). Los datos de expresión se tomaron del estudio público de Lapointe y cols. [34].

dos grupos de muestras. Este alto porciento permite relacionar estas primeras componentes principales con los fenómenos que más resaltan en los datos analizados, y a su vez, seleccionar los genes con mayores coeficientes en ambas componentes principales, para continuar los análisis. En este ejemplo particular, hay asociación de la primera componente con un fenómeno de represión de genes (30 genes de mayores coeficientes), y de la segunda componente con una sobreexpresión (dos genes de mayores coeficientes), donde uno de los genes sobreexpresados es el AMACR, del que se ha reportado una actividad elevada en tumores de próstata [35].

Análisis estadístico-matemáticos según los objetivos del experimento

Existe una relación estrecha entre los objetivos del experimento y el tipo de método estadístico que debe emplearse para dar respuesta a ellos [36] (Tabla 2). Los métodos estadísticos para análisis de datos de microarreglos se dividen en supervisados y no supervisados [37]. Los supervisados requieren la definición de clases o grupos experimentales. Estos incluyen los métodos orientados a identificar genes de comportamientos diferenciados entre las clases definidas (métodos de comparación) y aquellos orientados a predecir la membresía a una clase u otra (métodos de predicción). En este último caso se requiere un paso previo de selección de variables. Los métodos no supervisados (métodos de agrupamiento) van dirigidos fundamentalmente a identificar genes con comportamientos similares sin que se conozca la clase a que pertenecen (Figura 2).

31. Herrero J and Dopazo J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. J Proteome Res 2002;1:467-70.

32. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic Press; 1979.

33. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer-Verlag; 2002.

34. Lapointe J, Li C, Higgins JP, Van de Rijn M, Bair E, Montgomery K. Gene expression profiling identifies clinically relevant sub-types of prostate cancer. Proc Natl Acad Sci USA. 2004;101:811-6.

35. Kumar-Sinha C, Shah RB, Laxman B, Tomlins SA, Harwood J, Schmitz W, Conzelmann E, Sanda MG, Wei JT, Rubin MA, Chinnaiyan AM, *et al*. Elevated alpha-methy-lacyl-CoA racemase enzymatic activity in prostate cancer. Am J Pathol 2004 Mar; 164(3):787-93.

36. Simon R, Radmacher MD. Dobbin K. Design of studies using DNA Microarrays. Genet Epidemiol 2002;23:21-36.

37. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. Methods Mol Biol 2003;224:111-36.

Selección de genes significativos: métodos supervisados

Para determinar los genes cuyos perfiles de expresión se diferencian de manera significativa entre las condiciones estudiadas, es necesario el uso de métodos para la comparación de clases. Estos son métodos supervisados de aprendizaje que requieren como entrada la información del grupo o condición experimental a que pertenece cada muestra y como resultado identifican genes que están expresados diferencialmente en los grupos o condiciones definidos.

En general, el problema se reduce a seleccionar una prueba estadística adecuada y calcular los valores *p* de las pruebas. La elección de la prueba estadística depende del número de condiciones que se desea comparar. No se recomienda el uso del factor de cambio *FC*:

$$FC = \log_2 \left(\frac{\overline{x}_1}{\overline{x}_2} \right), \quad FC \geq 2 \quad \text{ó} \quad FC \leq -2$$

como medida de expresión diferencial, en este caso los cambios en la varianza dominan el análisis de expresión diferencial.

La prueba más aceptada para comparar dos condiciones es una *t* modificada [24, 38], como la de Tusher y col. [39], que está implementada en el software SAM (*Significance Analysis of Microarrays*, http://www-stat.stanford.edu/~tibs/SAM/). Actualmente existen versiones del SAM para facilitar la comparación de múltiples condiciones experimentales. En general las modificaciones a la prueba *t* se aplican al denominador. Tusher y col. proponen una prueba *d*, y que la modificación a la prueba *t* [40] consista en sumar un valor s_0, de manera que para cada gen se puede calcular:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sigma^2 + s_0}$$

tal que $s_0 > 0$, y σ es la variabilidad de su expresión dentro de las clases, y $\overline{x}_1 - \overline{x}_2$, la diferencia de las medias de expresión del gen entre ambas clases.

Se pueden usar diferentes tipos de métodos para determinar expresión diferencial [37, 38]: *1)* los que se basan en la normalidad de los datos como la prueba *t* para dos grupos o la prueba *F* para múltiples condiciones, *2)* las pruebas no paramétricas como la de Wilcoxon (para dos condiciones) y Kruskal Wallis (para múltiples condiciones), que se basan en la suma de rangos y no exigen normalidad u otros métodos no paramétricos menos conocidos [41-43], y *3)* procedimientos bayesianos [44-46]. En esquema se muestra una propuesta para el uso de estas pruebas estadísticas (Figura 3), y se diferencian las pruebas por el número de clases y por la hipótesis de normalidad. Después de aplicar la prueba, se obtiene una lista ordenada de genes según su estadística de expresión diferencial (valor de la prueba).

También se debe tener en cuenta que en este tipo de experimentos se realiza inferencia de miles de variables (genes), por tanto, en los resultados se necesitan genes con valores *p* pequeños, del orden de 10^{-3} a 10^{-4}. Una alternativa muy usada es el cálculo de los valores *p* no ajustados, para lo que se recomienda el uso de algoritmos de remuestreo [47]. La necesidad de este cálculo se explica porque, en general, los datos de microarreglos no cumplen con la hipótesis de normalidad. Si además se desea tener en cuenta la dependencia entre las variables observadas, como es conocido en el caso de genes funcionalmente relacionados, se deben realizar cálculos de valores *p* ajustados por pruebas de hipótesis múltiples con permutaciones [27].

Tabla 2. Resumen de los métodos estadísticos más utilizados según los objetivos del experimento

Objetivos	Métodos estadísticos más utilizados	Causa de la selección
Comparación de clases	t-test, **F-test**, Wilcoxon, Kruskal Wallis, SAM	F-test es una prueba que permite comparar dos o más condiciones experimentales, y logra mayor precisión que las pruebas no paramétricas (Wilcoxon, Kruskal Wallis) en datos de microarreglos
Predicción de clases	**kNN**, DLDA, Naive Bayes, QDA, LDA, LOCLDA, SVM	A pesar de que kNN es un método simple, en estudios de comparación de métodos discriminantes en datos de microarreglos, fue el de mejores resultados según Dudoit y cols [52].
Descubrimiento de clases	k-means, SOM, HCL, **SOTA**	SOTA es un método de agrupamiento que combina el SOM y el HCL. Tiene implementado un criterio de parada para la división de los grupos basado en una medida de variabilidad entre los elementos del grupo. Fue especialmente diseñado para datos de microarreglos por Herrero y cols [82]. En este caso particular también es llamado SOTArray

*Se destacan en negritas los métodos que recomendamos en cada caso y en la tercera columna se explican las razones de la selección.

Por último, para seleccionar los genes que consideraremos diferencialmente expresados, se prefija un criterio, que puede ser valores mayores que un umbral para la prueba estadística o valores menores que un umbral para los valores *p* ajustados. También se usan controles como el cálculo de la proporción de descubrimientos falsos, FDR (False Discovery Rate), y/o el FWER (Family Wise Error Rate). Se han utilizado diferentes modificaciones de estos procedimientos para tener un estimado de la proporción esperada de falsos positivos o negativos en los resultados [27, 48-50].

Búsqueda de una firma molecular: métodos supervisados

Para encontrar una *firma molecular* o lo que es lo mismo un conjunto reducido de genes cuyo perfil de expresión permite clasificar una nueva muestra, es necesario el uso de métodos para predicción de clases, los cuales también son métodos supervisados de aprendizaje. En este caso el objetivo es desarrollar un predictor de clases multivariado para asignar la membresía de un nuevo individuo a una clase u otra. En el estudio comparativo realizado por Dudoit y col. [51],

38. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002;18:546-54.

39. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116-21.

40. Welch BL. The generalization of 'students' problem when several different population variances are involved. Biometrika 1947;34:28-35.

41. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002;18:1454-61.

42. Zhao Y, Pan W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments.Bioinformatics 2003;19:1046-54.

43. Yan X, Deng M, Fung WK, Qian M. Detecting differentially expressed genes by relative entropy.J Theor Biol 2005; 234:395-402.

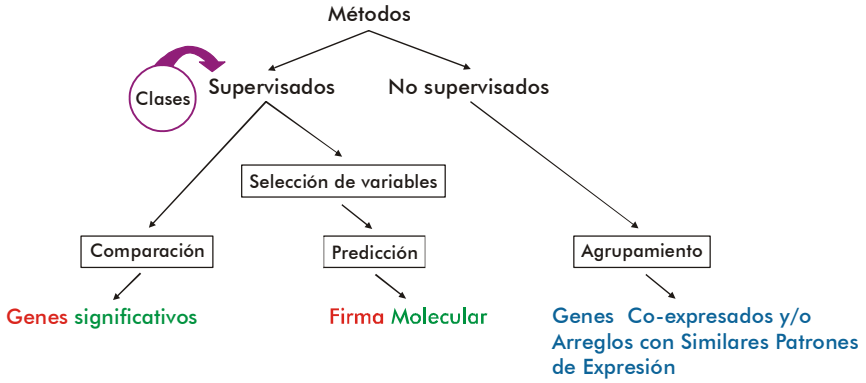

Figura 2. Esquema de la estructura general para el análisis estadístico de los datos de microarreglos. La comparación y predicción de clases requieren el conocimiento previo del grupo experimental (clase) en que esté ubicada la muestra estudiada e incluida en el análisis, no así los métodos no supervisados.

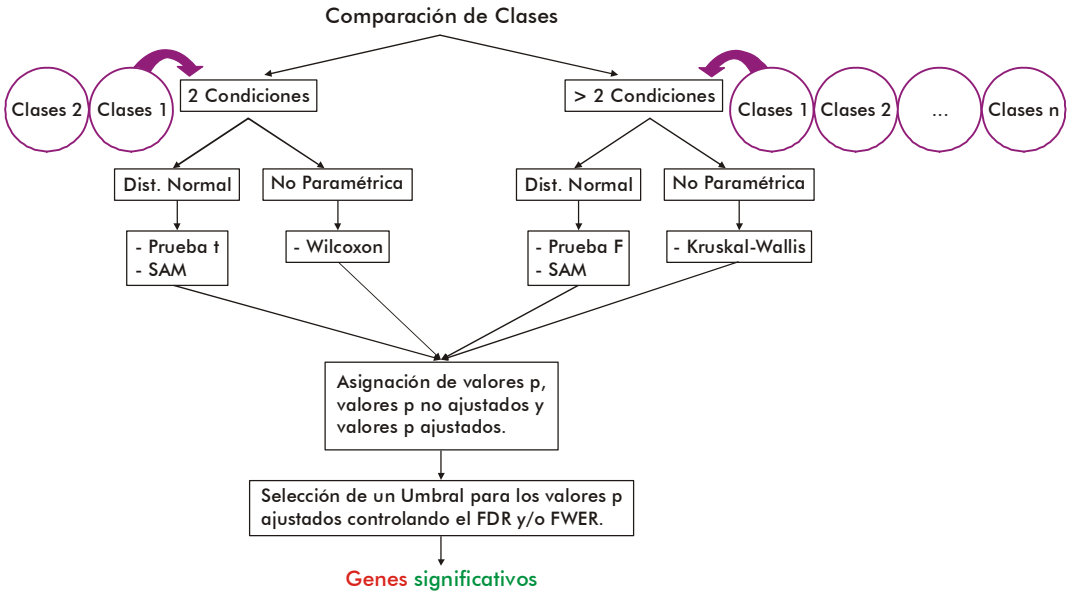

Figura 3. Propuesta de esquema de trabajo para el análisis estadístico y para encontrar los genes que tienen cambios estadísticamente significativos en una comparación de dos o más condiciones experimentales. Algunas pruebas estadísticas suponen la normalidad de los datos, mientras que no ocurre así en las no paramétricas. Luego, independientemente de la prueba utilizada, se calculan los valores *p* de cada una, se realizan permutaciones de las muestras entre los grupos estudiados y se obtienen los valores *p* no ajustados. Finalmente se calculan los valores *p* ajustados por pruebas de hipótesis múltiples con permutaciones. Los genes que tengan el valor *p* ajustado menor que determinado umbral, se consideran con cambios de expresión estadísticamente significativos entre las clases comparadas. Para la selección del umbral deben tenerse en cuenta las razones FDR/FWER.

el método *k* vecinos más cercanos *k* NN (*k* Nearest Neighbors) [52] resultó ser el de menor error de clasificación para responder a este objetivo.

Con anterioridad al desarrollo del predictor se requiere un paso de selección de variables (genes). Es razonable asumir que solo un subconjunto de los genes evaluados aporta información útil para distinguir las clases. Un método muy utilizado para esta reducción de dimensionalidad es seleccionar genes basados en la significación estadística de pruebas univariadas (prueba *t*, *F* o la prueba de suma de rangos de Wilcoxon), de diferencias entre las clases. Los genes con diferencias estadísticamente significativas se seleccionan para su inclusión en el modelo multivariado. Si se incrementa la exigencia del umbral de significancia, se obtiene un modelo más simple que contiene pocos genes, pero se corre el riesgo de perder genes importantes. Este procedimiento requiere tamaños de muestra grandes para poder identificar los genes más importantes, y así construir un modelo basado en la inclusión de un conjunto de genes, que prediga con mayor exactitud. Una estrategia es probar con umbrales de significancia bajos (entre 10^{-2} y 10^{-4}), y determinar la razón de clasificación errónea para los modelos resultantes usando validación cruzada. También se han usado pruebas de múltiples hipótesis y el análisis de componentes principales para realizar esta selección de variables [53, 54].

Agrupamiento de perfiles de expresión de genes: métodos no supervisados

El objetivo de la aplicación de los métodos de agrupamiento o *clustering* a datos de expresión es construir grupos de genes o muestras con perfiles de expresión similares utilizando una medida de distancia [51]. Las medidas de distancia más usadas son la *euclidiana* y la correlación de *Pearson*. En el caso de los métodos de agrupamiento jerárquicos se necesita además definir el método para determinar distancias entre conjuntos de genes [55].

Los métodos de agrupamiento por lo general no requieren la información del grupo, clase o condición experimental a que pertenece cada muestra que se incluye en el análisis, sino que por el contrario pueden sugerir un nuevo agrupamiento de las muestras basado en el grado de similitud entre los perfiles de expresión de los genes en estudio. Estos métodos, aplicados a datos de expresión, sirven para identificar grupos de genes coexpresados y patrones en los perfiles de expresión de las muestras, sin la necesidad de clases predefinidas que supervisen el análisis [56, 57].

El método de agrupamiento más empleado en datos de microarreglos es el agrupamiento jerárquico. Este método no supervisado deriva una serie de particiones de los datos; en este caso, cada dato será el perfil de expresión de una muestra o gen. Existen varios tipos de métodos de agrupamiento jerárquicos, tales como el aglomerativo y el divisivo, los divisivos funcionan mejor para dividir los datos en pocos grupos de varios elementos. El resultado de estos métodos es una estructura de árbol o dendograma. Existen alternativas a los métodos jerárquicos y de ellas el método *k-Means* es la más usada, pero tiene la desventaja de que requiere como entrada el número de grupos en que se considera estén separados los datos. La estimación de *k* es un problema conocido, siempre que se desea encontrar el mapeo de cualquier estructura de datos a una estructura de grupos, especialmente estudiado en datos de expresión de genes [58, 59]. Un criterio muy usado propone seleccionar a *k* como el número de grupos a partir del cual se observan pocas variaciones de

44. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8:37-52.

45. LönnStedt I, Speed TP. Replicated microarray data. Statistica Sinica 2002; 12:31-46.

46. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments.Stat Appl Genet Mol Biol 2004;3:Article 3.

47. Westfall PH, Young SS. Re-Sampling Based MultipleTesting New York: Wiley; 1993.

48. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96: 1151-60.

49. Storey J, Tibshirani R. Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays Data. Stanford University, Technical Report; 2001:28.

50. Tsai CA, Hsueh HM, Chen JJ. Estimation of false discovery rates in multiple testing: application to gene microarray data. Biometrics 2003;59:1071-81.

51. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. American Statistical Association 2002;97(457):77-87.

52. Hastie TJ, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

las ordenadas del gráfico FOM (Figure of Merit) [60]. Otros métodos se basan en evaluar la estabilidad de los grupos [61].

Uno de los errores más frecuentes en el análisis de datos de microarreglos es el uso del análisis por grupos para resolver problemas de predicción y comparación de clases [62]. Este análisis no da información cuantitativa válida desde el punto de vista estadístico sobre cuáles genes se expresan diferencialmente entre clases. Este sería solo un elemento exploratorio y se recomienda el uso de métodos supervisados, como los descritos en los puntos anteriores.

Softwares disponibles para el análisis de datos

Bioconductor: Es un proyecto internacional de software libre para el análisis e interpretación de datos genómicos, escrito en el lenguaje estadístico R (http://www.r-project.org), que incluye gran número de procesamientos computacionales para el análisis de datos de expresión de genes [63]. El proyecto Bioconductor (http://www. bio conductor.org) dispone de una serie de paquetes que contienen un amplio rango de aplicaciones estadísticas para varios tipos de análisis genómicos: ctc [64] para *clustering* de valores de expresión, multtest [27, 65] y maanova [66] para comparación de múltiples condiciones experimentales, marray [67] que contiene las funciones *loess/lowess* para realizar una regresión local, samr [68] que implementa el método SAM para determinar expresión diferencial, entre otros.

MeV (*The Institute for Genomic Research*, EE.UU.): este es uno de los programas más populares para el análisis de datos de microarreglos, debido a la gran variedad de métodos matemáticos que tiene implementado [69]. Este es un software libre de fácil instalación, escrito en lenguaje Java y con una interfaz muy amigable para el usuario. Entre sus opciones, permite las transformaciones básicas de los datos, el filtrado y la normalización, el agrupamiento de genes o de condiciones experimentales con el empleo de diferentes medidas de distancia y métodos tales como *k-Means* [70], HCL (*Hierarchical Clustering*) [55], SOM (*Self Organizing Maps*) [71, 72] y SOTA (*Self Organizing Tree Algorithm*) [73, 74]. Para la búsqueda de genes diferencialmente expresados en determinadas condiciones experimentales cuenta con métodos como la prueba *t*, ANOVA [75] y SAM con control del FDR. Este último es uno de los más utilizados en la literatura. MeV tiene otras funcionalidades como RN (*Relevance Networks*) [76], que a partir de una entrada de genes con su perfil de expresión, permite obtener las subredes de genes más interconectadas según un valor mínimo del coeficiente de correlación entre los nodos (genes) definido por el usuario. El software está disponible en http://www.tigr.org/software/tm4/.

GEPAS (Centro de Investigación Príncipe Felipe, Valencia): es una aplicación Web para el análisis de perfiles de expresión de genes [77-79], formada por varios módulos interconectados, que garantizan el preprocesamiento, la normalización (aplicación DNMAD para chips de ADNc y Expresso para Affymetrix) [80], la determinación de expresión diferencial con T-Rex [78, 81] y la ejecución de métodos de *clustering* (SOM, Som Tree [82], SOTArray). Con GEPAS es posible determinar también un predictor de clases con métodos como SVM (*Support Vector Machine*) [83], DLDA (*Diagonal Lineal Discriminant Analysis*), *k* NN [51] y otros implementados en la herramienta Prophet [84]. Para facilitar la interpretación y el significado biológico de conjuntos de genes relacionados, contiene el FatiGO [85] y Fatigo+ [86, 87], que permiten la asociación de los genes a los términos de Gene Ontology, GO [88] y a la base de datos KEGG [89] de vías metabólicas. La aplicación FatiScan [90] permite detectar bloques de genes funcionalmente relacionados (GO, KEGG) en listas ordenadas de genes. Este orden de los genes puede ser el resultado de un análisis de expresión diferencial o de otro criterio ya sea teórico o experimental (http://gepas.bioinfo.cipf.es/).

BRB-Arraytools (*National Cancer Institute-NCI*, USA): es un paquete profesional integrado para la visualización y el análisis estadístico de datos de expresión de genes que es posible instalar como un complemento (*plugin)* del Microsoft Excel [91]. Contiene casi todas las funcionalidades antes mencionadas con mayor énfasis que los demás programas en aspectos relacionados con el diseño experimental (http://linus.nci.nih.gov/BRB-ArrayTools.html).

Anotación funcional de los resultados a través de técnicas de minería de datos

Una vez realizados los análisis que llamaremos "etapa inicial del análisis", se obtienen grupos de genes relacionados, que se hace necesario vincular con otras fuentes de información pública. Teniendo en cuenta la significación estadística de los resultados, a partir de su interpretación y de un análisis integral, pueden surgir hipótesis que deben ser verificadas experimentalmente. La forma más aceptada para realizar este vínculo es a través de redes de genes/proteínas en sistemas que integran diferentes fuentes de datos biológicos, para observar el comportamiento de bloques de genes funcionalmente relacionados en lugar de cada gen en particular y las relaciones entre estos bloques.

Dopazo J [92] refiere dos generaciones de métodos para el análisis e interpretación de datos provenientes de metodologías de alto flujo. La más tradicionalmente usada es la llamada "análisis funcional basado en un umbral", y se realiza en dos etapas, primero se seleccionan los genes de interés basados en un umbral o valor de significación estadística, y luego se determina la frecuencia de los genes en términos biológicamente relevantes. La otra generación llamada "análisis funcional libre de umbral" está en desarrollo en la actualidad. Estos métodos analizan el comportamiento de bloques de genes relacionados funcionalmente, sin realizar un filtrado previo. Para mostrar la superioridad de esta segunda generación de procedimientos, Dopazo J utiliza un estudio de expresión de genes en diabetes mellitus realizado por Mootha y col. [93], en el cual al aplicar una prueba *t* para comparar dos grupos (17 controles con tolerancia normal a la gluco-sa frente a 26 casos, 8 con disminución de la toleran-cia a glucosa y 18 con diabetes mellitus tipo 2), no se obtienen genes diferencialmente expresados con un umbral de 0.05. El autor utilizó el programa Trex del paquete GEPAS y tampoco encontró expresión dife-rencial. Cuando se aplican métodos para determinar la

53. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-9.

54. West M, Blanchette C, Dressman H, Huang E, Ishida S, *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98:11462-7.

55. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863-8.

56. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680-6.

57. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al*. Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.

58. Belitskaya-Levy I. A generalized clustering problem, with application to DNA microarrays. Stat Appl Genet Mol Biol 2006;5:Article2.

59. Bolshakova N, Azuaje F. Estimating the number of clusters in DNA microarray data. Methods Inf Med 2006;45:153-7.

60. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics 2001;17:309-18

61. Smolkin M, Ghosh D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 2003;6:4:36.

62. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. J Natl Cancer Inst 2003;95:14-8.

63. Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, Dudoit S, *et al*. Bioconductor: Open software development for computational biology and bioinformatics. Bioconductor Project Working Papers. Working Paper 1. 2004. Available from: URL: http://www.bepress.com/bioconductor/paper1.

64. Lucas A, Gautier L. Cluster and Tree Conversion. Bioconductor 2006. Available from: URL: http://bioconductor.org/packages/2.2/bioc/html/ctc.html.

65. Pollard KS, Ge Y, Taylor S, Dudoit S. Resampling-based multiple hypothesis testing. Bioconductor's multtest Package. Bioconductor 2005. Available from: http://bioconductor.org/packages/2.2/bioc/html/multtest.html.

66. Wu H, Yang H, Churchill GA. R/MAANOVA: An extensive R environment for the Analysis of Microarray Experiments. Bioconductor 2008. Available from: http://bioconductor.org/packages/2.2/bioc/html/maanova.html.

67. Yang YH. Exploratory analysis for two-color spotted microarray data. Bioconductor 2008. Available from: URL: http://bioconductor.org/packages/2.2/bioc/html/marray.html.

68. Tibshirani R, Chu G, Hastie T. The samr Package. Bioconductor, 2005. Available from: http://www-stat.stanford.edu/~tibs/SAM/Rdist/index.html.

expresión diferencial en estos datos de expresión, no se obtienen genes expresados diferencialmente. Sin embargo, al usar la prueba de segmentación implementada en la aplicación FatiScan y otros programas considerados de segunda generación como GSEA [93, 94], PAGE [95] y SAFE [96], hubo coincidencia al hallar genes con diferencias que, sin ser significativas, tenían una vinculación con la diabetes. Estas son muestras de la superioridad del enfoque de la Biología de Sistema sobre el enfoque anterior, aunque se debe continuar investigando en esta nueva generación de procedimientos.

Verificación de los resultados

En la primera parte de este artículo se enumeran las principales fuentes de variabilidad y de introducción de errores en un experimento de microarreglos [97]. El estudio de decenas de miles de genes en cada experimento simultáneamente, favorece el aumento del número de falsos positivos en los resultados de los análisis estadísticos. De aquí que se utilicen otras técnicas experimentales para confirmar estos resultados.

El RT-PCR cuantitativo en tiempo real (Q-RT-PCR) es una de las técnicas experimentales más sensibles para detectar y cuantificar el ARNm en tejidos. Esto hace que sea uno de los métodos más robustos [98, 99] y comúnmente usado [100-103] para verificar la expresión de genes como resultado del análisis estadístico de datos de microarreglos. Otras técnicas como el análisis por *Northern blot* y los ensayos de protección de ribonucleasas también se han empleado para estos propósitos [104].

Aplicaciones de la tecnología en Oncología

Ejemplos en descubrimiento de clases

- Tamayo y col. usaron microarreglos para estudiar la expresión de genes en las células HL-60 y obtuvieron conjuntos de genes interesantes desde el punto de vista biológico, involucrados en la diferenciación, con el uso del SOM [72].

- Alizadeh y col. descubrieron nuevos subtipos de linfomas a partir del uso de métodos de *clustering* [57].

- Bittner y col. encontraron una subclasificación dentro de los melanomas, no identificada morfológicamente por otras vías [105]. El subconjunto se obtuvo al realizar un análisis matemático a la expresión de genes en una serie de muestras. Además se pudo ver cómo los genes que explicaban la existencia de este subgrupo estaban diferencialmente expresados en melanomas en etapa invasiva.

Ejemplos en comparación de clases

- Cáncer de próstata: esta ha sido una enfermedad muy estudiada con el uso de microarreglos. Estos estudios se pueden clasificar en cuatro tipos fundamentales: los que comparan tejidos normales frente a tumorales [34, 106-108], hiperplasia benigna de la próstata frente a tumorales [109], efectos de tratamientos, por ejemplo, tras el uso de Doxazosin [110] y los que estudian la evolución molecular del cáncer de próstata refractario [111, 112]. El interés fundamental de estos estudios ha sido encontrar los genes que tienen un comportamiento diferencial en estas condiciones experimentales y los genes que coexpresan entre ellos. Si, por otra parte, se estudian las coincidencias en los resultados de experimentos que analizan tejidos normales frente a tumorales, se puede ver que genes como CAMKK2, FASN, SIM2, CAV2, LIM y AMACR han tenido un comportamiento similar en su perfil de expresión en varios de ellos y se tienen evidencias estadísticamente significativas de su expresión diferencial en los grupos comparados.

Ejemplos en predicción de clases

- Cáncer de mama: Van't Veer y col. reportaron la identificación de un conjunto de 70 genes que sirven como predictores de metástasis cuando analizaron sus perfiles de expresión en tumores primarios de mama provenientes de 117 pacientes jóvenes [113]. El trabajo partió del análisis de los perfiles de expresión de 25 000 genes humanos. En un primer paso fueron identificados 5000 genes que mostraron diferencias significativas de su expresión al comparar todas las muestras de tumores frente a una muestra de referencia. Al aplicar un algoritmo de agrupamiento jerárquico no supervisado bidimensional, se pudieron agrupar los tumores de acuerdo a la similitud de expresión en los 5000 genes estudiados y los genes de acuerdo a la similitud de su expresión en el conjunto de tumores analizados. Como resultado, este análisis arrojó el agrupamiento de los tumores en dos grupos principales, uno en el que predominaban los pacientes con pronóstico negativo en los primeros 5 años y el otro en el que predominaba el pronóstico positivo en el mismo periodo, lo que evidenció el poder predictivo de los perfiles de expresión de los genes. Al relacionar los grupos de genes de perfiles similares con los datos histopatológicos, se encontró coincidencia con los reportes de la literatura. Posteriormente se seleccionaron los 78 pacientes que tenían nódulos linfáticos negativos para buscar una firma pronóstico en sus perfiles de expresión. De estos 78 pacientes, 44 habían estado libres de la enfermedad y 34 habían desarrollado metástasis pasados 5 años. Con el objetivo de identificar tumores con buen o mal pronóstico, se desarrolló un método supervisado de clasificación en tres etapas. Primero se identificaron 231 de ellos en el conjunto de los 5000 genes, cuyos perfiles de expresión correlacionaba mejor con el progreso de la enfermedad (coeficiente de correlación < -0.3 o > 0.3), en el segundo paso los genes se ordenaron por este coeficiente de correlación, y en tercer lugar se fue construyendo el predictor agregando secuencialmente conjuntos de 5 genes del extremo superior de la lista ordenada y calculando el error de clasificación usando validación cruzada cada vez que se agregaban genes. La exactitud mayor del predictor (83%) se obtuvo con 70 genes que se propusieron entonces como predictores. Al aplicar este predictor, 13 pacientes quedaron clasificados de forma errónea; de ellos, 5 pertenecían al grupo de buen pronóstico y 8 al de mal pronóstico. Estudios posteriores en 295 pacientes [114] que compararon la supervivencia de los pacientes según el comportamiento del predictor, confirmaron los resultados.

- Cáncer de pulmón: Chen y col. [115] encontraron una firma molecular de 5 genes (ERBB3, LCK, DUSP6, STAT1, MMD) asociada con supervivencia en NSCLC (*Non Small Cell Lung Cancer*). Se estu-

69. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, *et al*. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003; 34(2):374-8.

70. Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev 2000;14:963-80.

71. Kohonen T. Self-organizing formation of topologically correct feature maps. Biological Cybernetics 1982;43(1):59-69.

72. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, *et al*. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907-12.

73. Dopazo J, Carazo JM. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol 1997;44:226-33.

74. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001;17(2): 126-36.

75. Zar JH. Biostatistical analysis. 4th ed., New Jersey: Prentice Hall; 1999, p. 663.

76. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000;97:12182-6.

77. Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo, *et al*. GEPAS: A web-based resource for microarray gene expression data analysis. Nucleic Acids Res 2003;31:3461-7

78. Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, *et al*. Next station in microarray data analysis: GEPAS. Nucleic Acids Res 2006;34: 486-91.

79. Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Díaz-Uriarte R, *et al*. Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res 2005; 33:616-20.

80. Vaquerizas JM, Dopazo J, Díaz-Uriarte R. DNMAD: web-based Diagnosis and Normalization for MicroArray Data. Bioinformatics 2004;20:3656-8.

81. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;20:307-15.

82. Herrero J, Dopazo J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. J Proteome Res 2002; 1:467-70.

83. Vapnik V. Statistical learning theory. John Wiley and Sons 1999. New York.

84. Medina I, Montaner D, Tarraga J, Dopazo J. Prophet, a web-based tool for class prediction using microarray data. Bioinformatics 2007;23:390-1.

85. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20: 578-80.

diaron muestras de 125 pacientes con este tipo de cáncer en arreglos que contenían 672 genes asociados con actividad invasiva en un experimento previo [116] que analizó tejidos normales y de NSCLC. Los genes con un coeficiente de variación menor que el 3% se excluyeron del análisis, y quedaron 485 de los 672 genes. Las 125 muestras se asignaron aleatoriamente en un grupo de entrenamiento y uno de prueba. Con el objetivo de determinar los genes asociados con la muerte o recurrencia de la enfermedad se transformaron los valores de expresión y asignaron códigos según el nivel de intensidad para un análisis de regresión. Las razones *Hazard* de un análisis de regresión univariado mostraron la asociación del nivel de expresión de cada gen con la supervivencia, los genes con una razón Hazard menor que uno, se consideraron genes protectores y una razón mayor que uno, genes de riesgo. Se seleccionó una firma de 16 genes significativamente correlacionados con la supervivencia: 5 de estos 16 genes predijeron la supervivencia de pacientes con este tipo de cáncer de pulmón con un 96% de exactitud. La media de sobrevida de los 101 pacientes evaluados para la búsqueda del predictor fue de 20 meses. Los pacientes clasificados de mayor riesgo usando la firma molecular tuvieron un promedio de sobrevida menor que los de una firma de menor riesgo (20 frente a 40 meses, $p < 0.001$). Esta firma molecular se validó en otros 60 pacientes chinos y en 86 pacientes occidentales de un conjunto independiente de datos públicos de microarreglos en NSCLC. La presencia de una firma de alto riesgo en los tumores se asoció con un incremento en el riesgo de recurrencia y un decrecimiento en la supervivencia.

Conclusiones

En este trabajo se describen las etapas involucradas en la cuantificación y el análisis estadístico de los datos de un experimento de microarreglos. Se espera que en los próximos años se desarrollen nuevos algoritmos y métodos estadísticos que permitan una mejor interpretación de la información contenida en experimentos de microarreglos, además de que estos se complementarán con el uso de otras tecnologías de alto flujo a escala genómica. También deben aparecer soluciones a problemas conocidos como la alta dimensión de las matrices de expresión, que común mente contienen miles de filas (variables, genes) frente a cientos de columnas (observaciones, muestras) y al problema de la comparación e integración de datos de experimentos que utilicen diferentes tecnologías y controles. El desarrollo de métodos estadístico-matemáticos seguirá a la par del desarrollo tecnológico que, sin duda, acompañará en los próximos años la evolución de los microarreglos y las nuevas tecnologías de alto flujo. Su dominio y correcta aplicación serán un factor fundamental para la obtención de resultados de impacto.

86. Al-Shahrour, F, Minguez P, Vaquerizas, JM, Conde, L, Dopazo, J. Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments. Nucleic Acids Res 2005;33:460-4.

87. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas, *et al*. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res 2006;34: 472-6.

88. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25-9.

89. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004; 32:D277-80.

90. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 2005;21:2988-93.

91. Simon RM, *et al*. Analysis of gene expression data using BRB-Array Tools. Cancer Inform 2006;2:1-7.

92. Dopazo J. Functional interpretation of microarray experiments. OMICS. 2006;10: 398-410.

93. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; 34:267-73.

94. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005; 102:15545-50.

95. Kim SY, Volsky DJ. PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 2005;6:144.

96. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21: 1943-9.

97. Miranda J, Bringas R. Análisis de datos de microarreglos de ADN. Parte I: Antecedentes de la tecnología y diseño experimental. Biotecnol Apl (en prensa).

98. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, *et al*. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap resampling scheme. BMC Genomics 2007; 8:140.

99. Dallas PB, Gottardo NG, Firth MJ, Beesley AH, Hoffmann K, Terry PA, *et al*. Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR-how well do they correlate? BMC Genomics 2005;6(1):59.

100. Bektic J, Wrulich OA, Dobler G, Kofler K, Ueberall F, Culig Z, *et al*. Identification of genes involved in estrogenic action in the human prostate using microarray analysis. Genomics. 2004;83(1):34-44.

101. Vawter MP, Ferran E, Galke B, Cooper K, Bunney WE, Byerley W. Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree. Schizophr Res 2004;67(1):41-52.

102. Wiese AH, Auer J, Lassmann S, Nährig J, Rosenberg R, Höfler H, *et al*. Identification of gene signatures for invasive colorectal tumor cells. Cancer Detect Prev 2007;31(4): 282-95.

103. Jura J, Wêgrzyn P, Korostyński M, Guzik K, Oczko-Wojciechowska M, Jarzab M, *et al*. Identification of interleukin-1 and interleukin-6-responsive genes in human monocyte-derived macrophages using microarrays. 2008 (en prensa).

104. Kothapalli R, Yoder SJ, Mane S, Loughran TP. Microarray results: how accurate are they?. BMC Bioinformatics 2002;3:22.

105. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, *et al*. Molecular classification of cutaneous malignant melanoma by gene ex-pression profiling. Nature 2000; 406:536-40.

106. Ashida S, Nakagawa H, Katagiri T, Furihata M, Iizumi M, Anazawa Y, *et al*. Molecular Features of the Transition from Prostatic Intraepithelial Neoplasia (PIN) to Prostate Cancer: Genome-wide Gene-expression Profiles of Prostate Cancers and PINs. Cancer Res 2004;64:5963-72.

107. Zhao H, Ramos CF, Brooks JD, Peehl DM. Distinctive gene expression of prostatic stromal cells cultured from diseased versus normal tissues. J Cell Physiol 2007;210:111-21.

108. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodríguez J, Moskaluk CA, *et al*. Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer. Cancer Res 2001;61:5974-8.

109. Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, *et al*. Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling. Cancer Res 2001;61:4683-8.

110. Zhao H, Lai F, Nonn L, Brooks JD, Peehl DM. Molecular Targets of Doxazosin in Human Prostatic Stromal Cells. Prostate 2005; 62:400-10.

111. Stanbrough M, Bubley GJ, Ross K, Golub TR, Rubin MA, Penning TM, *et al*. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. Cancer Res 2006; 66:2815-25.

112. Tamura K, Furihata M, Tsunoda T, Ashida S, Takata R, Obara W, *et al*. Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. Cancer Res 2007;67:5117-25.

113. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AA, Mao M, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.

114. Van de Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, *et al*. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; 347:1999-2009.

115. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, *et al*. A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. N Engl J Med 2007;356: 11-20.

116. Chen JJ, Peck K, Hong TM, Yang SC, Sher YP, Shih JY, *et al*. Global Analysis of Gene Expression in invasion by a lung cancer model. Cancer Res 2001;61:5223-30.

Recibido en junio de 2008. Aprobado en diciembre de 2008.