Uso de chips de ADN *(microarrays)* en medicina: fundamentos técnicos y procedimientos básicos para el análisis estadístico de resultados

Víctor Moreno y Xavier Solé

Unidad de Bioestadística y Bioinformática. Servicio de Epidemiología y Registro del Cáncer. Instituto Catalán de Oncología. Hospital Duran i Reynals. L'Hospitalet de Llobregat. Barcelona. España.

La tecnología de *microarrays* de ADN permite realizar análisis genéticos sobre miles de genes simultáneamente. El análisis de estos experimentos supone un reto desde el punto de vista estadístico, ya que los métodos clásicos de análisis deben adaptarse a la enorme multiplicidad de hipótesis que se prueban. Además, la gran variabilidad observada en los experimentos y su elevado coste exigen un diseño cuidadoso. En esta revisión se explicará con detalle qué es un microarray de ADNc, cómo funciona y cuáles son sus principales usos. Seguidamente, se abordarán aspectos estadísticos del diseño experimental y de los diferentes apartados del análisis de un *microarray*, desde el procesamiento de la imagen y control de calidad de los datos hasta los tests para identificar genes de interés. Por último se comentarán diferentes técnicas de análisis multivariante que se pueden utilizar para analizar patrones en la expresión de los genes.

Palabras clave: Microarray de ADN. Análisis estadístico. Diseño de experimentos.

Use of DNA chips (microarrays) in medicine: technical foundations and basic procedures for statistical analysis of results

DNA microarray technology allows the assessment of genetic analyses on thousands of genes simultaneously. The statistical analyses of these experiments are challenging since a high number of multiple hypotheses are tested and classical statistical methods need to adapt to this situation. Furthermore, the great variability observed in the experiments and their high cost of them needs a careful design. In this review we will explain what is a cDNA microarray, how it works and its potential uses. Later we will deal with statistical issues of design and analysis, from the image processing and data quality control, to the statistical test of hypothesis to detect interesting genes. Finally we will comment on multivariate methods to detect patterns in gene expression.

Key words: DNA microarray. Statistical analysis. Experimental design.

Introducción

El genoma de los seres vivos es el conjunto de genes que se encuentran distribuidos en cromosomas. Los genes, a su vez, son secuencias de ADN que contienen toda la información necesaria para sintetizar las proteínas, moléculas esenciales para la vida que realizan prácticamente todas las funciones celulares. Cuando un gen se «activa» para dar lugar a su proteína correspondiente, diremos que ese gen se está expresando en esa célula. Es conocido que anomalías en la expresión de los genes pueden llevar a disfunciones celulares, provocando graves enfermedades como el cáncer, entre muchas otras. Los genes que tengan su expresión alterada en un tejido tumoral respecto a un tejido sano del mismo órgano, por ejemplo, serán claros candidatos a tener alguna implicación en el proceso neoplásico. Por lo tanto, la identificación de los genes desregulados es un paso importante para conocer las bases moleculares de muchas enfermedades de carácter genético.

Desde mediados de los años noventa existe la técnica de los *microarrays* de ADN, que permite monitorizar simultáneamente el nivel de expresión de miles de genes en un conjunto de células. Sin embargo, la potencia que nos ofrece esta herramienta implica nuevos retos en lo que se refiere al análisis estadístico. Los datos que se generan con *microarrays*, aparte de tener un gran volumen, se caracterizan por ser altamente variables, por lo que serán básicos tanto el análisis estadístico como el diseño experimental que se plantee para solucionar las diferentes cuestiones biológicas que nos propongamos.

En este trabajo explicaremos primero con más detalle qué es un *microarray* y cómo funciona, para después tratar sobre cuáles son sus principales usos. Seguidamente hablaremos de los diferentes diseños experimentales que se pueden utilizar, y pasaremos a tratar las diversas partes que componen el análisis de un *microarray*, desde el procesamiento de la imagen y control de calidad de los datos hasta el tratamiento estadístico para identificar genes de interés. Finalmente, hablaremos sobre las diferentes técnicas de análisis multivariante que se pueden utilizar para extraer el máximo conocimiento de nuestros datos. La figura 1 muestra un esquema con los aspectos más relevantes de un protocolo de experimentos con *microarrays*.

¿Qué es un *microarray* de ADN y cómo funciona?

Los *microarrays* de ADN son una herramienta que permite realizar análisis genéticos diversos basados en la miniaturización de procesos biológicos. La primera aplicación de esta tecnología fue para medir simultáneamente el nivel de expresión de miles de genes[1]. Las mejoras tecnológicas han perfeccionado la calidad y han ampliado el espectro de aplicaciones, de manera que los *microarrays* se han consolidado como herramientas útiles en investigación genética con aplicaciones en medicina[2,3]. El funcionamiento de los *microarrays* de expresión se basa en la capacidad de las moléculas complementarias de ADN de hibridar entre sí. Pequeñas cantidades de ADN, correspondientes a diversos genes cuya expresión se desea medir, son depositadas en una base de cristal. Para ello se utilizan robots de precisión que usan unas agujas especiales para obtener las moléculas de sus recipientes y depositarlas en las coordenadas adecuadas. A estas muestras de ADN depositadas en el *microarray* las denominaremos dianas. En un *microarray* típico, una su-

Correspondencia: Dr. V. Moreno.
Unidad de Bioestadística y Bioinformática. Servicio de Epidemiología y Registro del Cáncer.
Instituto Catalán de Oncología. Hospital Duran i Reynals.
Gran Vía, km 2,7.
08907 L'Hospitalet de Llobregat. Barcelona. España.
Correo electrónico: v.moreno@iconcologia.catsalut.net

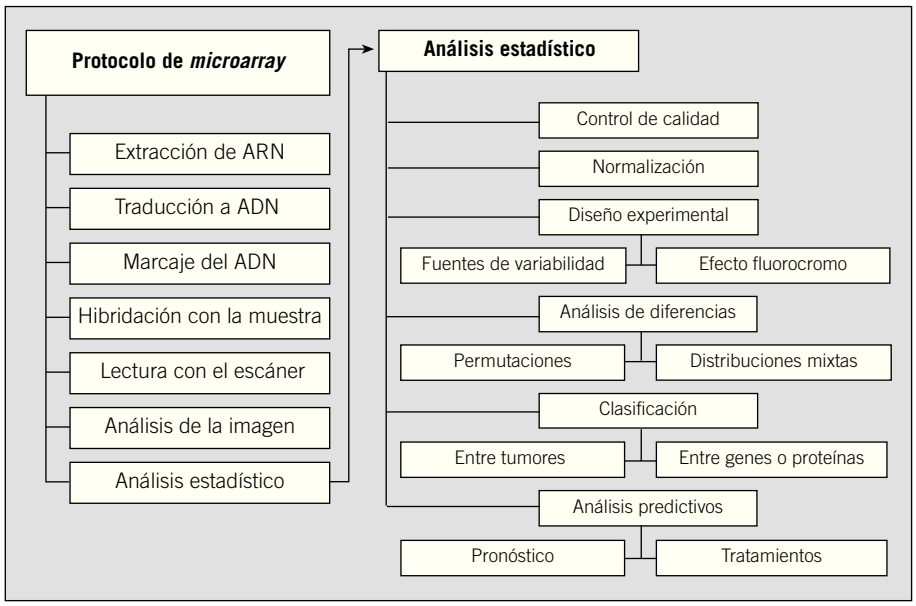

Fig. 1. Protocolo a seguir en un experimento de microarrays *y partes que componen su análisis estadístico.*

perficie de 2 2 cm puede contener más de 10.000 dianas en forma de pequeños puntos separados adecuadamente (fig. 2). De las células que queramos medir su expresión obtendremos una muestra de ARN que se convertirá en ADN complementario (ADNc) y se marcará con una molécula fluorescente. A esta muestra marcada la denominaremos sonda y se enfrentará a las dianas del *microarray*. Cada molécula de ADNc marcada de la sonda se moverá por difusión hacia la diana que contenga su molécula complementaria para hibridarse con ella y quedar fijada allí. Después de un tiempo para que la mayoría de las cadenas complementarias hibriden, el *microarray* se lava y se procede a hacer una medición relativa de la cantidad de ADN de la sonda que ha quedado fijada en cada diana.

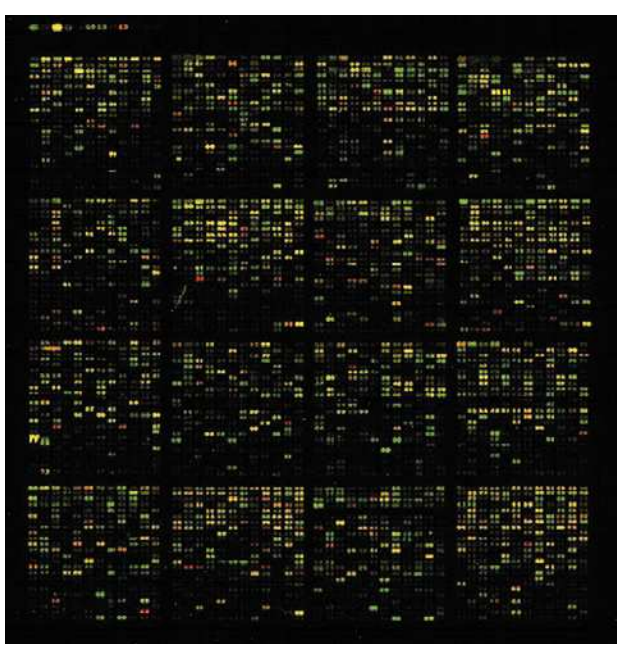

Fig. 2. Imagen de un microarray *de ADNc. Contiene 4.608 clones depositados por duplicado en un soporte sólido, que habitualmente suele ser de vidrio.*

Existe otra tecnología que emplea oligonucleótidos (secuencias cortas de ADN, de unas 15-30 bases)[4]. Estos oligonucleótidos, en lugar de ser depositados en el soporte mediante un robot, son sintetizados directamente sobre el soporte mediante una técnica denominada fotolitografía que es similar a la empleada para confeccionar circuitos microelectrónicos sobre silicio. Esta tecnología requiere una infraestructura muy sofisticada y su empleo por el momento está limitado a unas pocas empresas especializadas entre las que destaca Affymetrix. Para detectar la expresión de un gen se emplea una serie amplia de oligonucleótidos, alrededor de 30, por lo que estos *microarrays* contienen muchas más dianas, lo que es factible porque la fotolitografía permite obtener mayores densidades.

¿Para qué sirven los *microarrays* de ADN?

Las aplicaciones de los *microarrays* se amplían cada día, aunque por el momento hay 3 grandes áreas consolidadas:

El análisis del nivel de expresión génica

Ya se ha mencionado y se explicará con mayor detalle más adelante. En estos experimentos se obtienen datos sobre el nivel de expresión de miles de genes. A partir de estos datos, empleando un diseño experimental correcto y técnicas estadísticas adecuadas, se pueden realizar estudios de diagnóstico y caracterización de tumores u otros tejidos[5-8], identificación de los genes que modifican su expresión tras la administración de fármacos[9] o identificación de genes con valor pronóstico[10,11]. También se han empleado para asignar función a secuencias de ARN que se expresan pero cuya función era desconocida (EST) y para identificar grupos de genes que forman redes de regulación génica[12]. Otras aplicaciones son el diagnóstico de enfermedades infecciosas a partir de la detección del genoma del germen en tejidos[13,14].

Genotipificación

Una muestra de ADN obtenida de un tejido o fluido, adecuadamente amplificada, puede ser estudiada para detectar mutaciones en genes de interés o variantes génicas (poli-

morfismos en un nucleótido, SNP en la terminología de este campo). Esta metodología tiene usos potenciales para la detección de riesgo o susceptibilidad para presentar enfermedades[15,16]. Variantes de estos *microarrays* permiten secuenciar genes con mutaciones conocidas.

Detección del número de copias del ADN

Similar a la técnica de hibridación genómica comparada (CGH), se han diseñado *microarrays* para detectar ganancias o pérdidas alélicas en miles de secuencias, lo que permite obtener mapas cromosómicos mucho más detallados que la CGH tradicional[17]. Estas técnicas tienen interés potencial en el estudio del pronóstico de tumores, ya que éste se halla asociado al nivel de daño genómico. También puede ser útil para detectar nuevos oncogenes y genes supresores de tumores.

¿Cómo se usan los *microarrays* de expresión?

Se describirá en este apartado la metodología empleada en *microarrays* de ADNc (fig. 3). El objetivo del experimento es detectar genes que se expresan en un tejido. El proceso se inicia con la extracción del ARN de la muestra. El ARN es muy inestable y se degrada en pocos minutos, por lo que los tejidos deben ser frescos o congelados inmediatamente tras su obtención. El ARN se convierte en ADNc mediante una transcriptasa reversa y en este proceso se marca con un fluorocromo, es decir, con una molécula que posteriormente emitirá luz al ser excitada mediante una luz láser adecuada.

La sonda de ADNc marcado, que contiene una muestra de los genes que se expresan en el tejido de origen, se hibridará con las secuencias diana depositadas en el *microarray,* que son complementarias a las expresadas. Debido a que la eficiencia del marcado puede ser variable según los genes, y que la cantidad de ADN diana puede no ser igual de un *microarray* a otro, normalmente se realizan experimentos en competición. Éstos consisten en comparar sobre un mismo *microarray* 2 muestras de ADN de tejidos diferentes, cada una marcada con un fluorocromo distinto. Las 2 muestras se mezclan y se hibridan simultáneamente sobre el *microarray*. Este diseño experimental es ventajoso en cuanto que las comparaciones entre las 2 muestras de un *microarray* tienen menor variabilidad que las comparaciones de muestras de diferentes *microarrays* y también se reduce a la mitad el número de *microarrays* necesarios.

Una vez se ha producido la hibridación, el *microarray* se lee con un escáner láser y se obtienen 2 imágenes, una para cada fluorocromo usado, con puntos de luz cuyas intensidades variarán según el nivel de hibridación que se haya producido en cada diana. Estas imágenes se procesan mediante un software que cuantifica la señal de cada punto (diana) para cada fluorocromo (muestra) y elabora una base de datos que será analizada con técnicas estadísticas. Con frecuencia se elabora una imagen superpuesta de las dos en forma de seudocolor. Una de las imágenes se colorea en rojo y la otra en verde. De esta manera, los puntos amarillos tienen señal alta en las 2 muestras en cantidad similar. Los puntos rojos o verdes indican que la expresión de ese gen predomina en una de las muestras (fig. 2).

Diseño de experimentos

Como ya se ha mencionado, en un *microarray* de ADNc se analizan simultáneamente 2 muestras. Esta limitación no impone restricciones en el tipo de estudios que pueden realizarse, pero hace necesario que se utilicen diseños adecuados. Como la tecnología de *microarrays* es muy cara, algunos investigadores están tentados de realizar experimentos sin réplicas. De hecho, existe una serie amplia de técnicas que pretenden identificar qué genes se expresan de manera diferente en 2 tejidos a partir de un único *microarray.* Estos métodos consisten en definir un umbral en la razón de intensidades a partir del que se considera que existe un cambio significativo. El método más simple emplea un valor de umbral fijo que suele ser 2, es decir, la señal de expresión de un canal tiene una intensidad doble que la del otro. Este método no se basa en criterios estadísticos e ignora por tanto la variabilidad observada en la razón de intensidades. Otros métodos que utilizan criterios estadísticos son los de Chen, que define el umbral según la variabilidad observada[18]; el de Sabatti, que emplea información de un experimento de reproducibilidad[19], y el de Newton, que usa una aproximación bayesiana empírica[20]. Hoy día, sin embargo, se reconoce que los experimentos basados en un único *microarray* tienen poca validez, pues los datos tienen gran variabilidad que es necesario controlar con el empleo de un número de réplicas y de tamaños de muestra adecuados[21-26].

Otro aspecto importante para el diseño es el tipo de control a emplear y la distribución de muestras entre *microarrays*. Para aprovechar la reducción de variabilidad en las comparaciones dentro de un *microarray,* el diseño óptimo es aquel que enfrenta en un *microarray* 2 muestras que interesa comparar directamente. Por ejemplo, en la búsqueda de genes que se expresen diferencialmente en tumores respecto al tejido sano se hibridarán conjuntamente las muestras del mismo individuo (fig. 4a).

Si se desea comparar más de dos condiciones entre sí, los diseños (maneras de aparear las muestras) pueden ser variados. Se han utilizado con frecuencia diseños que

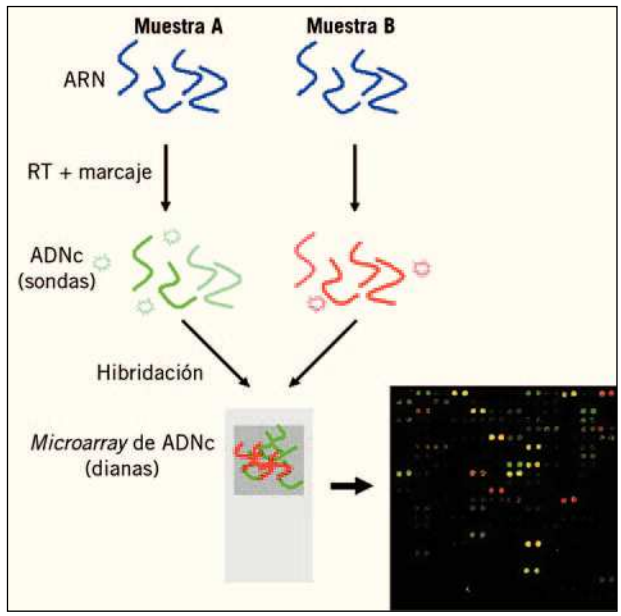

Fig. 3. Esquema del funcionamiento de un microarray *de ADNc. De 2 muestras diferenciadas, A y B, se extrae el ARN, que después será retrotranscrito a ADNc y marcado con unos fluorocromos (moléculas que emiten luz cuando son excitadas). Los 2 ADNc marcados con distintos fluorocromos, llamados sondas en la terminología de los* microarrays*, se hibridan conjuntamente de manera competitiva contra un conjunto de ADNc diana depositados en un soporte de vidrio* (microarray)*, obteniéndose así una imagen como la de la figura.*

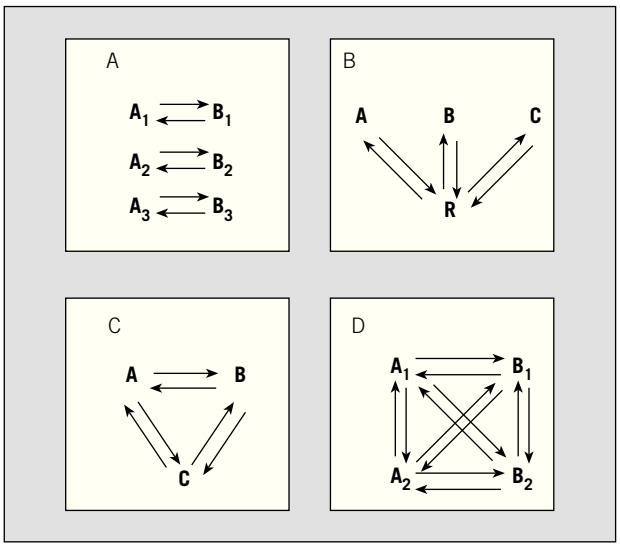

Fig. 4. Diferentes modelos experimentales que se pueden utilizar en los experimentos con microarrays *de ADNc. La dirección de la flecha indica el fluorocromo con que se marca la muestra (base: rojo y punta verde). A: varias muestras apareadas de dos condiciones (A/B). B: cada muestra (A, B, C) se hibrida con una referencia común (R). C: diseño circular de 3 muestras. D: diseño factorial, que permite evaluar simultáneamente 2 factores como tipo de tejido (A/B) y tratamiento (1/2).*

aparean cada muestra con un patrón común, que suele ser una combinación de varias de ellas (fig. 4b). Este diseño es fácil de entender y permite realizar comparaciones individuales entre las muestras, pero no es óptimo. Los diseños circulares (fig. 4c) o factoriales (fig. 4d) son más convenientes para obtener la mejor razón entre la varianza del error (residual) y el número de *microarrays* empleado[21,25].

Análisis de la imagen

La cuantificación de la señal a partir de las imágenes es un proceso muy importante ya que determina los valores que posteriormente se analizarán. En la actualidad existen múltiples herramientas, tanto comerciales como de libre distribución (ScanAlyze, de la Universidad de Stanford), diseñadas exclusivamente para analizar las imágenes de *microarrays*. Las imágenes a analizar suelen ser 2 archivos en formato TIFF en escala de grises de 16 bits, es decir, cada píxel puede tener una intensidad de señal entre 0 y 2^{16} –1 (65.535).

El proceso de análisis de la imagen se puede dividir en 3 etapas. En primer lugar se localizan los puntos a partir de los datos que proporciona el fabricante del *microarray* (cuántos puntos hay, cómo están agrupados, separación teórica entre los centros). A continuación se realiza la segmentación, que consiste en identificar qué píxeles corresponden a un punto y qué píxeles son fondo. Por último, se procede a la cuantificación, a menudo como el promedio o la mediana de las intensidades de los píxeles que forman el punto. En este apartado es importante que el software proporcione varias medidas que puedan ser empleadas como indicadores de la calidad del punto. Medidas típicas son los índices de circularidad, diámetro máximo, perímetro, homogeneidad de la señal dentro del punto, etc. También es necesario obtener una medida del fondo, es decir, el nivel de señal en los píxeles reconocidos como fuera de los puntos. Este valor normalmente se sustrae de la intensidad en el punto para obtener la intensidad neta. La razón entre la señal en el punto y en el fondo también es un índice de calidad importante.

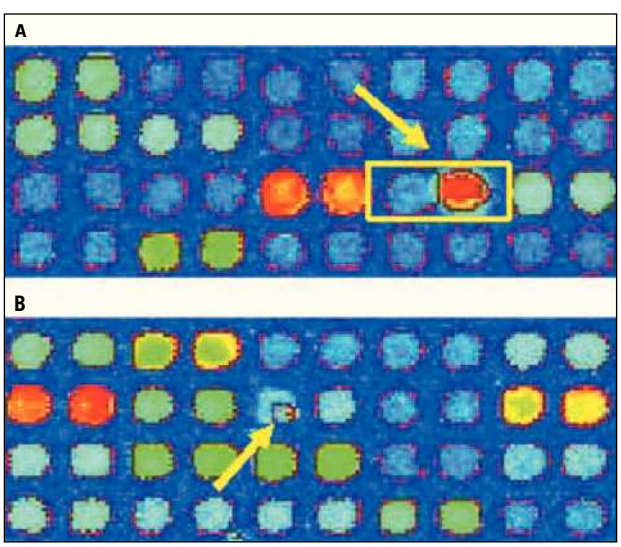

Fig. 5. Control de calidad de la hibridación. A: diferencias de intensidad considerables entre las dos réplicas de un clon. B: problemas en la segmentación debidas a una mota de polvo en el chip.

Análisis estadístico

Una vez se han analizado las imágenes y almacenado los datos, la última parte del proceso consiste en realizar su análisis estadístico. Estos análisis pueden tener diferentes objetivos y emplean técnicas específicas:

Control de calidad de los datos

Consiste en detectar valores incorrectos para su posterior exclusión de los análisis. Estos valores incorrectos pueden surgir por problemas en la calidad de los experimentos o por accidentes durante su manipulación, como rascadas en la superficie del *microarray*. Su detección puede realizarse empleando límites de tolerancia en los indicadores de calidad del punto. Es muy útil que el diseño del *microarray* incluya réplicas de una misma diana. La comparación de las réplicas permite identificar casos discordantes (fig. 5).

Normalización de los datos

El objetivo de la normalización es eliminar la variabilidad sistemática introducida por el proceso técnico que no está relacionada con el nivel de expresión[27]. Las 2 imágenes de un *microarray* se obtienen por separado, cada una con una longitud de onda diferente (normalmente rojo y verde) y una potencia que debe ajustarse de manera independiente para evitar saturación. El ajuste independiente hace que las 2 imágenes no sean comparables en cuanto a intensidad si no se normalizan previamente.

Existen múltiples métodos para normalizar. La mayoría supone que son pocos los genes que cambiarán su expresión, por lo que el promedio estimado por un método robusto debe centrarse. El método más recomendado emplea modelos de regresión de datos suavizados mediante técnicas no paramétricas que capturen la no linealidad como el *lowess*[28], dado que las diferencias entre imágenes suelen tener una distribución variable con la intensidad[27]. En la figura 6a se muestran los datos de un experimento en el que el ADNc de un mismo tejido se marcó con los 2 fluorocromos (R y G). Esperaríamos que la nube de puntos se situara alrededor de la recta con pendiente 1 que pasa por el origen.

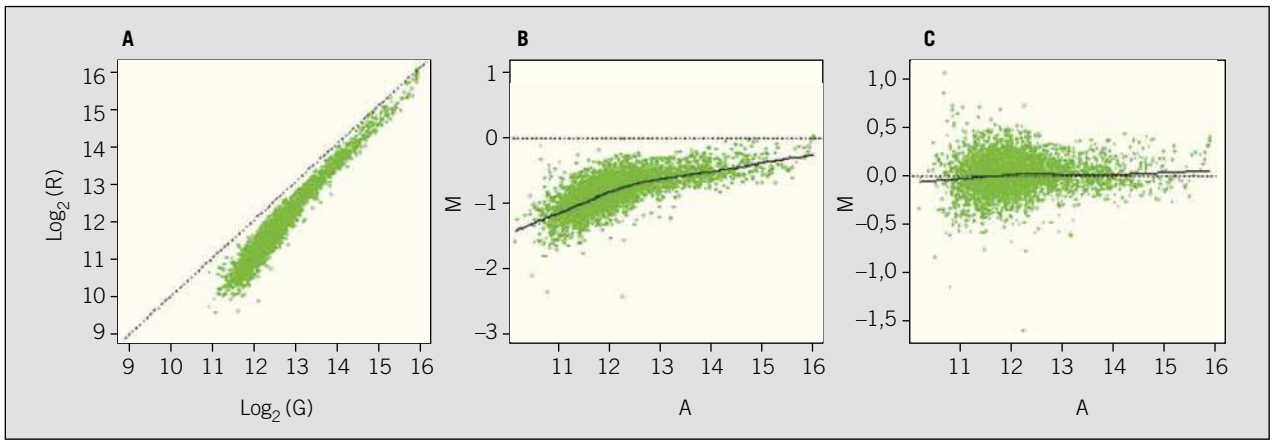

*Fig. 6. Normalización de las intensidades del canal rojo (R) y verde (G) obtenidos en un experimento de reproducibilidad (la misma muestra marcada con diferente fluorocromo). A: gráfica log(R) frente a log(G), donde se aprecia menor señal en la muestra marcada con el fluorocromo rojo. B: gráfica M = log(R/G) frente a A = log(R*G)/2 sin normalizar. Es simplemente una transformación de la anterior donde se aprecia mejor que las diferencias siguen una curva no lineal en función de la intensidad. C: gráfica M frente a A después de normalizar mediante regresión local robusta* (lowess).

Puede apreciarse que la nube de puntos está desplazada por debajo de la línea teórica y muestra una desviación no lineal. La magnitud de la dispersión corresponde a la variabilidad de la técnica, pues recordamos que las 2 muestras corresponden al mismo tejido. Para obtener una mejor visualización, normalmente se transforman los valores según las fórmulas: A = log(R*G)/2 y M = log(R/G). El valor A corresponde al logaritmo de la intensidad media (geométrica) de los 2 colores y el M al logaritmo de la razón de las intensidades. La figura 6b muestra los mismos datos transformados. La curva central a la nube de puntos muestra la estimación no paramétrica de la relación entre M y A, obtenida por el método *lowess* L = *lowess*(A,M). La normalización consiste simplemente en restar a cada punto la diferencia entre M y L. El resultado puede apreciarse en la figura 6c.

El modelo de normalización puede incluir covariables, además de la intensidad media, de las que dependa el valor de la señal. Se recomienda, por ejemplo, estratificar el modelo según la aguja del robot que produjo el *microarray*, o emplear información espacial para eliminar posibles heterogeneidades en la intensidad de la hibridación en diferentes zonas del *microarray* (normalización en 2 dimensiones). Otros métodos, útiles para normalizar múltiples *microarrays*, se basan en igualar no sólo el promedio de razón de intensidades de cada fluorocromo, sino la forma de la distribución (normalización por cuantiles). Este método se usa de forma rutinaria con *microarrays* de oligonucleótidos y puede emplearse también para los *microarrays* de ADNc.

Tratamiento de valores perdidos

Cuando se analizan datos de múltiples *microarrays* puede ser importante dar un tratamiento adecuado a los valores perdidos. Trabajar con los puntos con información completa en todos los *microarrays* puede suponer una pérdida importante de genes valorables. A menudo se emplean técnicas de imputación basadas en medias condicionales del gen respecto al conjunto de los *microarrays* o respecto a los valores de los puntos vecinos[29].

Análisis de la diferencias de expresión a nivel de ARN

Este apartado del análisis pretende determinar qué genes varían su nivel de expresión en función del tejido analizado (por ejemplo, normal y tumor). Dado que en un experimento típico se emplean relativamente pocos casos y se quiere investigar si existen diferencias en miles de variables (genes), las técnicas estadísticas clásicas no son adecuadas sin las correspondientes modificaciones. Por un lado, no puede asegurarse la normalidad de los datos que requieren las pruebas estadísticas clásicas. Por otro lado, la tasa de resultados falsamente positivos puede ser muy elevada si no se emplean correcciones que tengan en cuenta la multiplicidad de hipótesis que se prueban. Las soluciones propuestas para estos problemas son emplear tests de permutaciones para evaluar empíricamente el nivel de significación para cada gen y, posteriormente, controlar la tasa global de resultados falsos positivos mediante un ajuste de los valores p que tengan en cuenta las múltiples comparaciones realizadas.

Los tests de permutaciones construyen la distribución de probabilidad empírica a partir de los propios datos mediante muestreo múltiple[30,31]. Se emplea un test estadístico clásico como la t de Student u otra prueba adecuada según el tipo de variable o diseño de estudio. Se calcula, para un gen, el valor del test que compara los grupos (valor observado). Se repite el mismo test múltiples veces de manera que cada vez la asignación del grupo (normal o tumor) se cambia al azar. De esta manera se simula la situación en la que no hay diferencias, pues la asignación de cada muestra a uno u otro grupo es aleatoria. Finalmente, el valor p se calcula a partir del percentil que ocupa el valor observado en la distribución de valores obtenidos por permutación. Con unos cientos de permutaciones suele ser suficiente para obtener el valor p, pero pueden precisarse varios miles si se desea diferenciar entre valores pequeños (muy significativos). Este proceso se repite por separado para cada gen.

Los métodos para evitar resultados falsos positivos corrigen los valores p para controlar el nivel global de significación. Existen varios métodos, desde el más sencillo de Bonferroni, que consiste en considerar significativos sólo aquellos valores p inferiores al cociente entre alfa y el número de tests. Este método es muy conservador pues asume que cada test es independiente, lo cual probablemente es falso en el contexto del análisis de múltiples genes en *microarrays* de ADN, ya que la expresión de algunos genes puede estar correlacionada. Otros métodos de control del nivel global de significación emplean procedimientos adaptados y tienen en cuenta la correlación entre tests. Los más empleados en el contexto de *microarrays* utilizan también métodos de remuestreo como el min-P y max-T[32].

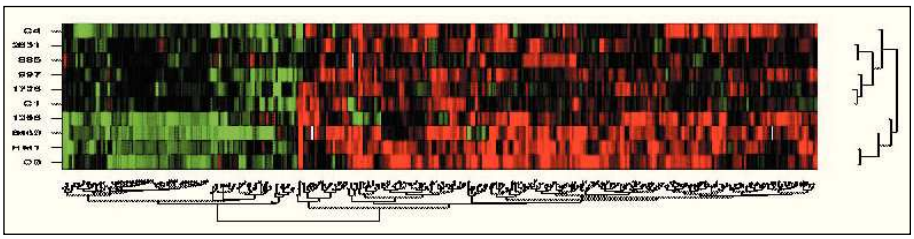

Fig. 7. Análisis de conglomerados (clusters) *de dos dimensiones para una selección de 449 genes de un* microarray. *Los genes están dispuestos en columnas. En filas se han seleccionado las razones de intensidades de 10 experimentos comparando tejido tumoral con el tejido normal del mismo individuo. Los genes con color rojo presentan una sobreexpresión en el tumor, mientras que los de tonalidad verde están más expresados en el tejido sano.*

También se ha propuesto una serie de procedimientos de análisis que se basan en métodos bayesianos puros o bayesianos empíricos. Algunos de estos procedimientos son interesantes ya que proponen que los resultados observados provienen de una mezcla de dos tipos de genes, los que no cambian su expresión (la mayoría) y los que cambian. Para estimar qué genes pertenecen a cada grupo se emplean mixturas de distribuciones que se rigen por una serie de parámetros. Con frecuencia los valores de estos parámetros se estiman a partir de la información que aporta el propio experimento (métodos bayesianos empíricos), pero otras veces los propone el investigador, que puede haberlos obtenido de otros estudios (métodos bayesianos puros). En cualquier caso, estos métodos son interesantes pues tienden a suavizar los valores extremos, que suelen ser la causa de resultados falsamente positivos[33].

Clasificación de muestras o genes

A partir de los genes que muestren expresión diferencial se puede intentar buscar patrones con el objetivo de clasificar las muestras[34-38]. Posteriormente, tras evaluar las características de las muestras agrupadas, se pueden identificar «genes prototipo» que definan los grupos. También se pueden buscar grupos de genes que muestren un patrón de expresión diferencial similar. Esto puede ser útil para asignar función a genes o secuencias que se expresan que hasta ahora la tienen desconocida. Estos análisis también se han empleado para identificar redes de regulación génica[39,40].

Existen múltiples métodos de clasificación automática. Los más utilizados se basan en técnicas de análisis de conglomerados o *clusters* jerárquicos, que pueden emplear diferentes distancias y algoritmos[35]. Estas técnicas generan un gráfico en forma de árbol (dendograma) con la jerarquía obtenida. El problema suele ser decidir por dónde cortar el árbol para definir el número de grupos identificados. Como alternativa, se puede definir *a priori* el número de grupos que se desea identificar y usar métodos que reparten las observaciones de manera óptima entre los grupos. Entre estas técnicas se encuentran el «*k-means*» y los «*Self Organizing Maps*»[36,41]. Otros métodos se basan en mixturas paramétricas *(«model based clustering»)*[42]. Los análisis de *clusters* se pueden aplicar para agrupar muestras o agrupar genes. También pueden realizarse las 2 agrupaciones simultáneamente, lo que permite interpretar más fácilmente los resultados. La figura 7 muestra una doble clasificación (muestras en columnas y genes en filas) donde se aprecian 2 grupos de muestras y 2 de genes. La imagen en seudocolor muestra en rojo genes con expresión aumentada y en verde genes con expresión disminuida. A menudo las técnicas de *clusters* se combinan con técnicas de reducción de la dimensionalidad como el análisis de componentes principales[43] y la regresión de mínimos cuadrados parciales *(«partial least squares»)*[44].

Cuando las muestras están caracterizadas por variables que interesa diferenciar, los métodos de clasificación supervisada son más eficaces. Ejemplos de aplicación son la discriminación en función del tipo de tejido (normal o tumoral) o en función del tipo celular o pronóstico. Entre los métodos de clasificación supervisada destacan los de regresión discriminante, con múltiples versiones (lineal, cuadrática, logística, etc.), las redes neurales, los árboles, etc. Una excelente revisión de estos métodos aplicados al análisis de *microarrays* puede encontrarse en el libro de Hastie et al[45].

Hay una amplia disponibilidad de software de uso libre para analizar datos de *microarrays*. Merece una mención especial el proyecto Bioconductor (www.bioconductor.org), que contiene numerosas herramientas para análisis gráfico y estadístico basadas en el software R (www.r-project.org).

Discusión

Los *microarrays* para análisis genéticos se están consolidando como una tecnología útil, a pesar de que es relativamente reciente y todavía adolece de limitaciones técnicas como una gran variabilidad. La tecnología mejora día a día, por lo que los problemas de escasa reproducibilidad se solventarán a corto plazo.

El campo que ha recibido con mayor entusiasmo esta tecnología es la oncología, donde ya se han publicado interesantes resultados en cuanto a clasificación molecular y predicción de pronóstico en varios tumores[5-7,37,46,47]. Las aplicaciones a otros campos de la medicina como cardiología[48], neumología[49] o reumatología[50], entre otras, también son prometedoras.

Uno de los principales retos que afronta esta tecnología es evitar un excesivo entusiasmo en el momento de reportar resultados para evitar frustraciones por falsos positivos. Cada experimento con *microarrays* evalúa miles de hipótesis simultáneamente, por lo que un análisis ligero, sin las debidas correcciones estadísticas, puede generar resultados falsamente significativos. La identificación de señales interesantes en un *microarray* debe ser verificada con otras técnicas y, como en otros campos del conocimiento, es necesario que experimentos en muestras independientes repliquen los resultados antes de adoptarlos como válidos. Por el momento los *microarrays* se utilizan fundamentalmente en investigación, pero dentro de poco aparecerán aplicaciones clínicas para diagnóstico o evaluación de riesgo.

Agradecimientos

Esta revisión recoge la experiencia de la Unidad de Bioestadística y Bioinformática del Instituto Catalán de Oncología en colaboración con el Grupo de Microarrays de ADN del Instituto, que investiga en aplicaciones de los *microarrays* al estudio del diagnóstico y pronóstico del cáncer colorrectal. Los investigadores del grupo son G. Capellà, M.A. Peinado, M. Grau, E. Vendrell, A. Obrador, G. Tarafa, E. Guinó, J. Valls, X. Solé y V. Moreno. El grupo cuenta con financiación del Fondo de Investigaciones Sanitarias (FIS 96/0797, FIS 00/0027, FIS 01/1264) y de la CICYT (SAF 99/0103, SAF 00/81-C2). Este grupo es miembro de las Redes de Temáticas de Investigación Cooperativa en Cáncer (C03/10) y en Epidemiología y Salud Pública (C03/09), financiadas por el Instituto Carlos III, Ministerio de Sanidad y Consumo.

REFERENCIAS BIBLIOGRÁFICAS

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467-70.
2. Aitman TJ. DNA microarrays in medical practice. BMJ 2001;323:611-5.
3. Petricoin EF, 3rd, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, et al. Medical applications of microarray technologies: a regulatory science perspective. Nat Genet 2002;32(Suppl):474-9.
4. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21:20-4.
5. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.
6. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, et al. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. Cold Spring Harb Symp Quant Biol 1999;64:71-8.
7. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 2002;18:1216-26.
8. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. J Comput Biol 2000; 7:559-83.
9. Cunningham MJ, Liang S, Fuhrman S, Seilhamer JJ, Somogyi R. Gene expression microarray data analysis for toxicology profiling. Ann N Y Acad Sci 2000;919:52-67.
10. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937-47.
11. 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.
12. Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imato S, et al. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. DNA Res 2003;10:19-25.
13. Chizhikov V, Wagner M, Ivshina A, Hoshino Y, Kapikian AZ, Chumakov K. Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization. J Clin Microbiol 2002;40:2398-407.
14. Kozal MJ, Shah N, Shen N, Yang R, Fucini R, Merigan TC, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753-9.
15. Irizarry K, Kustanovich V, Li C, Brown N, Nelson N, Wong W, et al. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat Genet 2000;26:233-6.
16. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244-55.
17. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23:41-6.
18. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. Ratio statistics of gene expression levels and applications to microarray data analysis. Bioinformatics 2002;18:1207-15.
19. Sabatti C, Karsten SL, Geschwind DH. Thresholding rules for recovering a sparse signal from microarray experiments. Math Biosci 2002;176:17-34.
20. 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.
21. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res 2001;77:123-8.
22. Lee ML, Lu W, Whitmore GA, Beier D. Models for microarray gene expression data. J Biopharm Stat 2002;12:1-19.
23. Lee ML, Whitmore GA. Power and sample size for DNA microarray studies. Stat Med 2002;21:3543-70.
24. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 2000;97:9834-9.
25. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579-88.
26. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002;18:546-54.
27. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002;32(Suppl):496-501.
28. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. JASA 1979;74:829-836.
29. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520-5.
30. Good P. Permutation tests. 2nd ed. New York: Springer, 2000.
31. Tsai CA, Chen YJ, Chen JJ. Testing for differentially expressed genes with microarray data. Nucleic Acids Res 2003;31:e52.
32. Westfal PHY. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: John Wiley & Sons, 1993.
33. Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002;23:70-86.
34. Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. Pharmacogenomics 2003;4:41-52.
35. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863-8.
36. 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.
37. Van Ruissen F, Jansen BJ, De Jongh GJ, Van Vlijmen-Willems IM, Schalkwijk J. Differential gene expression in premalignant human epidermis revealed by cluster analysis of serial analysis of gene expression (SAGE) libraries. Faseb J 2002;16:246-8.
38. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.
39. Reinke V. Functional exploration of the C. elegans genome using DNA microarrays. Nat Genet 2002;32(Suppl):541-6.
40. Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol 2003;4:R6.
41. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999;451:142-6.
42. Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. Genome Biol 2002;3:RESEARCH0009.
43. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics 2001;17:763-74.
44. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18:39-50.
45. Hastie TT R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2001.
46. Anbazhagan R, Tihan T, Bornman DM, Johnston JC, Saltz JH, Weigering A, et al. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. Cancer Res 1999;59:5119-22.
47. Alaiya AA, Franzen B, Hagman A, Dysvik B, Roblick UJ, Becker S, et al. Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles. Int J Cancer 2002;98:895-9.
48. Barrans JD, Allen PD, Stamatiou D, Dzau VJ, Liew CC. Global gene expression profiling of end-stage dilated cardiomyopathy using a human cardiovascular-based cDNA microarray. Am J Pathol 2002;160:2035-43.
49. Geraci MW, Moore M, Gesell T, Yeager ME, Alger L, Golpon M, et al. Gene expression patterns in the lungs of patients with primary pulmonary hypertension: a gene microarray analysis. Circ Res 2001;88:555-62.
50. Thornton S, Sowders D, Aronow B, Witte DP, Brunner HI, Giannini EH, et al. DNA microarray analysis reveals novel gene expression profiles in collagen-induced arthritis. Clin Immunol 2002;105:155-68.

Este suplemento ha sido posible gracias a la colaboración desinteresada del INSTITUTO DE FORMACIÓN NOVARTIS y a su esfuerzo mantenido por el desarrollo y actualización del conocimiento científico entre los profesionales de la salud.