Luces y sombras del análisis de expresión génica utilizando microarrays. Un ejemplo en cerdo ibérico

A.I. Fernández, C. Óvilo, A. Fernández, C. Barragán, M.A. Toro, C. Rodríguez, L. Silió

Departamento de Mejora Genética Animal, INIA, Madrid
E-mail: avila@inia.es

Resumen
La tecnología de los microarrays de expresión es la herramienta ideal para el estudio de patrones de expresión de miles de genes de forma simultánea. Sin embargo existe gran variabilidad de resultados atribuible a los aspectos técnicos y de análisis estadístico. En este trabajo presentamos algunos de los problemas surgidos en el estudio de las diferencias de expresión en hígado de cerdos ibéricos para los tratamientos sexo y alimentación empleando microarrays de Affymetrix. Los datos de expresión normalizados fueron analizados siguiendo dos aproximaciones de la metodología de los modelos mixtos. Para ambos tratamientos las diferencias de expresión detectadas fueron dependientes del modelo de análisis y solo un pequeño número de genes diferencialmente expresados fueron coincidentes en ambas estrategias estadísticas. Algunas de estas diferencias de expresión fueron validadas por PCR cuantitativa. Además identificamos errores de diseño y falta de anotación de las sondas del array. Los resultados de este estudio nos han permitido detectar diferencias de expresión de algunos genes de interés, pero también remarcan la necesidad de realizar estudios complementarios que confirmen las diferencias de expresión reveladas a través de la tecnología de los microarrays.

Palabras clave: Microarray, Expresión diferencial, PCR cuantitativa, Cerdo

Summary
Lights and darkness of gene expression analysis using microarrays: an example in Iberian pigs
Expression microarray technology is the ideal tool for the study of thousands of gene expression patterns simultaneously. However there is a great variability of results attributed to technical and statistical analysis aspects. In this work we present several of the arisen problems of a differential expression study in liver of Iberian pigs under the treatments sex and feeding level using Affymetrix microarray. Normalized expression data were analyzed following two approaches of the mixed model methodology. In both treatments the detected differential expressions were dependent of the statistical model and just a small number of genes were coincident between both statistical strategies. Some of the expression differences were confirmed by quantitative PCR. Besides, we have identified design mistakes and missing annotation of the array probes. The results of this study have allowed us to detect differential expression of interesting genes, but it pointed out the necessity of carrying out complementary studies in order to confirm the differential expressions revealed using microarrays technology.

Key words: Microarray, Differential expression, Quantitative PCR, Pig

Introducción

La tecnología de los microarrays de expresión es la herramienta ideal para el estudio de patrones de expresión de miles de genes de forma simultánea en un mismo individuo, tejido o célula. Los estudios en la especie humana y ratón, basados en el uso de microarrays, han permitido identificar diferencias de expresión de genes implicados en diversos procesos biológicos, como las respuestas a enfermedades o a cambios ambientales e incluso han permitido identificar nuevos genes y funciones génicas. En el caso de las especies domésticas, los primeros trabajos trataron de hacer uso de microarrays de humano a través de hibridaciones cruzadas (Hernández *et al.*, 2003), aunque en los últimos años se han publicado los primeros resultados de análisis de microarrays de expresión específicos de especies como la vaca o el cerdo. El estudio de diferencias de expresión génicas usando microarrays implica múltiples pasos desde el diseño del experimento hasta la identificación de diferencias de expresión estadísticamente fiables y biológicamente explicables. Se tiende a establecer protocolos estandarizados que eviten la aparición de falsos positivos, falsos negativos y que permitan obtener resultados comparables entre distintos estudios. Existe una gran variabilidad de resultados atribuible a los aspectos técnicos de esta metodología como son la plataforma de hibridación usada (microarrays caseros, comerciales, ADN copia, oligos cortos, oligos largos, uno o dos canales), el protocolo de extracción de ARN (mayor o menor pureza y cantidad), la metodología de síntesis de ADN copia, de marcaje, hibridación, impresión y escaneado. Asimismo, en cuanto al análisis de datos procedentes de microarrays existen diferentes estrategias para llevar a cabo su normalización, la determinación de las diferencias de expresión, y el establecimiento de umbrales de falsos descubrimientos. Esta diversidad de métodos estadísticos contribuye también a la dificultad en comparar resultados publicados. El principal reto del análisis estadístico de los datos es que se contrastan miles de genes con un muy reducido número de muestras, debido al alto coste de esta técnica. El reducido número de muestras hace necesaria una validación de resultados a través de la cuantificación relativa de la expresión utilizando PCR cuantitativa (qPCR). Dadas sus características, mayor sensibilidad y especificidad y menor coste, la qPCR permite además de precisar la estima de las diferencias de expresión, validar los resultados en un mayor número de muestras.

En cuanto a la anotación e interpretación biológica de los resultados, la mayor fuente de información en las especies de animales domésticos procede de la especie humana. En estos momentos se están realizando importantes esfuerzos por desarrollar herramientas bioinformáticas con el propósito de facilitar el proceso de anotación (ej. EasyGo, David database, NetAffy). Sin embargo en la actualidad parecen seguir existiendo errores de asignación y falta de anotación, esto junto con el desconocimiento de muchas de las rutas génicas que controlan los procesos biológicos hace de la interpretación biológica de los resultados obtenidos una tarea compleja.

El presente trabajo se propone explicar las aportaciones y dificultades que supone el uso de esta metodología a partir de un sencillo estudio de análisis de la expresión génica diferencial en cerdo para dos tratamientos (sexo y alimentación) empleando microarrays.

Material y métodos

Las muestras de hígado analizadas en este trabajo corresponden a cerdos Ibéricos pertenecientes a cinco familias de cuatro her-

manos completos de la misma camada. Cada una de las familias está compuesta por dos individuos de cada sexo donde un macho y una hembra de cada familia fue alimentado con una dieta casi *ad libitum* y el otro macho y hembra de cada familia con una dieta restringida al 75% de la ingesta *ad libitum*. Las muestras de ARN de hígado fueron extraídas con el kit RiboPure (Ambion) que garantizaba la cantidad (medida con el equipo NanoDrop) y calidad óptima (medida con el bioanalizador Agilent 2100), para realizar estudios de cuantificación de la expresión. Muestras de ARN de hígado de ocho individuos, dos familias, fueron hibridadas con chips porcinos de Affymetrix (Affymetrix Porcine Genechip TM) que incluyen 20.201 genes representados por 23.937 sets de sondas y cada set está constituido por 11 sondas diferentes diseñadas a lo largo del transcrito. La síntesis del correspondiente ADN copia, marcajes, hibridaciones y escaneado se realizó a través del servicio del hospital Vall d`Hebrón (Barcelona). La calidad de las hibridaciones y normalización de los datos se llevaron a cabo utilizando el paquete affyPLM y la función RMA del programa Bioconductor.

El análisis estadístico de los datos normalizados se realizó utilizando la metodología de los modelos mixtos siguiendo dos estrategias diferentes:

I. Análisis conjunto de todos los genes (Byrne *et al.*, 2005), utilizando el modelo:

y_{ijkl}= *media*+ gen_i+ $sexo_j$ + *nivel alimentación*$_k$+ *(gen x sexo)*$_{ij}$ + *(gen x nivel alimentación)*$_{ik}$+ *error*

II. Análisis gen a gen con una aproximación bayesiana (Programa GEAMM, desarrollado por Casellas *et al.*, 2008), siguiendo el modelo:

y_{jkl}= *media*+ $sexo_j$ + *nivel alimentación*$_k$+ *error*

El umbral de falso descubrimiento (FDR=0.05) se determinó como describieron Benjamini y Hochberg en 1995.

La anotación de las sondas se llevó a cabo en primer lugar utilizando un archivo suministrado por Affymetrix y posteriormente confirmando la anotación con las diversas herramientas disponibles en la red principalmente NetAffy, EasyGo y David database. La validación de resultados se realizó por qPCR utilizando el método de detección Sybr Green en un ABI 7500 Fast y utilizando dos genes (*GADPH* y *BM2*) como controles endógenos. Los cálculos de la cantidades relativas de expresión se efectuaron utilizando el programa geNorm (http://medgen.ugent.be/genorm).

Resultados y discusión

Calidad de las hibridaciones

Todas las muestras de ARN utilizadas en el análisis superaron los requisitos establecidos en cuanto a calidad (RNA Integrity Number >8) y cantidad (7µg) para llevar a cabo las hibridaciones en los chips de Affymetrix. Asimismo, todas las hibridaciones pasaron los controles de calidad de hibridación, degradación y marcaje establecidos por Affymetrix para llevar a cabo el análisis de las diferencias de expresión, lo que permite minimizar la posibilidad de obtener falsos positivos debidas a este tipo de variaciones técnicas.

Análisis de los datos de expresión y anotación

El análisis de las diferencias de expresión génica entre machos y hembras en tejido hepático utilizando el modelo I de análisis conjunto permitió detectar 306 sets de sondas diferencialmente expresados (DE), de

los que un 10% no está anotado. Mediante el análisis bayesiano con el modelo II gen a gen también se detectó un elevado número de sets de sondas DE, concretamente 324. En ambos sets DE, se observó una sobrerrepresentación de genes del cromosoma Y, además de identificar sondas DE de interés como aquellas que corresponden a genes de la familia proteínica citocromo P450, proteínas que participan en la metabolización de compuestos xenometabólicos asociados a las diferencias sexuales en cuanto a resistencia y/o tolerancia a tóxicos y en la generación del escatol (determinante del olor sexual en machos). Sin embargo, entre ambos análisis (conjunto de genes /gen a gen) sólo coincidieron 120 sets de sondas DE.

Como ejemplo ilustrativo de los problemas de interpretación de los resultados, entre los sets DE se identificaron dos diseñados para un mismo gen, *EIF2S3*, de los que sólo uno se mostraba estadísticamente significativo en el análisis conjunto, mientras que ambos sets eran estadísticamente significativos utilizando el modelo gen a gen. Las diferencias de expresión que presentaban los dos sets de sondas del gen se daban en sentidos opuestos, mientras que para uno de los sets (EIF2S3-I) los machos presentaban sobreexpresión del orden de 20-30 veces más que las hembras, para el otro set (EIF2S3-II) las hembras mostraban 1,5 veces más expresión que los machos. Esto sugería la existencia de dos tránscritos diferentes para el mismo gen, información no suministrada por la anotación de Affymetrix. El análisis comparativo de la secuencias de ambos tránscritos corrobora esta hipótesis, ya que entre ambas existe una región común, región codificante, y una región específica de cada transcrito para el extremo 3´ no codificante. Sin embargo en el diseño de los dos sets en el microarray hay sondas complementarias tanto a la región común, como a la región específica de los transcritos, por lo tanto es posible que se infravaloren las diferencias de expresión estimadas ya que existe la posibilidad de que se produzcan hibridaciones tanto de un tránscrito como del otro en ambos sets.

En cuanto a las diferencias de expresión génica en tejido hepático entre los animales de los dos niveles de alimentación se detectaron 231 sets de sondas DE utilizando el modelo de análisis conjunto, de los que el 1% no está anotado. En el análisis gen a gen se detectó un número bastante más reducido de sets de sondas DE, en concreto 87, de las que el 3% tampoco tiene anotación. En este caso, los resultados se mostraron más dispares que los obtenidos para el análisis entre sexos, es decir, los resultados eran más dependientes del tipo de análisis estadístico, y no sólo presentaban diferencias en cuanto a la cantidad de sondas identificadas como DE, sino que de entre éstas sólo 10 sets de sondas coincidían como DE en ambos análisis. De entre los sets de sondas DE en el análisis conjunto, se identificaron una gran proporción de genes relacionados con el metabolismo de grasas, carbohidratos y proteínas. Sin embargo, al utilizar el modelo de análisis gen a gen, la proporción de genes que "a priori" (errores y falta de anotación y/o función) se relacionan con el metabolismo de grasas, carbohidratos y proteínas es mucho menor.

Validación por qPCR

Una etapa imprescindible en los análisis de las diferencias de expresión con microarrays es la validación de algunos resultados a través de qPCR, especialmente cuando son dependientes del modelo de análisis. El primer paso al medir la cantidad de expresión por qPCR es la selección y validación de genes endógenos, estos son los genes que se utilizan como controles internos y que per-

miten corregir las diferencias en las medidas de expresión no debidas al tratamiento sino a diferencias técnicas (cantidades, pipeteo, eficacia síntesis de ADN copia, etc.). En general, los genes endógenos lo son en determinados tejidos, poblaciones y estados fisiológicos (Vandesompele *et al.*, 2002) y por tanto es imprescindible su validación en el/los tejidos a analizar antes de llevar a cabo el estudio de las diferencias de expresión, y generalmente se considera más fiable el uso de más de un gen endógeno. En este estudio se seleccionaron seis genes (*BM2*, *TBP*, *TOP2B*, *GADPH*, *ACTB*, *EEF2*) considerados endógenos en la especie porcina y se estimó su estabilidad (geNorm) y eficiencia en la PCR, de forma que se determinó que los genes *GADPH* y *B2M* eran los que mejor se comportaban como genes endógenos en el tejido hepático, y por tanto fueron los que se usaron como controles internos.

De entre los sets de sondas DE se realizó una selección de genes a validar por qPCR a través de diferentes criterios. La posibilidad de que se hubiesen infravalorado las diferencias de expresión entre sexos de los tránscritos del gen *EIF2S3*, debido al diseño de las sondas del array para ambos tránscritos, unido a los resultados dispares entre los análisis estadísticos, hizo aconsejable su validación por qPCR. Para ello se diseñaron dos parejas de cebadores en las regiones específicas de cada tránscrito, lo que permitió determinar específicamente la cantidad de expresión relativa de cada uno, además de extender el análisis al resto de animales no incluidos en las hibridaciones. Los resultados del análisis por qPCR aparecen representados de forma gráfica en la figura 1, donde se muestran los niveles de expresión relativa a los genes endógenos. Se trata de dos transcritos cuya expresión depende del sexo y estas diferencias condicionadas por el sexo son estadísticamente significativas. De hecho, EIF2S3-I no parece expresarse en hembras, al menos no es detectable (figura 1 A), mientras que sí en machos. Sin embargo, la sobreex-

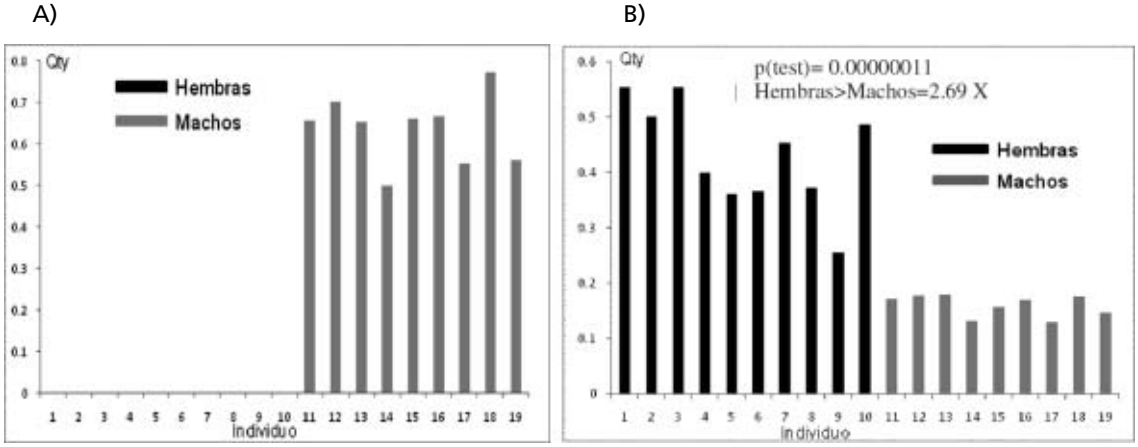

Figura 1. Representación gráfica de las medidas relativas de expresión en machos y hembras de los dos tránscritos del gen *EIF2S3* medidas a través de qPCR. A) Expresión del tránscrito EIF2S3-I. B) Expresión del tránscrito EIF2S3-II.

Figure 1. Graphic representations of the relative expression measures in males and females of the two EIF2S3 *gene transcripts measured by qPCR. A) EIF2S3-I transcript expression. B) EIF2S3-II transcript expression.*

presión del transcrito EIF2S3-II en hembras es 2,69 veces mayor que en machos, valor superior al estimado con los microarrays (1.5 veces). Estos resultados pueden atribuirse a un mal diseño de los sets de sondas de estos transcritos en el array.

De entre los sets de sondas DE entre niveles de ingesta utilizando el modelo de análisis conjunto, se identificaron los genes *fatty acid binding protein 3* (*FABP3*) y *stearoyl-CoA desaturase* (*SCD*) que codifican para importantes enzimas del metabolismo de ácidos grasos y son considerados interesantes genes candidato para caracteres de calidad en cerdo. Sin embargo, en los resultados del modelo de análisis gen a gen ninguno de ellos se detectó como DE. Por ello, dado su interés como genes candidato se validaron a través de qPCR. Los resultados de esta validación aparecen representados de forma gráfica en las figuras 2 A y B, donde se muestran los niveles de expresión relativa a los genes endógenos tanto del gen *FABP3* (A) como del *SCD* (B). En estos genes no se detectan diferencias de expresión significativas condicionadas por la restricción de la dieta. Estos resultados, junto con los obtenidos para el análisis entre sexos, cuestionan que el modelo de análisis conjunto de los genes sea el más apropiado para nuestro diseño experimental.

En el presente estudio se ha empleado la tecnología de microarrays de expresión con el objetivo de detectar diferencias de expresión génica en tejido hepático porcino entre machos y hembras, o debidas al nivel de ingesta. Los resultados han permitido detectar diferencias de expresión en genes interesantes por diversos aspectos (ej. relacionados con el metabolismo de nutrientes, generación del escatol). Sin embargo se aprecia que la información aportada por esta tecnología plantea problemas de interpretación estadística, técnica y biológica. Se requiere realizar tareas complementarias como son el examen exhaustivo de la anota-

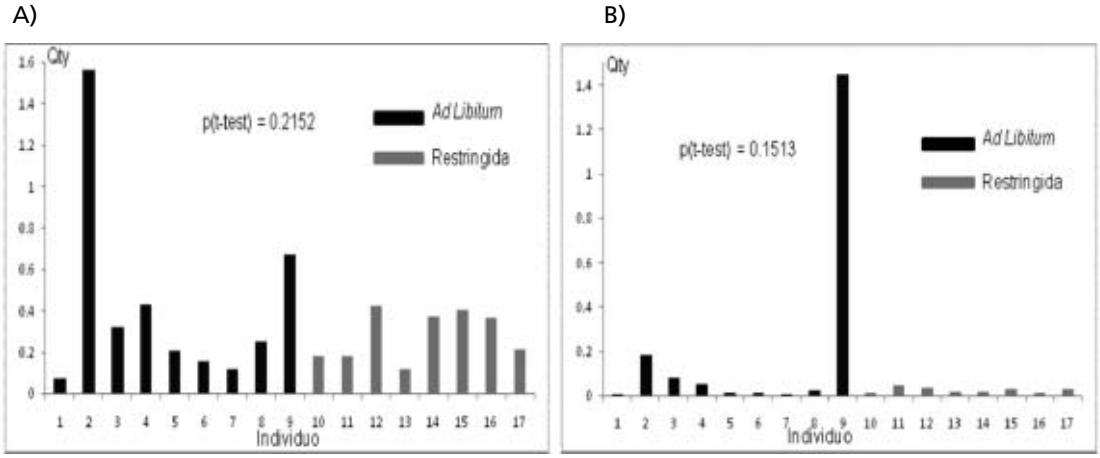

Figura 2. Representación gráfica de las medidas relativas de expresión de los genes *FABP3* y *SCD* medidas a través de qPCR en cerdos alimentados con dieta ad libitum y restringida.
A) Expresión del gen *FABP3*. B) Expresión del gen *SCD*.
Figure 2. Graphic representation of the relative expression measures of FABP3 *and* SCD *genes measured by qPCR in pigs fed with ad libitum and restrinted diets.*
A) FAPB3 *gene expression. B)* SCD *gene expression.*

ción de los sets de sondas y de las secuencias utilizadas para su diseño, así como la validación de algunos de los resultados mediante qPCR, lo que va a permitir determinar de manera más precisa el nivel de expresión, además de contrastar los resultados obtenidos con diferentes enfoques estadísticos.

Agradecimientos

Este trabajo ha sido llevado a cabo en el marco del proyecto GEN03-20658-C05 04, con la colaboración técnica de M. Nieto y de Comercial Pecuaria Segoviana (Sepúlveda).

Bibliografía

Benjamini Y, Hochberg Y, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc, Series B 57, 289-300.

Byrne KA, Wnag YH, Lehnert SA, Hasper GS, McWilliamd SM, Bruce HL, Reverter A, 2005. Gene expression profiling of muscle tissue in Brahman steers during nutritional restriction. J. Anim Sci, 83, 1-12.

Casellas J, Ibáñez-Escriche N, Martínez-Giner M, Varona L, 2008. GEAMM v.1.4: a versatile program for mixed model analysis of gene expression data. Anim Genet, 39(1), 89-90.

Hernández A, Karrow P, Mallard BA, 2003. Evaluation of immune responses of cattle as a means to identify high or low responders and use of a human microarray to differentiate gene expression. Genet Sel Evol, 35, S67-81.

Vandesompele A, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F, 2003. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology, 3(7), 0034.I-0034.II.

(Aceptado para publicación el 28 de abril de 2008)