Diseño experimental aplicado al diseño de hibridación en chips multi-arreglo

Julio A. Di Rienzo[1], Lucila Peluffo[2], David Blesa[4], Paula Fernández[2,3], Francisco García[4], Verónica Lía[2], Darío Príncipi[2], Federico Ehrenbolger[2], Ana Conesa[4], Joaquín Dopazo[4], Norma Paniego[2,3] y Ruth A. Heinz[2,3]

[1]Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba
[2]CONICET; [3]Instituto de Biotecnología, CICVyA-INTA Castelar, Buenos Aires
[4]Centro de Investigación Príncipe Felipe, Valencia, España

Resumen

Una iniciativa conjunta del Instituto de Biotecnología, CICVyA, INTA Castelar, Argentina y el Centro de Investigación Príncipe Felipe, Valencia, España, permitió el desarrollo de una matriz de alta densidad de oligonucleótidos (chip) para el girasol (*Helianthus annuus*),incluyendo aproximadamente 42K unigenes. El propósito de esta iniciativa fue el análisis de los perfiles de expresión génica en respuesta a factores bióticos y abióticos. Estos estudios se llevan a cabo en una red de laboratorios financiados a través del proyecto ANPCyT PAE 37100, que representan los sectores públicos, gubernamentales y privados en la Argentina.
La disponibilidad de este chip ha dado y dará origen una serie de proyectos de genómica funcional en girasol. El chip, basado en tecnología Agilent, cuenta con un diseño de cuatro micromatrices (arreglos) por 44 K sondas lo que permite hibridar simultáneamente cuatro muestras. La disponibilidad de cuatro arreglos por chip, representa una ventaja desde el punto de vista del diseño experimental, pero al mismo tiempo introduce una innovación que debe ser considerada tanto al momento del diseño de las hibridaciones como en el análisis estadístico posterior.

Keywords: Bioinformática, Microarrays, Genómica Funcional, Girasol, Diseño de Experimentos.

Introducción

En los primeros tiempos del desarrollo de las micromatrices de ADNc o de oligonucleótidos (microarrays), los aspectos metodológicos relativos al análisis de experimentos que usaban estas tecnologías se centraron principalmente en la problemática de la lectura de las imágenes, y en el pre-procesamiento de los datos leídos. La búsqueda de genes con expresión diferencial significativa se basó mayoritariamente en el ajuste de modelos sencillos y la mayor atención estuvo centrada en el control de falsos positivos y en la estimación de la varianza del error experimental. Debido a la simplicidad de los primeros experimentos, no se prestó mucha atención al diseño experimental así como al modelado de

datos. Una posible excepción a este modelo, lo constituyeron los desarrollos para las micromatrices de dos colores que, por la naturaleza de la plataforma, obligaban a tomar algunos cuidados de diseño (i.e. diseños de intercambio por fluoróforo o "dye-swap"). En los últimos, años la madurez de las técnicas de microarreglos y el desarrollo de nuevas plataformas estandarizadas basadas en oligonucleótidos, dio impulso al desarrollo de proyectos de expresión génica más ambiciosos. Estos proyectos implementan experimentos con una mayor riqueza en lo que se conoce como *estructura de tratamientos* y también en su *estructura de parcela*. Las réplicas biológicas suelen provenir de parcelar ordenadas según un diseño experimental (estructura de parcela) y los tratamientos pueden ser el resultado de múltiples factores de clasificación, incluyendo la presencia de datos longitudinales (estructura de tratamientos). Esta complejidad tiene una contraparte estadística necesaria tanto a la hora del diseño como del análisis.
Este tipo de experimento no es nuevo para los estadísticos. Los han estado diseñando y analizando desde mediados del siglo XX, en distintas aplicaciones del área agronómica e industrial. Sin embargo, no han sido tan comunes en los experimentos con microarreglos y esto se refleja en la falta de adecuadas implementaciones de software para analizarlos. La disponibilidad de un chip multi-arreglo, como el disponible para los experimentos de genómica del girasol, introduce una innovación que debe ser seriamente considerada a la hora de diseñar el plan de hibridaciones y compatibilizarlo con el diseño utilizado para la obtención del material biológico.
A continuación se presenta un ejemplo relacionado con esta problemática en un experimento conducido con un chip multi-arreglo para evaluar el efecto de la inoculación del hongo necrotrófico *Sclerotinia sclerotiorum* a distintos tiempos post infección sobre el perfil de expresión génico en dos líneas de girasol con comportamiento contrastante en cuanto a su susceptibilidad a la enfermedad.

Elementos del Trabajo y Metodología

El experimento cuyo diseño de hibridaciones se discute es un tri-factorial donde se evaluaron dos líneas de girasol: RHA801, tolerante (T) y HA89 susceptible (S), dos tratamientos de inoculación: esporas de hongos (I) y la suspensión o espray de agua (N) y dos momentos post infección: 2 y 4 días. Este arreglo tiene un total de 8 tratamientos y se quiere repetir 3 veces.

El chip utilizado es un multi-arreglo con 4 campos por chip. Desde la óptica estadística el chip es un *bloque con 4 unidades experimentales*. Debido a que el experimento tiene 8 tratamientos y los bloques son de tamaño 4 necesitaremos utilizar el concepto de *bloques incompletos* para diseñar la asignación de los tratamientos a los bloques en cada repetición. Es obvio que para tener una repetición del experimento se necesitan dos chips (8 unidades experimentales). Esto implica que algunos pares de tratamiento van a quedar en bloques diferentes y por lo tanto su comparación va quedar "confundida" con la diferencia entre bloques. Por ello, hay que controlar qué se confunde en cada repetición siguiendo un *plan de confundimientos*[1]. A esto se suma que el experimento tiene una estructura factorial de tratamientos y por lo tanto el plan de confundimientos debe prestar atención a ello.

En un experimento tri-factorial con los factores: Línea, Inoculo y Tiempo, es deseable poder evaluar los efectos fijos que se detallan en la Tabla 1.

Tabla 1: Lista de efectos a evaluar en el experimento tri-factorial donde los factores son Línea, Inóculo y Tiempo

Efecto
Línea
Inóculo
Tiempo
Línea*Inóculo
Línea*Tiempo
Inóculo*Tiempo
Línea*Inóculo*Tiempo

Si designamos a la línea tolerante como T y a la susceptible como S, a la exposición al hongo como I (inoculado) y no inoculado como N y a los tiempos de evaluación como 2 y 4, correspondientes a los días 2 y 4 días, tenemos los siguientes 8 tratamientos definidos por la combinación de niveles de cada factor:

TI2 TI4 TN2 TN4 SI2 SI4 SN2 SN4

Para comprender el fundamento del plan de confusión debe entenderse como se estiman los efectos de la Tabla 1.

La estimación de los efectos listados en la Tabla 1 se realiza a través de *contrastes*. Un contraste es una combinación lineal, en ese caso aplicada a las medias de tratamientos. A continuación se muestran los coeficientes del contraste necesario para estimar el efecto Línea. Éste compara las medias de los tratamientos que contienen a la línea tolerante (T) vs aquellos que contienen a la línea susceptible (S):

$$contrasteTvs.S = \sum_{i=1} c_i \hat{\mu}_i \;\; donde \;\; \mathbf{c}' = \begin{matrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{matrix}.$$

De manera similar si quisiéramos estimar el efecto Inóculo necesitaríamos un contraste que compare los tratamientos que recibieron el inóculo (I) vs los tratamientos que no lo recibieron (N):

$$contrasteTvs.S = \sum_{i=1} c_i \hat{\mu}_i \;\; donde \;\; \mathbf{c}' = \begin{matrix} 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{matrix}.$$

Si ubicáramos los tratamientos en los bloques según el signo de los coeficientes de un contraste, entonces, ese contraste y por lo tanto el efecto que estima, quedará confundido con la diferencia entre bloque y no podrá, en términos prácticos, estimarse.

La Tabla 2 muestra los contrastes necesarios para estimar cada uno de los términos enunciados en la Tabla 1. Es interesante observar que cada efecto tiene un único contraste asociado debido a que los factores sólo tienen dos niveles cada uno. Esto tiene implicancias en el diseño de las hibridaciones del chip de girasol que se discutirán más adelante.

Tabla 2: Lista de efectos a evaluar en el experimento tri-factorial

	Tratamientos							
Efectos	TI2	TI4	TN2	TN4	SI2	SI4	SN2	SN4
Línea	-1	-1	-1	-1	1	1	1	1
Inóculo	-1	-1	1	1	-1	-1	1	1
Tiempo	-1	1	-1	1	-1	1	-1	1
Línea*Inóculo	1	1	-1	-1	-1	-1	1	1
Línea*Tiempo	1	-1	1	-1	-1	1	-1	1
Inóculo*Tiempo	1	-1	-1	1	1	-1	-1	1
Línea*Inóculo*Tiempo	-1	1	1	-1	1	-1	-1	1

Como el experimento quiere repetirse 3 veces, podemos elegir en cada repetición (dos chips) un efecto diferente para confundir, de manera tal que en el experimento completo todos los efectos puedan ser estimados. Obviamente aquellos efectos confundidos serán estimados con menor precisión y es por ello que debe establecerse qué efectos se quieren sacrificar parcialmente. Si confundimos la interacción triple y las interacciones dobles Inóculo*Tiempo y Línea*Tiempo entonces el plan final de hibridaciones es el que se presenta en la Tabla 3.

Tabla 3: Plan final de hibridaciones

Repetición	Chip	Efectos confundidos	Asignación de tratamientos			
1	1	Línea*Inóculo*Tiempo	TN4	TI2	SN2	SI4
1	2		TN2	TI4	SN4	SI2
2	3	Inóculo*Tiempo	TN2	TI2	SN4	SI4
2	4		TN4	TI4	SN2	SI2
3	5	Línea*Tiempo	SI2	TN4	SN4	TI2
3	6		TI4	SN2	TN2	SI4

Cuando existe un diseño experimental asociado a la obtención de las muestras biológicas este debe diseñarse teniendo en cuenta el ulterior diseño de hibridación y en la medida de los posible tratar que los bloques, ya sean completos o no de ese diseño, contengan un número de unidades que sea múltiplo de 4 de tal forma en que permita, con cierta facilidad organizarse un diseño que contemple ambas etapas del proceso experimental total. Lo más importante es no perder la información sobre el diseño que da origen a las muestras biológicas (trazabilidad).

Una estructura de bloqueo puede aparecer *a posteriori* del diseño de hibridación debido a cuestiones de logística en el proceso mismo de la hibridación. En estos casos es posible que algunas réplicas se realicen en distintas rondas de hibridación considerando las capacidades del instrumental y la planificación del laboratorio. Esta información debe también anotarse y contemplarse a la hora del análisis ya que esto introduce otra "capa" en la estructura de

parcela. En el experimento que analizamos, cada repetición (dos chips) se hibridó en distintas rondas de hibridación.
Una vez que se tienen las lecturas normalizadas de expresión génica, la forma tradicional de analizar experimentos con microarreglos de alta densidad es la utilización de librarías de Bioconductor[2] como *limma*[3]. El enfoque se basa en el ajuste, gen a gen, de un modelo lineal de efectos fijos para asignar p-valores para las distintas hipótesis de interés. Eventualmente, *limma* permite especificar algunas estructuras de correlación entre observaciones de manera tal que puede contemplarse la presencia de efectos aleatorios como los bloques. Sin embargo las facilidades de *limma* son limitadas para manejar estructura de parcelas es compleja. La estrategia de análisis que se recomienda en este trabajo es ajustar un modelo lineal de efectos mixtos. Estos modelos se llaman mixtos porque tienen una parte conformada por efectos considerados fijos (reproducibles) y otra parte dada por efectos aleatorios. En estos modelos la estructura de parcela se introduce a través de efectos aleatorios, mientras que la estructura de tratamientos se modela en la parte fija. Además de tener una gran plasticidad para incorporar los detalles del diseño experimental, permiten manejar de manera transparente la presencia de datos faltantes. El modelo ajustado fue el siguiente:

$$Y^{g}_{ijklmn} = \mu + \lambda_i + \delta_j + \gamma_k + \lambda\delta_{ij} + \lambda\gamma_{ik} + \delta\gamma_{jk} + \lambda\delta\gamma_{ijk} + s_l + c_{m(l)} + \varepsilon_{ijklmn}$$

Y^{g}_{ijklmn} = representa la expresión génica del g-esimo gen en la i-ésima línea, j-ésimo inóculo, k-ésimo tiempo, l-ésima horneada de hibridación, m-ésimo bloque dentro de horneada, n-ésima observación dentro del bloque.

μ = media general común a todas las observaciones

λ_i = efecto de la i-ésima línea, i=1,2.

δ_j = efecto del j-ésimo inóculo, j=1,2.

γ_k = efecto del k-ésimo tiempo, k=1,2.

$\lambda\delta_{ij}$ = interacción de la i-ésima línea - j-ésimo inóculo.

$\lambda\gamma_{ik}$ = interacción de la i-ésima línea - k-ésimo tiempo.

$\delta\gamma_{jk}$ = interacción del j-ésimo inóculo - k-ésimo tiempo.

$\lambda\delta\gamma_{ijk}$ = interacción i-ésima línea - j-ésimo inóculo - k-ésimo tiempo.

s_l = efecto aleatorio de la l-ésima horneada, l=1,...,3.

$c_{m(l)}$ = efecto aleatorio del m-ésimo bloque dentro de la l-ésima horneada, m(l)=1,2.

ε_{ijklmn} = error experimental.

Una vez ajustado el modelo se probaron las hipótesis de efectos nulos para los distintos efectos fijos. La asignación de genes con efectos significativos se determinó de manera secuencial. Primero se filtraron los genes que resultaron significativos para la interacción tripe, luego se filtraron por las interacciones dobles y finalmente se examinaron, en los genes remanentes los que presentaban efectos significativos para los efectos principales.

Como el interés estaba centrado en el efecto de la inoculación, sólo se consideraron las interacciones: Línea*Inóculo*Tiempo,Inóculo*Tiempo, Línea*Inóculo y el efecto principal Inóculo. Una vez identificados los genes que resultaron significativos ($p<0.001$ – no corregido para controlar FDR) se realizó un análisis de agrupamiento de los perfiles medios para los genes asociados a cada efecto significativo considerado. El algoritmo utilizado para ello fue k-means. Debido a que interés estaba en agrupar perfiles de expresión según su forma, previo al análisis de agrupamiento, el perfil de cada gen fue centrado por la media general de expresión del gen y escalado por la desviación estándar residual.

La lectura de los microarreglos se realizó mediante la función *read.maimages*(), la corrección por ruido de fondo mediante la función *backgroundCorrect()* utilizando el método *rma* y normalización de los arreglos se realizó utilizan la función *normalizeBetweenArrays()* especificando el método *quantile*, todas funciones de la librería *limma*. El modelado estadístico se realizó mediante la implementación de un algoritmo basado en la función *lme* de la librería *nlme*[4] de R[5]. Todo el trabajo con R se realizó bajo la interfaz de fgStatistics[6]. El cálculo de los valores medios de tratamiento mediante la librería lsmeans[7].

Resultados

Mediante el procedimiento descripto anteriormente se pudieron identificar 36 genes que muestran patrones diferenciales de expresión para las combinaciones Línea*Inóculo*Tiempo. Sus perfiles de expresión se muestran en la Figura 1. Para el efecto Línea*Inóculo se identificaron 19 genes cuyos perfiles se muestran en la Figura 2, y 7 genes mostraron expresión diferencial para la interacción Tiempo*Inóculo (Figura 3). Finalmente 11 genes mostraron cambios en su expresión debida al inóculo independientemente de los otros factores considerados en el estudio. De estos genes, tres mostraron niveles de expresión menores como resultado de la infección mientras que el resto se sobre expresó. Todos estos genes están actualmente siendo objeto de un análisis avanzado a partir de las evidencias por función y proceso biológico, como paso previo a una etapa de validación experimental

Discusión

El propósito de este trabajo es dirigir la atención sobre algunos aspectos metodológicos del diseño de los planes de hibridación cuando se utilizan chips multi-arreglo como el recientemente incorporado en los proyectos de genómica funcional del girasol. Es importante que los investigadores involucrados en estos proyectos perciban que el diseño de hibridaciones no es trivial, que debe considerarse en el marco más amplio de un diseño conjunto tanto para la obtención del material biológico como para el plan de hibridaciones propiamente dicho y que la trazabilidad de los datos es crucial para la formulación de un modelo estadístico apropiado.
También se quiere destacar la necesidad de incorporar a los modelos mixtos en la práctica rutinaria del análisis de la expresión génica. Estos modelos no sólo ofrecen la plasticidad necesaria para contemplar estructuras de tratamiento complejas sino también y especialmente la posibilidad de modelar la estructura de parcela con todo el nivel de detalle

que requiera. Más aún, los modelos mixtos pueden ser utilizados no sólo para encontrar genes candidatos sino también como un procesamiento previo a la aplicación de otras técnicas de análisis avanzadas. Aplicar cualquier procesamiento a los datos de expresión génica sin eliminar previamente la estructura de parcela podría con facilidad conducir a conclusiones erróneas. Los modelos mixtos permiten, de manera simple, obtener expresiones génicas donde toda la variación atribuible a la estructura de parcelas ha sido eliminada.

Agradecimientos

Este trabajo es soportado parcialmente por los proyectos: PAE 37100-PME 207-024, PAE 37100-PICT 2007-240607, PICT 02226 y 0960, AEBIO 245732, Fondo AECID D/016099/08, PIP CONICET 11220090100576, PICT 32905 e INTA AEBIO 245001 y 245711.

Referencias

[1] Hinkelmann K, Kempthorn O. (2005). Design and Analysis of Experiment. Vol 2. Advanced Experimental Design. John Wiley & Sons.
[2] http://www.bioconductor.org/
[3] Smyth, G. K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005.
[4] Pinheiro, J.C., and Bates, D.M. (2000) "Mixed-Effects Models in S and S-PLUS", Springer.
[5] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
[6] Di Rienzo J.A, (2009). fgStatistics, Statistical software for the analysis of experiments of functional genomics. http://sites.google.com/site/fgStatistics/
[7] Di Rienzo, J.A., Romero, MC. (2010). lsmeans. An R library for the calculation of least square means form lm, gls and lme models. R-Forge repository. install.packages("lsmeans",repos="http://R-Forge.R-project.org")

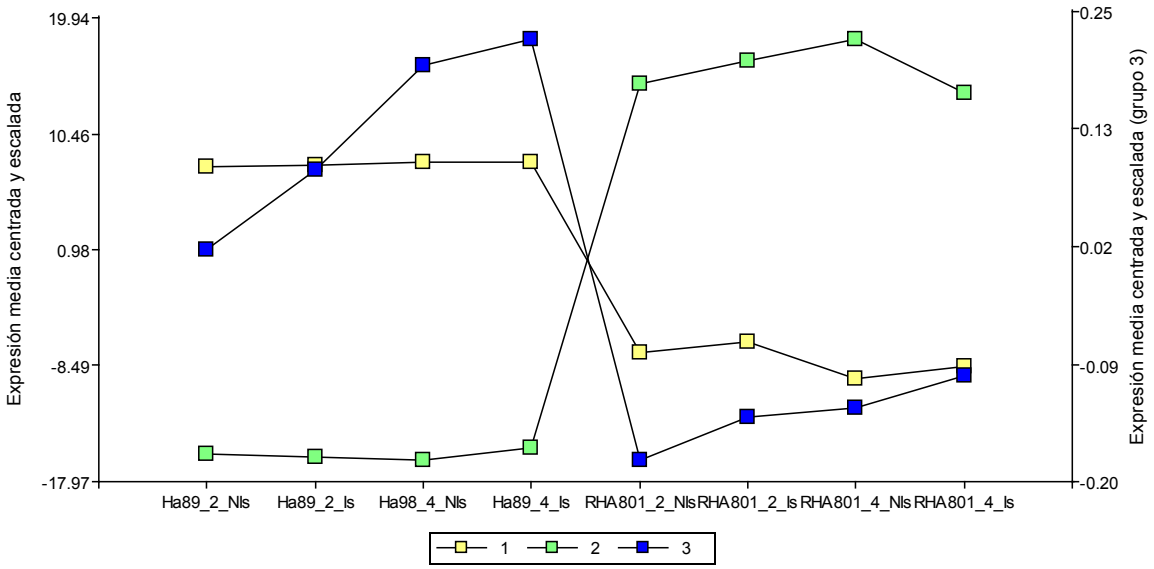

Figura 1: Perfiles de expresión génica de 3 tres grupos de genes que mostraron interacción Línea*Inóculo*Tiempo significativa.

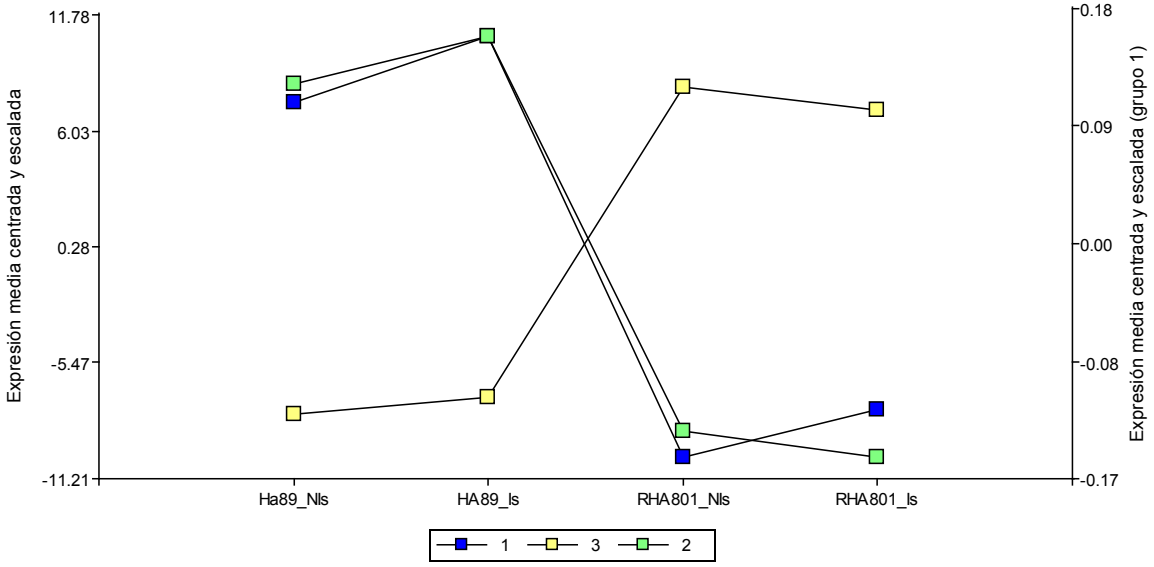

Figura 2: Perfiles de expresión génica de 3 tres grupos de genes que mostraron interacción Línea*Inóculo significativa.

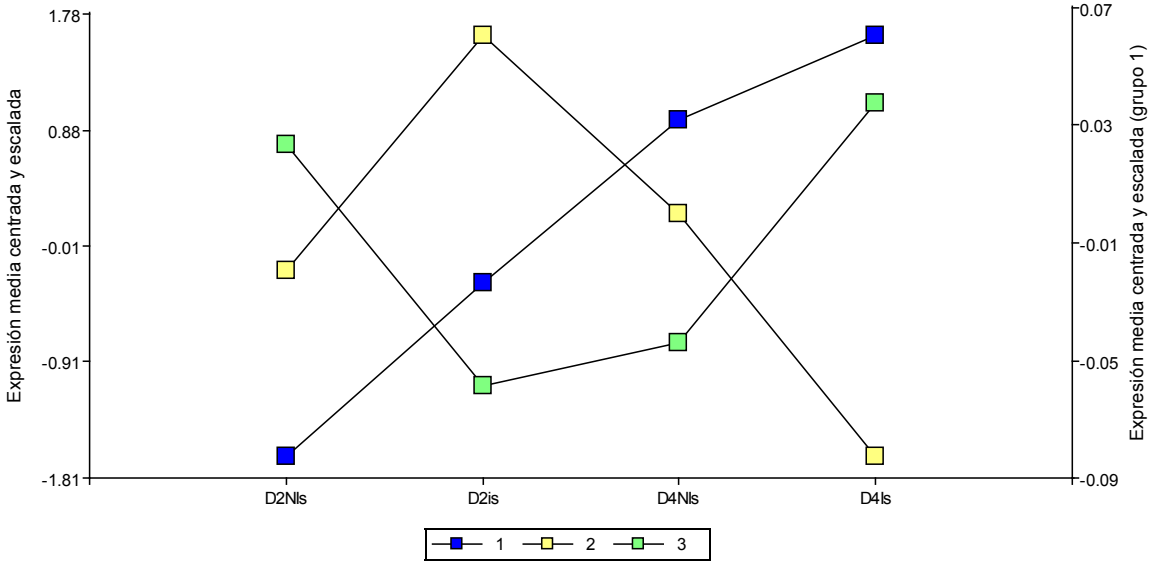

Figura 3: Perfiles de expresión génica de 3 tres grupos de genes que mostraron interacción Tiempo*Inóculo significativa.