Papa, Lucia[1]
Quaglino, Marta B.[1]
Dianda, Daniela F.[1]
Orellano, Elena G.[2]
Daurelio, Lucas D.[2]
[1]*Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario.*
[2]*IBR-CONICET, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario.*

IDENTIFICACIÓN DE PATRONES DE EXPRESIÓN GÉNICA EN PLANTAS RUTÁCEAS BAJO ESTRÉS BIÓTICO MEDIANTE ANÁLISIS DE CONGLOMERADOS

Resumen

El objetivo de este trabajo es analizar la expresión génica obtenida de diferentes interacciones entre plantas y patógenos utilizando la tecnología de los microarreglos. En un primer paso se normalizaron los datos, con el fin de identificar y remover fuentes de variación sistemáticas. Una vez normalizados los datos, se realizaron comparaciones de interés entre las interacciones planta-patógeno analizadas, con el objetivo de identificar aquellos genes que se expresaron de manera diferencial. Para esto fue utilizado el ajuste de modelos lineales para la creación de contrastes evaluados a través de la prueba “t” y el método “fold change”, específico para este tipo de datos. Los resultados de las diferentes comparaciones se unificaron en una gran base de datos. De dicha base fueron detectados aquellos genes diferencialmente expresados en al menos un tratamiento. En un paso posterior, con el grupo de genes diferencialmente expresados se realizó un análisis de conglomerados a fin de identificar patrones de comportamiento o co-expresión de los genes. Tanto el análisis como la construcción de los algoritmos para el armado y manipulación de las bases de datos se llevó a cabo utilizando el software estadístico R.

Palabras Claves: conglomerados, expresión génica, fold change, microarreglos.

Abstract

The aim of this paper is to analyze gene expression obtained from different plant-pathogen interactions, using microarrays technology. The first step was a normalization to identify and remove various sources of systematic variation. Once standardized data, comparisons of interest between plant-pathogen interactions were tested, in order to identify those genes that are differentially expressed. This analysis was made by adjusting linear models for creating contrasts evaluated through the "t" test and the specific method "fold change". The results of different comparisons were unified in a big database. Those genes differentially expressed in at least one comparisons of treatment were detected in the unified basis. Next, a cluster analysis was performed using the differentially expressed genes in order to identify patterns of behavior or co-expression of genes. Both, the analysis and the construction of algorithms for assembly and manipulation of databases, were performed using the statistical software R.

Key words: cluster, gene expression, fold change, microarrays.

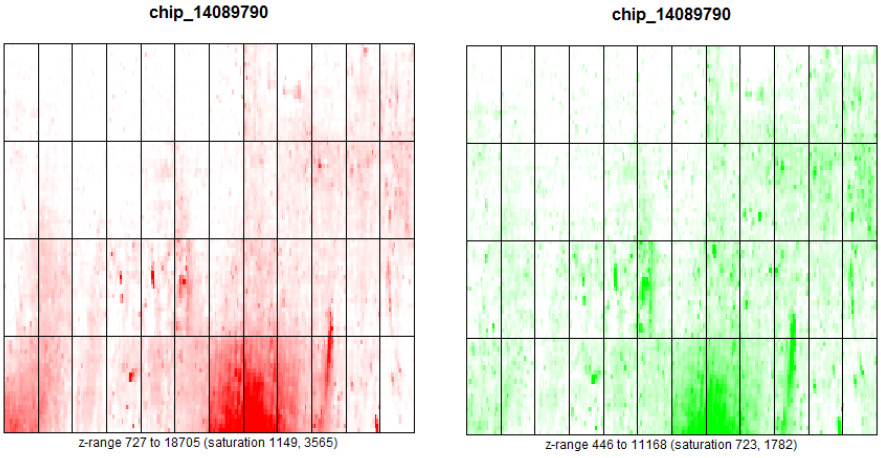
1. Introducción

El estudio de la expresión génica o medición de los niveles de ácido ribonucleico mensajero[1] (ARNm) es utilizado para responder preguntas biológicas acerca de una gran variedad de organismos en una gran diversidad de condiciones. Para esto ha sido necesario encontrar técnicas que faciliten el estudio de grandes cantidades de datos de este tipo y que obtengan como resultado información útil. En este sentido, una de las áreas más importantes de la Bioinformática es la denominada genómica funcional, la cual incluye el análisis de expresión génica.

Una de las tecnologías más utilizadas hoy en día para el análisis de expresión génica masiva es la de los microarreglos. El uso de microarreglos ha emergido como una potente técnica para la medición de datos de expresión génica y la comparación de la abundancia relativa de ARNm generado en distintas pruebas biológicas. Los resultados obtenidos mediante esta técnica implican un gran número de datos que deben ser analizados utilizando diferentes métodos estadísticos y bioinformáticos.

El microarreglo, también llamado DNA chip o Biochip, es un soporte sólido construido normalmente en cristal o en membrana de nylon sobre el cual se fijan fragmentos de ácido desoxirribonucleico[1] (ADN) conocidos en ubicaciones perfectamente especificadas, correspondientes a los genes cuyos niveles de expresión serán cuantificados. Después de realizar los experimentos bajo las condiciones de interés sobre las células a analizar, se extrae el ARNm generado por los genes de dichas células. Paso seguido estas moléculas de ARNm son utilizadas como molde para una retrotranscripción (proceso contrario a la transcripción donde se genera una molécula de ADN a partir de una de ARN de molde), generando un tipo especial de ADN conocido como ADNc o ADN copia[1]. En el proceso de retrotranscripción las moléculas de ADNc se marcan con moléculas fluorescentes que permitirán detectar una señal lumínica, colocándose posteriormente en los microarreglos para comenzar el proceso de hibridación. La hibridación del ADN es un proceso comúnmente utilizado para detectar un gen particular o un segmento de un ácido nucleico. Después del proceso de hibridación aquellas moléculas de ADNc presentes en la solución marcada se unirán a las respectivas de ADN complementario fijadas en la matriz del microarreglo mientras que se eliminan todas las cadenas que no se han podido unir a través de lavados (sólo las moléculas que hibridan permanecerán en el biochip), y se procede al revelado mediante un escáner láser.

Figura 1. Sección de la imagen obtenida de un microarreglo mediante el escaneado del mismo.



Fuente: Daurelio y col.

El resultado final es una matriz en la que cada una de las celdas (denominadas “spots”) está marcada con una determinada intensidad de color rojo o verde las cuales serán medidas utilizando diferentes canales, como se puede observar en la Figura 1. Dicha intensidad del color simboliza el grado de expresión génica de dicho gen frente a una determinada condición experimental, dado que en los microarreglos de dos colores se hibridan en simultáneo los

[1] Ver definición de ADN, ARNm y ADNc en anexo

ADNc marcas de dos muestras a comparar (por ejemplo un tratado y un control), una de ellas es marcada con una molécula de color rojo y la otra de color verde. La cantidad de cada color será proporcional a la cantidad de ARNm del respectivo gen en cada una de las muestras, donde en caso de cantidades similares se observará un color amarillo y en caso de ausencia de expresión no se observará señal.

Debido al gran volumen y la variación intrínseca de los datos obtenidos en experimentos de microarreglos de genes, los métodos estadísticos se han convertido en una importante herramienta de análisis con el fin de extraer sistemáticamente información biológica y evaluar posibles asociaciones.

En primera instancia se realizan sobre las imágenes de los microarreglos exploraciones visuales o descriptivas con la finalidad de detectar errores groseros que indiquen que no pueden ser utilizadas. Luego se realiza el denominado pre-procesado de los datos que incluye la corrección del fondo y la remoción de fuentes de variación sistemáticas distintas de la diferencia de expresiones de las intensidades de fluorescencia medidas o "normalización".

Finalmente se procede al análisis de los datos pre-procesados. En estudios de microarreglos de genes, existe interés en encontrar genes diferencialmente expresados entre dos o más condiciones, descubrir patrones de expresión característicos, predecir la respuesta a un tratamiento, identificar genes co-regulados o expresándose en la misma ruta metabólica, etc. Para cada uno de estos intereses, existen múltiples métodos de análisis estadístico que es posible aplicar. En este trabajo se aplican técnicas de aglomeración con el fin de agrupar genes en función de sus patrones de expresión, detectando grupos de genes que co-expresen según los distintos tratamientos.

2. Materiales y métodos

Daurelio y col. (2013) iniciaron un proyecto de investigación para caracterizar a nivel transcripcional vías metabólicas, familias génicas y sistemas reguladores de la expresión de plantas Rutáceas involucrados en las respuestas a patógenos bacterianos. El objetivo de este proyecto es contribuir en la interpretación de los mecanismos moleculares durante dichas interacciones planta-patógeno y encontrar soluciones alternativas a la cancrosis de los cítricos producida por la bacteria *Xanthomonas citri* subsp. *Citri (Xcc).* Para ello fueron analizados los transcriptomas de diferentes plantas del género de las Rutáceas frente a distintas especies de bacterias fitopatógenas del género *Xanthomonas* spp a través de la técnica de microarreglos. Estas interacciones abarcan diferentes respuestas de una planta frente a un patógeno. Para cada uno de los tratamientos se amplificaron tres muestras de ARN total obtenidas en forma independiente y tres muestras pertenecientes a los controles o "referencia". El diseño experimental implica la hibridación de cada muestra independiente frente a una referencia común. Esta referencia se obtuvo a partir de una mezcla equimolar de ARN proveniente de todas las muestras a analizar. Las muestras se marcaron con el fluoróforo Cyanina5 (Cy5 que produce color rojo), mientras que la referencia se marcó con el fluoróforo Cyanina3 (Cy3, que produce color verde).

2.1. Medición de la expresión génica utilizando microarreglos

El conjunto de todos los ARNm transcritos en un sistema biológico se conoce como *transcriptoma*. Aunque la mayoría de las proteínas sufre modificación después de la traducción y antes de convertirse en funcionales, la mayoría de los cambios en el estado de una célula están relacionados con los cambios en los niveles de ARNm de algunos genes, por lo que la medición sistemática del transcriptoma es una aproximación muy importante para analizar el estado celular.

Si bien existen varias tecnologías que permiten medir los cambios del nivel de expresión de los genes en gran parte o todo el transcriptoma, las más utilizadas son los microarreglos, de los cuales los más frecuentes son los de ADNc y de oligonucleótidos. Aunque ambos exploran la hibridación del ADN, difieren en la forma en que las secuencias de ADN se colocan sobre la matriz y en la longitud de las mismas, como también en el análisis posterior de los datos. Como ventaja más importante, el uso de microarreglos ofrece la valiosa oportunidad de poder estudiar

a la vez el comportamiento de una gran cantidad de genes bajo diferentes condiciones experimentales. Como punto negativo de esta tecnología se debe considerar el elevado costo de producir microarreglos para los experimentos por lo que habitualmente un estudio cuenta con un bajo número de ellos, lo que lleva a tener sólo dos o tres réplicas biológicas en la mayoría de los casos.

En el enfoque de microarreglos de ADN de dos colores, después de la hibridación, un escáner láser de colorante mide la fluorescencia de cada color en una rejilla fina de píxeles. Valores altos de fluorescencia indican cantidades más altas de ADNc hibridado, que a su vez indica una mayor expresión génica en la muestra. El algoritmo de análisis de imágenes produce resúmenes de fluorescencia para cada spot, y para las áreas circundantes sin manchas (fondo de la imagen). Para cada ubicación (spot) en el microarreglo, una salida típica consiste en al menos cuatro cantidades: una para la intensidad de cada color, tanto para el punto como para el fondo y, a veces, estos son acompañados por medidas que refieren a la calidad de la mancha para detectar problemas técnicos, o a la variabilidad en la intensidad de los píxeles.

Convencionalmente los dos colores utilizados son el rojo y el verde, y las intensidades de cada color son simbolizadas con las letras R y G respectivamente. El uso de dos intensidades permite la medición de la expresión génica relativa a través de dos fuentes de ADNc, controlando la cantidad de manchas de ADN, que puede ser variable, así como alguna otra variación experimental. Esto ha llevado al énfasis en el cálculo y uso de las razones de intensidad en cada punto comúnmente llamada tasa de intensidades: $\frac{R}{G}$.

En experimentos de microarreglos, los primeros valores cuantificados están contenidos en las imágenes producidas por el escáner. Las intensidades de los píxeles, guardadas en estos archivos, son transformadas a datos numéricos y estos pueden ser considerados como los datos crudos.

2.2. Normalización de los datos.

En un experimento ideal, las normalizaciones no deberían ser necesarias. Sin embargo, los sesgos técnicos son inevitables y deben ser corregidos. Para determinar si es necesaria la normalización, se puede trazar un diagrama de dispersión de intensidades R versus G. Bajo el supuesto de que un gen se expresa de la misma forma bajo cualquier condición, es de esperar que un gráfico de intensidades R vs. G presente una nube de puntos a la cual se podría ajustar una recta con pendiente cercana a uno. De dicho gráfico sobresaldrían o serían "outliers" aquellos genes que presenten expresión diferencial en alguna de las condiciones. Con base en esta justificación, si los datos observados no se distribuyen de la forma anteriormente descripta, todas aquellas desviaciones con respecto al ideal y que no sean puramente debidas a la expresión diferencial, podrían deberse a errores sistemáticos vinculados con varios factores, como pueden ser: diferencias en la eficiencia de la incorporación de los tintes, diferencias en la cantidad de ARNm en las muestras, diferencias en los parámetros de escaneado, efectos espaciales y otros, lo cual justificaría la aplicación de un método de normalización a los datos.

Existe un método gráfico mejorado, conocido como gráfico MA (MA-plot), en el cual se grafica en el eje vertical al logaritmo en base 2 de la razón de intensidades$M = log_2\left(\frac{R}{G}\right) = log_2(R) - log_2(G)$ y en el eje horizontal al valor medio de los logaritmos en base 2 de las intensidades, $A = \frac{1}{2}[log_2(RG)] = \frac{1}{2}[log_2(R) + log_2(G)]$. Un gráfico MA ideal será aquel en el cual la nube de puntos se distribuye en forma aleatoria alrededor del 0 en el eje vertical, quedando como valores atípicos sólo aquellos que correspondan a genes diferencialmente expresados.

El proceso de normalización de datos provenientes de microarreglos de dos canales puede separarse en dos componentes: ubicación o localización (l) y escala (s). Sean $\{Y_i\}$ y $\{X_i\}$ los valores observados de variables que miden la expresión génica del gen i en un experimento de microarreglos, donde i= 1, 2,..., N y N es el número de genes en el microarreglo. Asumiendo que las fluctuaciones entre ambos grupos de datos son modeladas como una relación lineal que involucra operaciones de escalamiento y desplazamiento, se tiene que:

$$Y_i = \alpha X_i + \beta + \eta$$

Donde α y β son los parámetros de escala y de desplazamiento respectivamente y η representa los errores entre ambas muestras.

El proceso de normalización consiste en una transformación del conjunto de datos $\{X_i\}$ en un nuevo conjunto de datos $\{X'_i\}$, más confiables para la realización del análisis.

$$X'_i = \left(\frac{Y_i - b}{a}\right)$$

Donde a y b son los estimadores de los parámetros α y β respectivamente.

A partir de aquí, los métodos de normalización se distinguirán según el método utilizado para la estimación de los parámetros, los cuales pueden ser basados en regresión lineal, de media de razón unitaria o basada en regresión local

En este trabajo se aplicó el método de normalización basado en regresión local, el cual permite, en el contexto de experimentos con microarreglos, capturar la dependencia de la tasa de registro de intensidad (log-ratio) $M = log_2 \frac{R}{G}$ respecto de la intensidad promedio $A = log_2 \sqrt{RG}$, garantizando que los valores normales calculados no son impulsados por un pequeño número de expresiones diferenciales de genes con tasas de registro extremas.

2.3. Técnicas de análisis de datos de expresión génica.

Un experimento de microarreglos simple se puede llevar a cabo para detectar las diferencias en la expresión de los genes bajo dos condiciones distintas. Cada condición puede ser representada por una o más muestras de ARN. Usando microarreglos de ADNc de dos colores, las muestras se pueden comparar directamente en el mismo microarreglo o indirectamente mediante la hibridación de cada muestra con una muestra de referencia común.

En ambos casos la hipótesis nula que se postulará es que no hay diferencia en la expresión entre las muestras. Cuando las condiciones se comparan directamente, esto significa que la verdadera relación entre la expresión de cada gen en las dos muestras debe ser uno. Cuando las muestras se comparan indirectamente, las relaciones entre las expresiones de la muestra y de la referencia deben ser similares en las dos condiciones.

Suele ser más conveniente utilizar logaritmos en base 2 de los coeficientes de expresión que los propios coeficientes porque los efectos sobre la intensidad de las señales de microarreglos tienden a ser multiplicativos.

A continuación se presentan dos métodos diferentes para identificar los genes que presentan expresión diferencial. Uno de ellos es un método estadístico inferencial, basado en la realización de contrastes lineales; el otro, denominado “fold change”, , es una regla práctica muy utilizada en el ámbito de la investigación de expresiones génicas.

Fold change

El método más simple para la identificación de genes diferencialmente expresados es evaluar la relación entre el registro de dos condiciones (o la media de relaciones cuando hay repeticiones) y considerar que todos los genes que difieren en más de un punto de corte arbitrario, se expresaron diferencialmente. Por ejemplo, si el valor de corte elegido es una diferencia de dos, se considerarán como genes diferencialmente expresados aquellos para los cuales la expresión bajo una condición es más de dos veces mayor o menor que bajo la otra condición. Esta prueba, llamada 'fold change' es equivalente a la razón $\frac{R}{G}$ observada y no es una prueba estadística, no hay un valor asociado que pueda indicar el nivel de confianza en la designación de los genes como diferencialmente expresados. Como se enuncia anteriormente, para un mejor análisis e interpretación de la razón $\frac{R}{G}$ es común la utilización del $log_2(fold\ change) = M$.

Contrastes

Con el fin de detectar estadísticamente aquellos genes diferencialmente expresados,

puede recurrirse a la realización de pruebas de hipótesis, siendo la prueba "t" para la evaluación de contrastes lineales la habitualmente utilizada en este tipo de investigaciones.

Sea un conjunto de n microarreglos que producen un vector de respuesta $\boldsymbol{y_g^T} = (y_{g1}, \dots, y_{gn})$ para el gen g, donde tales respuestas suelen ser el logaritmo de la razón de intensidades

$$y_{gi} = M_{gi} = log_2 \frac{R_{gi}}{G_{gi}}$$

Asumiendo que las respuestas fueron correctamente normalizadas, se supone que $E(\boldsymbol{y_g}) = \boldsymbol{X\alpha_g}$ donde $\boldsymbol{X}$ es la matriz de diseño y $\boldsymbol{\alpha_g}$ el vector de coeficientes correspondiente a los tratamientos analizados. Se asume que $Var(\boldsymbol{y_g}) = \boldsymbol{W_g}\sigma_g^2$ siendo $\boldsymbol{W_g}$ una matriz definida positiva de pesos conocidos y σ_g^2 la variancia para el gen g. A partir de aquí, pueden plantearse contrastes de interés biológico entre los tratamientos, los cuales pueden definirse como $\boldsymbol{\beta_g} = \boldsymbol{C^T\alpha_g}$, siendo $\boldsymbol{C}$ la matriz de contrastes a probar. De modo que es de interés ahora probar si $\boldsymbol{\beta_g}$ es igual al vector nulo. Ajustando el modelo lineal con todas las observaciones de cada gen, se obtienen los coeficientes estimados $\boldsymbol{\hat{\alpha}_g}$, el estimador de la variancia del gen s_g^2 y la matriz de covariancias estimadas $\boldsymbol{Var}(\boldsymbol{\hat{\alpha}_g}) = \boldsymbol{V_g}s_g^2$ donde $\boldsymbol{V_g}$ es una matriz definida positiva, independiente de s_g^2. Los contrastes estimados serán $\boldsymbol{\hat{\beta}_g} = \boldsymbol{C^T\hat{\alpha}_g}$ con matriz de covariancias estimadas igual a $\boldsymbol{Var}(\boldsymbol{\hat{\beta}_g}) = \boldsymbol{C^TV_g}s_g^2$.

Si bien las respuestas $\boldsymbol{y_g}$ no se asumen necesariamente normales, ni la estimación debe ser necesariamente por Mínimos Cuadrados, los contrastes estimados se suponen aproximadamente normales con media $\boldsymbol{\beta_g}$ y matriz de covariancias $\boldsymbol{C^TV_gC}\sigma_g^2$. Las variancias del error s_g^2 se asume que siguen una distribución aproximada chi-cuadrado. Como se ha anunciado anteriormente, en los estudios de microarreglos los tamaños de muestra suelen ser demasiado pequeños, lo cual es una desventaja al momento de utilizar este método. Varios autores han demostrado que es posible mejorar el uso de las pruebas t para evaluar expresión diferencial en experimentos de microarreglos mediante el uso de variancias estimadas a partir de todos los datos de la muestra. Dado que para los miles de genes se ajustará un mismo modelo lineal, puede sacarse provecho a esto, identificando cómo los coeficientes β_{gj} y las variancias σ_g^2 varían a través de los distintos genes. Como resultado de este procedimiento se obtiene la estadística t moderada $\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{v_{gj}}}$ que representa el clásico método Bayesiano mediante el cual la variancia muestral es sustituida por la variancia a posteriori, calculada con todos los datos y sigue una distribución aproximada t-Student con $d_0 + d_g$ grados de libertad. Donde $d_0 + d_g$son los grados de libertad del error de un modelo lineal para el gen g.

Las hipótesis a probar serán:

H_0: *El gen no está diferencialmente expresado.*
H_1: *El gen está diferencialmente expresado.*

Luego, dado un nivel de significación (α, fijado por el investigador) y los grados de libertad $(d_0 + d_g)$, la regla de decisión será:

Rechazar H_0 si $|\tilde{t}| \geq t_{(d_0+d_g;\alpha/2)}$, donde $t_{(d_0+d_g;\alpha/2)}$ es el valor de la variable t de Student con $d_0 + d_g$ grados de libertad, que acumula una probabilidad de $(1 - \frac{\alpha}{2})$.

Corrección de los p-valores

Al llevar a cabo una prueba de hipótesis, se está frente a la posibilidad de cometer dos tipos de errores:

- Error tipo I: Rechazar la hipótesis nula cuando esta es cierta. (falso positivo)
- Error tipo II: No rechazar la hipótesis nula cuando es cierta la hipótesis alternativa. (falso negativo)

Al probar múltiples hipótesis simultáneamente, como en el caso de los experimentos de

microarreglos, la situación se vuelve más compleja, ya que cada gen tiene posibles errores tipo I y II, de modo que la tasa de error global resulta extremadamente elevada.

Por este motivo se han propuesto diversos métodos que proporcionan correcciones para las probabilidades asociadas de cada test, de modo tal que se controle la tasa global de falsos descubrimientos (FDR por su sigla en inglés). Uno de tales métodos, propuesto por Siemes (1986) y que fue el utilizado en este trabajo, está basado en el valor esperado de la FDR cuando se evalúan simultáneamente m hipótesis. Considerando que se han probado m hipótesis $H_1, H_2, \dots, H_M$ basadas en sus correspondientes p valores $P_1, P_2, \dots, P_m$, sean $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ los p valores ordenados de menor a mayor, y sea $H_{(i)}$ la hipótesis nula correspondiente al p valor $P_{(i)}$. Sea k el máximo valor de i tal que $P_{(i)} \leq \frac{i}{m} q^*$, donde q^* representa la tasa de descubrimientos falsos deseada, entonces, el procedimiento propone rechazar todas las $H_{(i)}$ para $i = 1,2, \dots k.$.

Hasta aquí se han presentado los dos métodos usuales en experimentos de microarreglos para identificar los genes diferencialmente expresados. Dado que ambos métodos tienen el mismo objetivo pero no necesariamente conducen a los mismos resultados, es posible construir una representación gráfica de los resultados de ambos métodos para cada gen, la cual se conoce como gráfico de volcán.

Gráfico de volcán

El gráfico de volcán es una representación eficaz y fácil de interpretar que resume la información que proveen tanto el fold change como las pruebas t. Es un gráfico de dispersión del valor negativo del logaritmo en base 10 de los p-valores de la prueba t específica para cada gen versus el logaritmo en base 2 del fold change, es decir, M. Genes con expresión diferencial estadísticamente significativa según la prueba específica t se encontrarán por encima de una línea horizontal arbitraria que determine cuáles son los valores de p que se consideran significativos. Los genes con grandes valores de fold change se encontrarán fuera de un par de líneas verticales que indicarán los valores de fold change que se consideren significativos.

Análisis de conglomerados (clúster)

El verdadero poder de análisis de microarreglos no solo se asocia al análisis de los experimentos individuales, sino al análisis de muchas hibridaciones a la vez para identificar patrones comunes de expresión génica. Sobre la base de la comprensión total de los procesos celulares, los genes que se encuentran en una vía particular, o que responden a un reto ambiental común, deberían ser co-regulados y, en consecuencia, mostrar patrones de expresión similares. En este sentido, existe un gran grupo de métodos estadísticos que se pueden utilizar con el objetivo de identificar los genes que muestran patrones similares de expresión, entre ellos, los métodos de agrupamiento en conglomerados o *clustering*.

En cualquier algoritmo de agrupamiento, el cálculo de una distancia entre dos objetos es fundamental para luego armar los grupos. La búsqueda de grupos de genes “similares” en el análisis de datos de microarreglos no es la excepción. Dicha búsqueda y posterior agrupación se basará en definir qué genes presentan expresiones “cercanas” o “similares”. En este trabajo, se utilizó la distancia euclídea, definida como: $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ donde x_i e y_i son los valores de expresión medidos, respectivamente, para los genes X e Y en el i-ésimo experimento, y la suma se extiende sobre los n experimentos analizados.

Es posible aplicar diversas técnicas de agrupamiento para la identificación de patrones en los datos de expresión génica. La mayoría de las técnicas de agrupamiento son jerárquicas; la clasificación resultante tiene un número creciente de clases anidadas y el resultado se asemeja a una clasificación filogenética[2]. Las técnicas de agrupamiento se pueden clasificar como de división o de aglomeración. Un método de división comienza con todos los elementos en un gran grupo que se separa gradualmente hacia abajo en grupos cada vez más pequeños. Las técnicas de aglomeración comienzan (habitualmente) con racimos de un solo miembro cada

2 Término que refiere a todo aquello propio o vinculado a la Filogenia. La filogenia, cuya palabra tiene un origen griego que implica nacimiento, origen o procedencia, es la determinación de la historia evolutiva de los organismos.

uno que se van uniendo gradualmente formando grupos más grandes.

La agrupación jerárquica tiene la ventaja de ser simple y sus resultados pueden ser fácilmente visualizados. Sin embargo, un problema potencial con muchos métodos de agrupamiento jerárquico es que, como los conglomerados crecen en tamaño, el vector de expresión que represente a dicho conglomerado ya no podrá representar ninguno de los genes en particular del grupo. En consecuencia, en cuanto el conglomerado crece, los patrones de expresión individuales de cada gen se vuelven menos relevantes. Además, si una mala asignación se realiza al comienzo del proceso, no podrá ser corregida a futuro. No obstante, dicha técnica se ha convertido en una de las más utilizadas para el análisis de datos de expresión génica y es un enfoque de aglomeración en la que se unen los perfiles de expresión individuales formando grupos, que se unirán aún más hasta que el proceso finalice, formando un solo árbol jerárquico. Éste proceso trabaja de una manera simple:

1. Se calcula la matriz de distancias para todos los pares de genes a ser agrupados.
2. Se busca en la matriz de distancias a los dos genes o agrupaciones más similares. Inicialmente, cada grupo constará de un solo gen. Esta es la primera etapa en el proceso de aglomeración.
3. Los dos genes seleccionados se combinarán para producir un nuevo conglomerado que ahora contendrá al menos dos objetos.
4. Se vuelven a calcular las distancias entre este nuevo grupo y todos los demás genes. Existen en este punto diferentes métodos que permiten calcular la distancia entre un grupo y un gen o entre grupos, el cual debe ser seleccionado al momento de realizar el agrupamiento. Uno de los más comunes, y el utilizado en el presente trabajo es el del vecino más cercano (neighbor joining) donde la distancia es aquella entre los genes más cercanos de cada grupo o bien del gen más cercano del grupo al otro gen.
5. Por último, los pasos 2, 3 y 4 se repiten hasta que todos los objetos estén en un único conglomerado.
6. Determinado el número de grupos deseados, se realiza el corte. Dicho número puede prefijarse, o bien, realizar un corte según una distancia máxima entre genes dentro del grupo, y así obtener una cierta cantidad de grupos.

3. Resultados

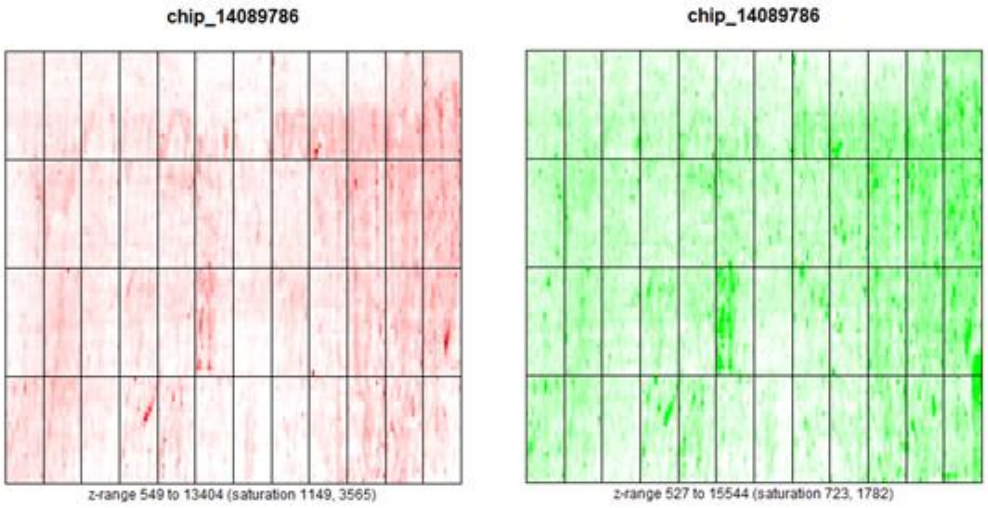
El trabajo realizado por Daurelio y col. incluyó el análisis de las interacciones de plantas del género de las Rutáceas y distintas especies de bacterias fitopatógenas del género Xanthomonas spp., definidos en la Tabla 1.

El estudio de las interacciones o tratamientos se llevaron a cabo en un microarreglo de dos colores, con tres repeticiones para cada experimento. Los análisis de conglomerados para identificar gupos de genes que co-expresen en los distintos tratamientos, fueron aplicados sobre los datos correspondientes a tres de los nueve tratamientos detallados: VLov, VXcc1 y VCtr1. En un primer análisis visual de los datos crudos se observan las imágenes obtenidas de ambos canales del microarreglo, en las cuales es posible ver diferentes intensidades en los colores, llevando a creer a priori que existen genes que se han expresado de manera diferencial (Figura 2).

Tabla 1. Interacciones planta-patógeno analizadas.

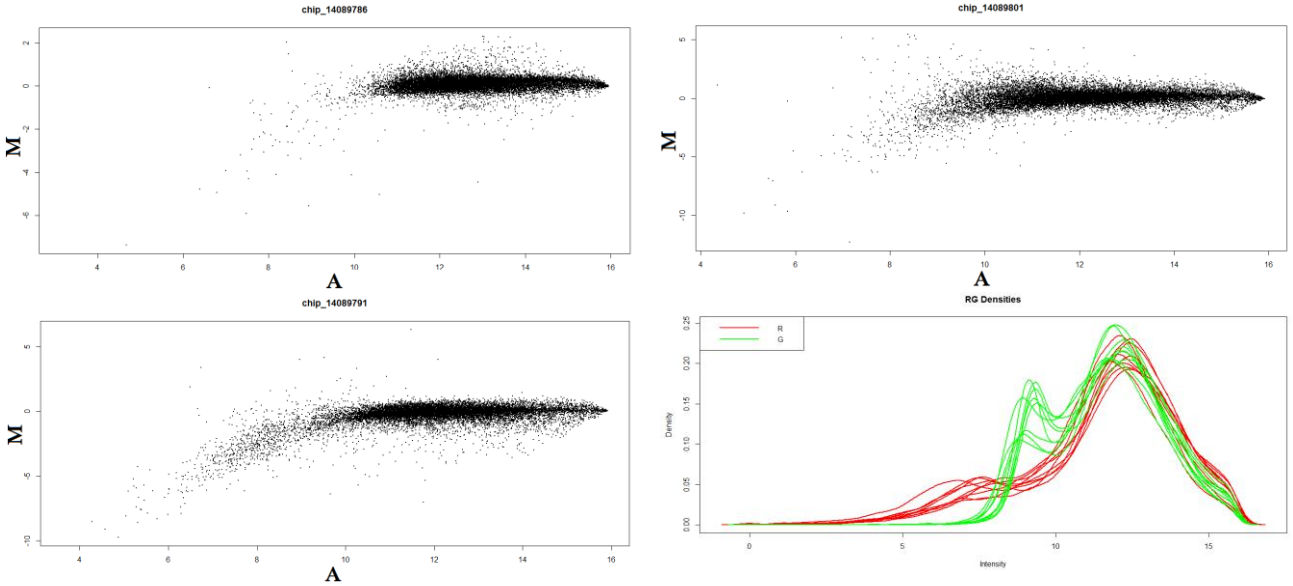
Interacción	Planta	Patógeno y respuesta que produce	Identificador de la interacción
1	*Citrus sinensis* "V", naranjo.	*Xanthomonas citri* subsp. *citri,* "Xcc", bacteria que enferma cítricos.	VXcc
2	*Citrus sinensis* "V", naranjo.	*Xanthomonas campestris* pv. *vesicatoria,* "Xcv", bacteria que no enferma cítricos, genera una respuesta de defensa por parte de la planta.	VXcv
3	*Citrus sinensis* "V", naranjo.	Xcc mutante en genes de secreción del sistema tipo III (SSTT), "Mut", que no enferma cítricos pero despertaría la defensa de tipo basal.	VMut
4	*Citrus sinensis* "V", naranjo.	Diluyente, con 8 horas de exposición post tratamiento, "Ctr", control negativo.	VCtr
5	*Fortunella margarita* "K", quinoto o kumquat.	"Xcc", no enferma quinoto, genera una respuesta de defensa por parte de la planta.	KXcc
6	*Fortunella margarita* "K", quinoto o kumquat.	Diluyente con 8 horas de exposición post tratamiento, "Ctr", control negativo.	KCtr
7	*Citrus sinensis* "V", naranjo.	Xcc mutante en el gen *lov*, "Lov", enferma cítricos pero con un incremento de la defensa de tipo basal.	VLov
8	*Citrus sinensis* "V", naranjo.	Diluyente con 24 horas (1 día) de exposición post tratamiento, "Ctr1", control negativo.	VCtr1
9	*Citrus sinensis* "V", naranjo.	"Xcc" con 24 horas de exposición post tratamiento, "Xcc1".	VXcc1

Figura 2. Imágenes obtenidas para un tratamiento con "Xcc" para 24 horas, en color rojo para las muestras y verde para las referencias Referencia.



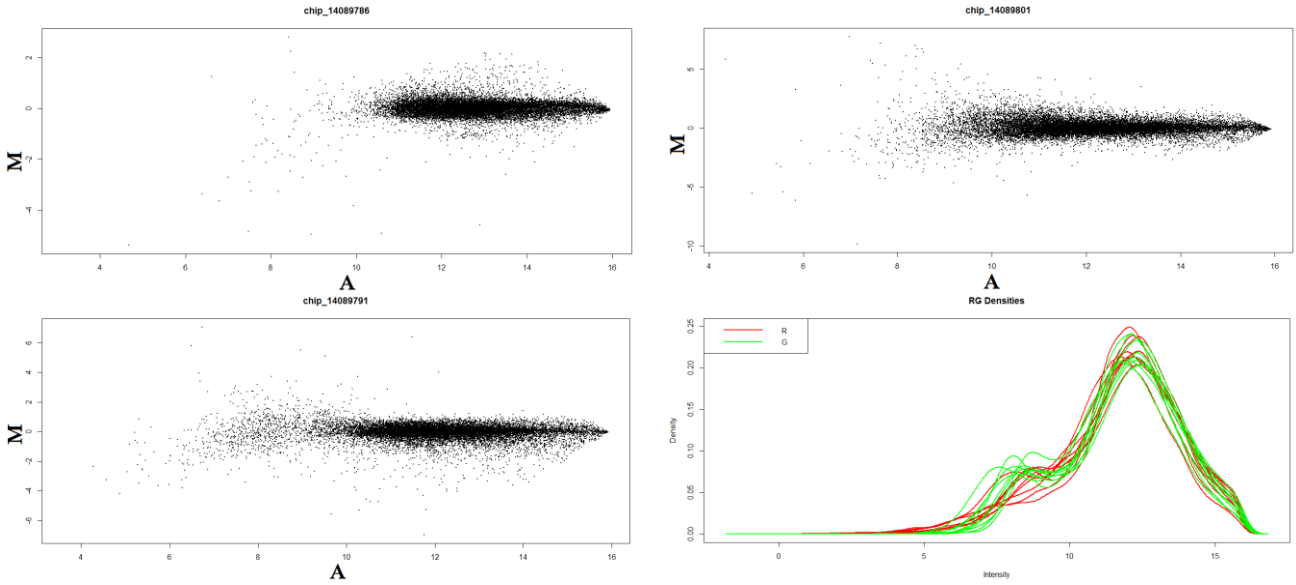
Mediante la observación de los gráficos MA de la Figura 3 es posible detectar que si bien la mayoría de los datos de distribuyen de manera aleatoria alrededor del cero, hay un gran número de ellos que se encuentran muy por encima o por debajo de dicho valor. Por otro lado, el gráfico de densidad hace posible pensar en una asimetría por izquierda en la distribución de las densidades de ambas intensidades. A fin de reducir los errores sistemáticos, y encontrar aquellos genes que realmente se expresaron de manera diferencial a causa del tratamiento y no de variables ajenas al mismo, se aplica una normalización de los datos mediante una regresión lineal local.

Figura 3. Gráficos MA y de densidades para los datos crudos.



Nota: Se presenta un MA para una réplica biológica correspondiente a cada interacción Ctr1, VXcc1 y VLov y las densidades para todas las réplicas.

Figura 4. Gráficos MA y de densidades para los datos normalizados.



Nota: Se presenta un MA para una réplica biológica correspondiente a cada interacción Ctr1, VXcc1 y VLov y las densidades para todas las réplicas.

En la Figura 4 se presentan los gráficos MA para los datos normalizados, en los cuales se observa que el número de puntos alejados de la línea del valor cero son menos, además de mostrarse menos alejados de la misma. En la misma figura se observan las curvas correspondientes a los nueve microarrelgos de las interacciones Ctr1, VXcc1 y VLov, en rojo para las muestras y en verde las respectivas de las referencias. Dicho gráfico, para los datos normalizados muestra una curva un poco más suave, la cual sigue presentando una asimetría a izquierda, pero de menor intensidad y haciendo que las curvas de la muestra y de la referencia sean similares. Todo esto indica que la aplicación de una normalización a las intensidades originalmente observadas fue correcta.

Una vez normalizados los datos, se ajustó un modelo lineal mediante el paquete “limma” del software estadístico R, el cual está definido como:

$$\boldsymbol{y_g} = \begin{pmatrix} y_{g1} \\ y_{g2} \\ y_{g3} \\ y_{g4} \\ y_{g5} \\ y_{g6} \\ y_{g7} \\ y_{g8} \\ y_{g9} \end{pmatrix} = \boldsymbol{X\alpha_g} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} VCtr1 \\ VLov \\ VXcc1 \end{pmatrix}; para\ el\ gen\ g, con\ g = 1, \dots, 22176.$$

Se debe tener en cuenta que en cada experimento, las intensidades medidas fueron comparadas con una intensidad referencia, por lo tanto, la variable respuesta y_{gi} corresponde al $\log_2$ de la razón entre la intensidad del gen g en el tratamiento i (R_{gi}) y la intensidad de dicho gen en la referencia i (G_{gi}), por lo tanto $y_{gi} = M_{gi} = log_2 \frac{R_{gi}}{G_{gi}}$.

A partir del modelo estimado, se plantean los siguientes seis contrastes de interés, con el objetivo de evaluar si cada uno de los genes está diferencialmente expresado en alguno de los experimentos:

- ✓ VXcc1-VCtr1=0
- ✓ VLov-VCtr1=0
- ✓ VXcc1-VLov=0
- ✓ VXcc1=0
- ✓ VLov=0
- ✓ VCtr1=0

Se crea la matriz de contrastes, con forma:

$$C = \begin{pmatrix} -1 & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Luego, el objetivo es probar si el vector $\boldsymbol{\beta_g} = \boldsymbol{C^T \alpha_g} = \begin{pmatrix} -VCtr1 + VXcc1 \\ -VCtr1 + VLov \\ -VLov + VXcc1 \\ VXcc1 \\ VLov \\ VCtr1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$.

Con esto, se prueba si cada uno de los genes está diferencialmente expresado en alguno de los experimentos.

Al realizar las pruebas de contrastes para los 22176 genes, se obtiene como resultado una base de datos para cada una de las comparaciones, compuestas por 22176 filas y doce variables, que contienen información acerca de la ubicación del gen en el microarreglo, nombre y medidas descriptivas de la razón de las expresiones del gen en los tratamientos que se analizan. Entre estas, se encuentran:

- ID o Spot (identificación del gen)
- Valor de M observado
- Valor de A observado
- Valor de la estadística t observada
- p valor correspondiente a la prueba, corregido por el método FDR
- Desvío estándar del gen en esa comparación, SE_g

De las seis bases obtenidas presentan interés biológico aquellas en las que un tratamiento es comparado con otro, por lo cual se consideran en adelante sólo las tres bases obtenidas de los contrastes entre tratamientos, las cuales se suman a las nueve comparaciones realizadas previamente por Daurelio y col. de las cuales se dispone de los datos procesados. Considerando todos los datos, se cuenta en el presente trabajo con los resultados de las siguientes doce comparaciones o contrastes de interés biológico:

- ✓ VXcc-VCtr=0
- ✓ VXcv-VCtr=0
- ✓ VMut-VCtr=0
- ✓ VLov-VCtr1=0
- ✓ VXcc1-VCtr1=0
- ✓ VMut-VXcv=0
- ✓ VXcc-VMut=0
- ✓ VXcc-VXcv=0
- ✓ VXcc1-VLov1=0
- ✓ (VMut-VCtr)-(KXcc-KCtr)=0
- ✓ (VXcc-Vctr)-(KXcc-Kctr)=0
- ✓ (VXcv-VCtr)-(KXcc-KCtr)=0

Estas doce bases de datos fueron unidas en una única base que contiene el ID del gen y las variables M, p-value y SE de cada comparación, para todos los genes, con el fin de realizar un análisis multivariado de los datos y estudiar la co-expresión de los genes en los distintos experimentos.

La Figura 5 muestra los gráficos de volcán para dos de las comparaciones realizadas.

Figura 5.

Gráfico de volcán para la comparación VXcv-VCtr.

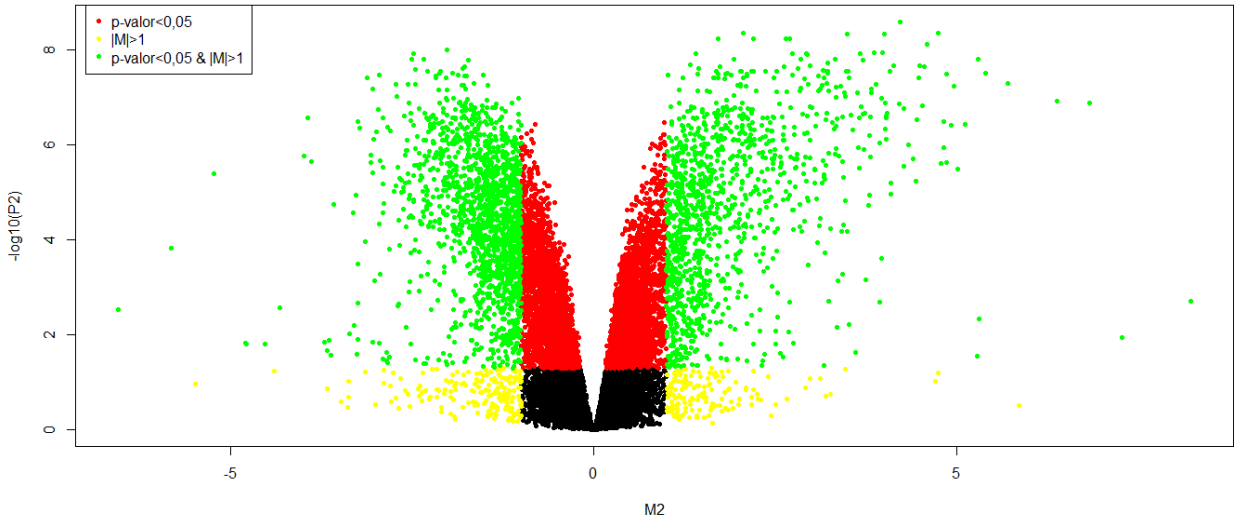
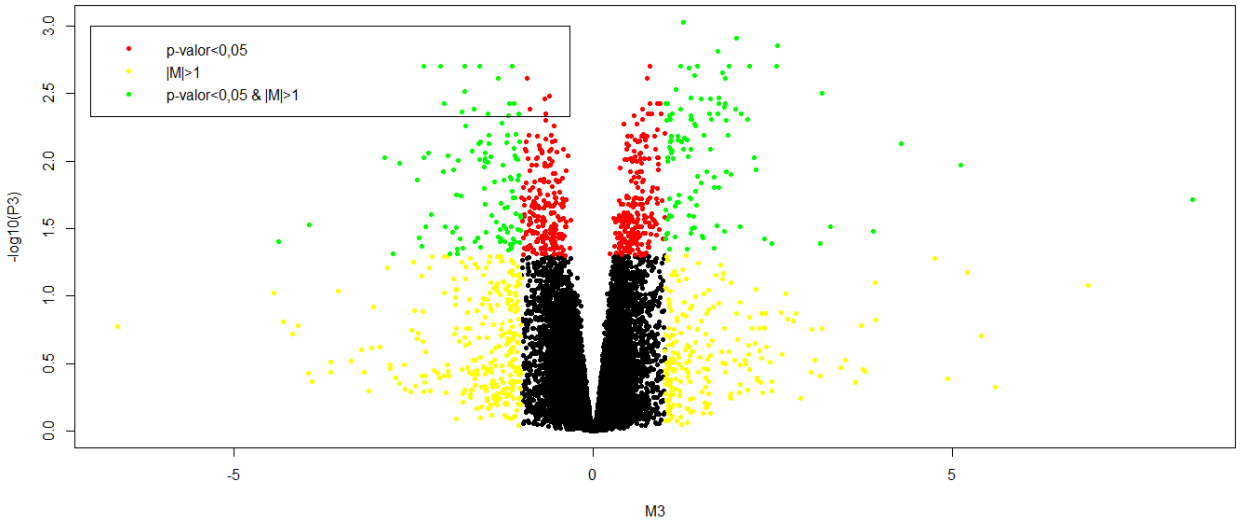


Gráfico de volcán para la comparación VMut-VCtr



La base de datos generada con todas las comparaciones fue analizada con la finalidad de obtener aquellos genes que presentaran expresión diferencial en al menos una de las comparaciones. De los 22176 genes analizados, 12350 fueron detectados como diferencialmente expresados en al menos un tratamiento considerando los resultados de la prueba t. La Tabla 2 muestra el resumen de los resultados obtenidos tanto de las pruebas de hipótesis como del fold change, y la combinación de ambos, en términos de número y porcentaje respecto del total, de genes expresados de manera diferencial en cada una de las comparaciones. Se observa que las comparaciones VXcv-VCtr, VXcv-VMut y VXcc-VXcv son las que presentaron un mayor porcentaje de genes expresados de manera diferencial (entre el 30 y el 40%). El máximo número de genes expresados diferencialmente fue de 8681, esto muestra que el uso de las pruebas t, aún con un pequeña muestra, es fundamental para reducir el tamaño de la base final a analizar. Esta reducción permitirá realizar experimentos más específicos, con más repeticiones, y pruebas estadísticas con mayor potencia.

Tabla 2. Resumen del análisis de genes expresados diferencialmente en la base de datos.

	Genes con |M|>1		***Genes con p<0,05***		***Genes con |M|>1 y p<0,05***	
Comparación	**Número**	**Porcentaje**	**Número**	**Porcentaje**	**Número**	**Porcentaje**
1.VXcc-VCtr	639	2,88%	429	1,93%	140	0,63%
2.VXcv-VCtr	3088	13,92%	8681	**39,15%**	2738	**12,35%**
3.VMut-VCtr	724	3,26%	662	2,99%	216	0,97%
4.VLov-VCtr1	2125	9,58%	4023	18,14%	1622	7,31%
5.VXcc1-VCtr1	718	3,24%	973	4,39%	286	1,29%
6.VXcv-VMut	2252	10,16%	7131	**32,16%**	1918	**8,65%**
7.VXcc-VMut	404	1,82%	17	0,08%	12	0,05%
8.VXcc-VXcv	2298	10,36%	6753	**30,45%**	1974	**8,90%**
9.VXcc1-VLov	877	3,95%	1123	5,06%	373	1,68%
10.(VMut-VCtr)-(KXcc-KCtr)	1432	6,46%	680	3,07%	364	1,64%
11.(VXcc-VCtr)-(KXcc-KCtr)	1482	6,68%	946	4,27%	449	2,02%
12.(VXcv-VCtr)-(KXcc-KCtr)	2538	11,44%	5145	23,20%	1696	7,65%

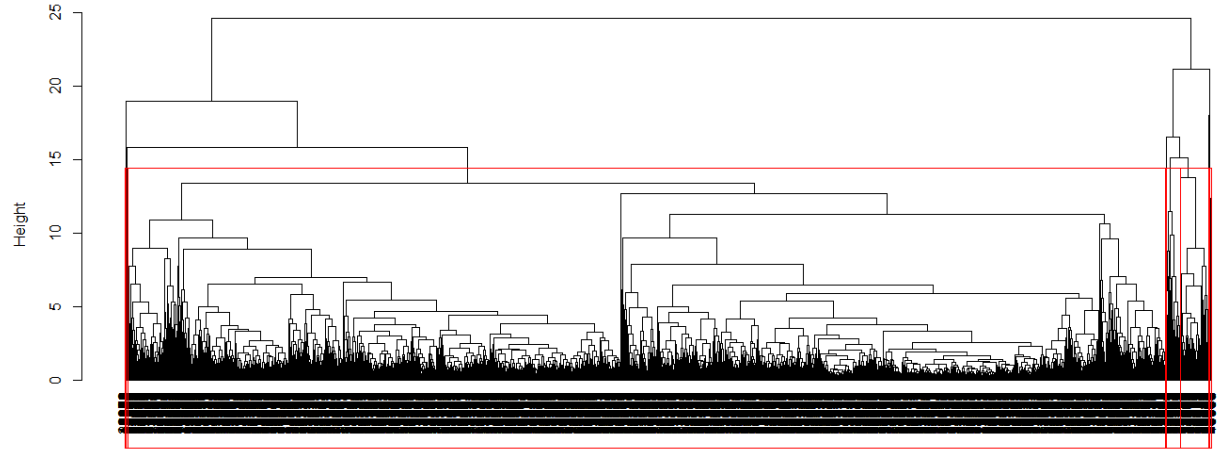
El análisis exploratorio, utiliza por lo general los dos criterios, tanto el del fold change como las pruebas estadísticas y si se consideran ambos, también se observa que los mayores porcentajes de genes que cumplen con ambas condiciones se dan en las mismas tres interacciones mencionadas (Tabla 2). Estas tienen como interacción común a VXcv, entre la planta de naranjo y la bacteria que no enferma cítricos sino que genera una respuesta de defensa por parte de la planta Xcv, y a esto se puede deber que un gran número de genes se hayan expresado diferencialmente.

El análisis multivariado llevado a cabo consta de un Análisis Clúster jerárquico con el fin de encontrar y agrupar aquellos genes que se expresaron de manera similar en los diferentes experimentos. En un primer paso, con el fin de agrupar sólo a aquellos genes que se hayan expresado significativamente de manera diferencial, se aplica un filtro a la base completa, mediante el cual se obtiene una nueva base con los 12350 genes que presentaron un valor de p menor a 0,05 ($\alpha = 5\%$).. Luego se procedió a eliminar aquellos genes que contienen datos faltantes en alguna comparación (necesario para calcular la matriz de distancias), y la base definitiva quedó compuesta por 10350 genes. Con dicha base se realizó un análisis clúster, en el cual se agruparon los genes según los valores de M observados, obteniéndose el dendograma presentado en la Figura 6. Sobre el mismo se realizó un corte utilizando la herramienta *rect.hclust,* definiendo una distancia máxima de 15 unidades y se pudo observar la formación de ocho grupos posibles de genes conformados por distintas cantidades de genes, las cuales se detallan en la Tabla 3.

Tabla 3. Cantidad de genes en cada grupo con una distancia máxima de 10 unidades entre genes.

Grupo	1	2	3	4	5	6	7	8
Cantidad	9891	266	136	19	7	6	20	5

Figura 6. Dendograma para genes diferencialmente expresados en al menos una comparación.



Si bien el objetivo del análisis es encontrar aquellos genes que están diferencialmente expresados, la decisión no debe basarse exclusivamente en el resultado de las pruebas de hipótesis, sino también en el valor de fold change obtenido para cada gen en cada comparación. Siguiendo esta línea, se aplicaron otros filtros considerando los valores del fold change con el fin de obtener bases de datos más pequeñas, con genes que presentaron una tasa de cambio más importante.

A continuación, en las figuras 7 y 8, se presentan los resultados de realizar un análisis clúster con los datos obtenidos al filtrar aquellos genes que presentaron un valor absoluto de M mayor a 4 y posteriormente valores absolutos de M mayores a 5, es decir, que la expresión en un experimento fue 16 o 32 veces la misma expresión en el otro experimento. Las bases obtenidas contaban con 92 y 32 genes respectivamente.

Figura 7. Dendograma para genes diferencialmente expresados en al menos una comparación con valores absolutos de M mayores a 4.

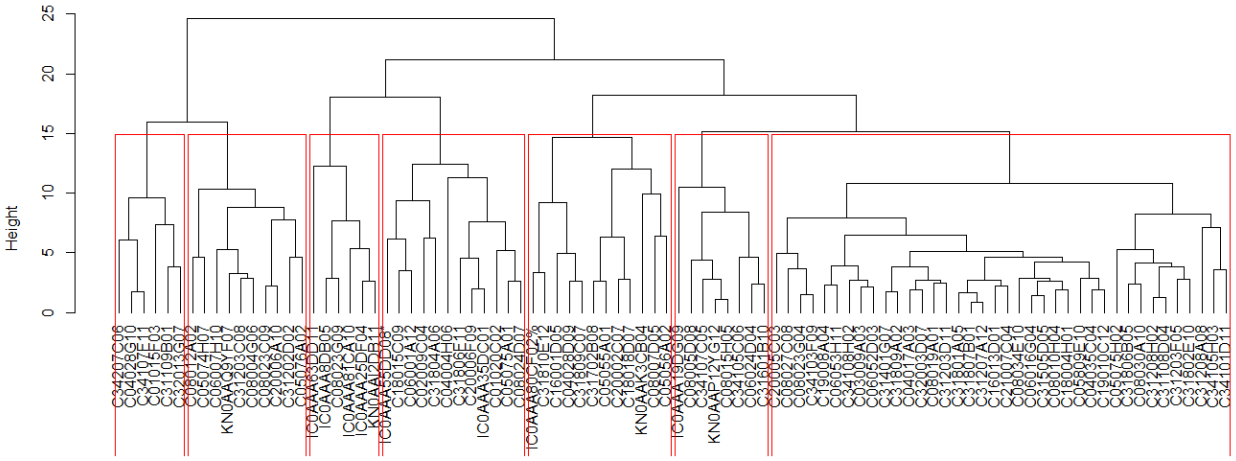
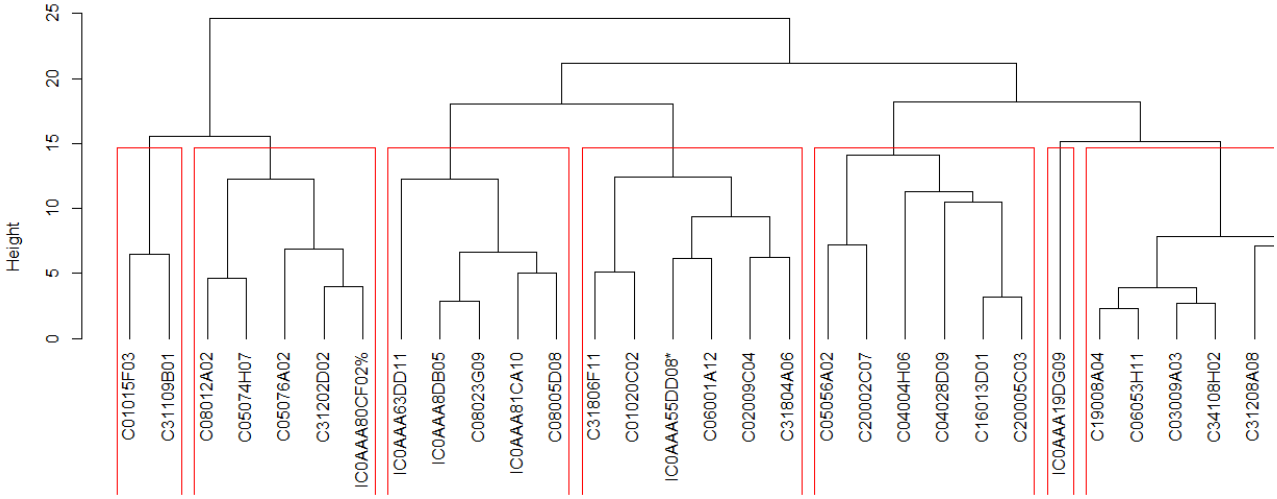


Figura 8. Dendograma para genes diferencialmente expresados en al menos una comparación con valores absolutos de M mayores a 5.



A partir de estos resultados, los especialistas en el tema pueden determinar cuál es la función biológica de cada gen que conforma cada uno de los grupos y evaluar así, si genes que tienen funciones similares se expresaron de manera similar.

4. Conclusiones

En este trabajo se ha presentado una reseña de los procedimientos utilizados para el análisis de datos de expresión génica obtenida a partir de la metodología de microarreglos y una aplicación de los mismos a datos reales correspondientes a un problema de investigación en el que se intenta identificar grupos de genes con patrones de expresión similares.

El paso principal previo al análisis es la normalización de los datos, con el objetivo de eliminar el ruido o error sistemático que pueden presentar los datos. Este procedimiento fue llevado a cabo utilizando la metodología de regresión lineal local, aunque cabe destacar que es posible la elección de otras metodologías para el mismo fin.

En un paso posterior, se ajustaron modelos lineales y a partir de éstos se plantearon contrastes de interés biológico. Las pruebas de contrastes se llevaron a cabo a partir de una estadística t modificada, la cual utiliza todos los datos para el cálculo de una variancia común, debido a la escasa cantidad de repeticiones en cada uno de los experimentos. Los escasos tamaños muestrales en el contexto de los experimentos con microarreglos se deben a un impedimento principalmente económico, ya que los mismos son muy costosos. No obstante, la salvedad que se hace es que si bien el método es estadístico y se realiza una inferencia, la misma es tomada en forma exploratoria, para reducir el tamaño de la base original y realizar, en estudios posteriores, análisis con mayor profundidad y mayores tamaños de muestra de los genes que hayan resultado de interés.

Por otro lado, dado el gran tamaño de las bases de datos obtenidas, una de las formas prácticas de presentar los resultados es de manera gráfica. En ese sentido, se ha presentado y ejemplificado el uso de los gráficos MA, específicos para este tipo de aplicaciones.

Por último, se aplicó a algunos subgrupos de genes de interés un análisis multivariado de aglomeración, el análisis clúster jerárquico, el cual permitió agrupar a los genes en función de sus patrones de expresión, a partir de lo cual los especialistas pueden investigar si los genes que se expresaron de manera similar comparten también funciones biológicas similares.

5. Referencias bibliográficas

Benjamini Y, Hochberg Y. (1995) *Controlling the False Discovery Rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological), 57(1):28

Cai XZ, Zhou X, Xu YP, Joosten MH y de Wit PJ. (2007). *Cladosporium fulvum CfHNNI1 induces hypersensitive necrosis, defence gene expression and disease resistance in both host and nonhost plants.* Plant Mol Biol 64: 89-101.

Chisholm ST, Coaker G, Day B y Staskawicz BJ. (2006). *Host-Microbe interactions: Shaping the evolution of the plant immune response*. Cell 124: 803-814.

Daurelio LD, Romero MS, Petrocelli S, Merelo P, Cortadi AA, Talón M, Tadeo FR, Orellano EG. (2013). *Characterization of Citrus sinensis transcription factors closely associated with the non-host response to Xanthomonas campestris pv. vesicatoria*. Journal of Plant Physiology, 170(10): 934-42.

Gordon K. S, Matthew R, Natalie T James Wettenhall, Wei Shi and Yifang Hu. (2014) *Linear Models for Microarray Data. User's Guide*. Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

Jacoby WG. (2009). *Loess: a nonparametric, graphical tool for depicting relationships between variables*. Electoral Studies, 19: 577-613.

Jones JD y Dangl JL. (2006). *The plant immune system*. Nature 444: 323-32.

Savio Rodríguez Baena, D. (2006). *Análisis de datos de Expresión Genética mediante técnicas de Biclustering.* Memoria del período de investigación. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla.

Storey JD, Tibshirani R. (2003) *SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays.*

Parmigiani G, Garret E, Irizarri R y Zeger S (eds.) (2003). *The analysis of gene expression data: Methods and Software*.

Utah State University –Spring 2014. *Introduction to Filtering with Gene Expression Data*. STAT 5570: Statistical Bioinformatics Notes 3.2.

Wan J, Dunning FM y Bent AF. (2002). *Probing plant-pathogen interactions and downstream defense signaling using ADN microarrays.* Functional and Integrative Genomics 2: 259–273.

Wise RP, Moscou MJ, Bogdanove AJ y Whitham SA. (2007). *Transcript profiling in host-pathogen interactions.* Annual Review of Phytopathology, 45: 329-369.

Xiangqin C y Churchill GA. (2003) *Statistical tests for differential expression in cDNA microarray experiments.* Genome Biology, 4: 210.

6. Anexo

• El ácido desoxirribonucleico, abreviado como ADN, es un ácido nucleico que contiene instrucciones génicas usadas en el desarrollo y funcionamiento de todos los organismos vivos conocidos, y es responsable de su transmisión hereditaria. La función principal de la molécula de ADN es el almacenamiento a largo plazo de información. Muchas veces, el ADN es comparado con un plano o una receta, o un código, ya que contiene las instrucciones necesarias para construir otros componentes de las células, como las proteínas y las moléculas de ARN. Los segmentos de ADN que llevan esta información génica son llamados genes, pero las otras secuencias de ADN tienen propósitos estructurales o toman parte en la regulación del uso de esta información génica.

• El ácido ribonucleico mensajero o ARNm es el que contiene la información génica (el código genético) procedente del ADN del núcleo celular, es decir, el que determina el orden en que se unirán los aminoácidos de una proteína y actúa como plantilla o patrón para la síntesis de dicha proteína.

• El ADN complementario o ADN copia (ADNc) es una hebra de ADN de doble cadena, una de las cuales constituye una secuencia totalmente complementaria del ARNm a partir del cual se ha sintetizado. Se suele utilizar para la clonación de genes propios de células eucariotas en células procariotas.