Capítulo 1

Métrica mixta

En este capítulo utilizaremos los resultados obtenidos en el capítulo anterior para desarrollar una variante al procedimiento córte de árbol dinámico que integre la información contenida en el espacio de expresión con la información contenida en el espacio de conocimientos GO, para favorecer la búsqueda de estructuras biológicamente coherentes. La idea principal de esta variante es la de montar sobre la heterogeneidad topológica (transcripcional) un desorden de pesos a partir de la información biológica, y con ello obtener una métrica mixta que contenga información de ambos espacios.

1.1. Hacia una métrica mixta

Existen muchas formas de mezclar las métricas del espacio de expresión y del espacio GO. Una de las métricas mixtas investigadas previamente por el grupo [1] fue la de mezcla convexa de distancias, donde se define una nueva distancia d_{mix} a partir de las distancias de expresión y GO y un parámetro α que controla la contribución de cada métrica al algoritmo:

$$d_{mix} = \sqrt{\alpha d_x^2 + (1-\alpha)\tilde{d}_{GO}^2} \tag{1.1}$$

con $\tilde{d}_{GO} = \frac{\langle d_x \rangle}{\langle d_{GO} \rangle} d_G O$, de tal manera que coincidan los valores medios de ambas distribuciones de distancia.
Esta métrica busca el consenso a partir del parámetro global α, que parametriza de manera continua una métrica en contraposición con la otra.
Para este trabajo, implementamos una métrica mixta que en lugar de buscar el consenso entre las métricas, penaliza las correlaciones no soportadas por las distintas ontologías e incentiva aquellas que si lo están .

1.1.1. Modificación de la similaridad de correlación

Para obtener una métrica que permita detectar heterogeneidades dentro de los grupos, modificamos la similaridad de correlación utilizando la información obtenida mediante el índice KTA local en las redes de 30 primeros vecinos mutuos.
Para ello, tomamos cada grupo de cada tratamiento y calculamos el KTA de todo el grupo, además del KTA local de cada arista del grupo. Estas dos cantidades, el KTA de grupo, que llamaremos KTA_{fondo}, y el KTA local de la arista entre dos nodos i y j, que llamaremos $KTAl_{ij}$, se relacionan en una cantidad llamada $stress$ que utilizaremos como criterio para identificar si debemos penalizar o incentivar una correlación en el espacio de expresión. El $stress$ se define como:

$$stress = \frac{KTA_{fondo}}{KTAl_{ij}} \tag{1.2}$$

Si $KTAl_{ij} > KTA_{fondo}$, se obtiene un $stress < 1$, mientras que si $KTAl_{ij} < KTA_{fondo}$, obtenemos que $strees > 1$. Por lo tanto, si realizamos la transformación no lineal de la similaridad de correlación:

$$w_{ij} = simcor_{ij}^{stress} \tag{1.3}$$

obtenemos unos nuevos pesos para la similaridad de correlación en donde aquellas relaciones que son similares en GO ($KTAl_{ij} > KTA_{fondo}$) serán incentivadas en el espacio de expresión y a la inversa, las que no son similares en GO serán penalizadas en el espacio de expresión. La figura 1.1 presenta la distribución de $stress$ para el tratamiento 'Frío' en función del tamaño de grupo. Se observa que el 92 % de los valores de $stress$ se ubica entre 0,8 y 1,2.

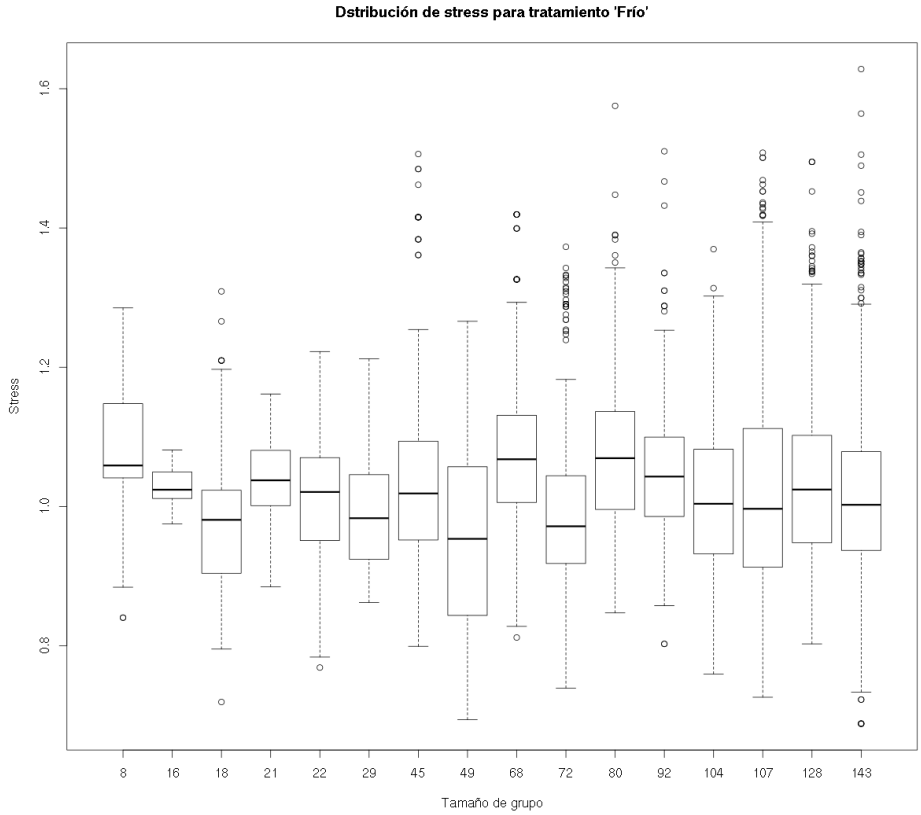

Figura 1.1: Distribución de *stress* para el tratamiento 'Frío' y ontología CC en función del tamaño de grupo.

Este rango en los valores del *strees* no permite penalizar adecuadamente las similaridades en el espacio de expresión, lo que implica que será necesario modificar la escala del *strees* proporcionalmente por medio de un parámetro β, de tal forma que los pesos de la ecuación 1.3 se transformen como:

$$\tilde{w_{ij}} = w_{ij}^{\beta} \tag{1.4}$$

Para encontrar el beta adecuado, haremos uso de la información contenida en la red.
Como analizamos en el capítulo anterior, las redes de k primeros vecinos mutuos no poseen una distribución de grado con cola pesada. Una cola pesada es indicativo de que la distribución de grado sigue una ley de potencias, es decir, $p(k) \approx k^{-\gamma}$, y por lo tanto la red tiene una topología de tipo libre de escala. Muchas redes reales son de este tipo. Horvath, en [2], propone buscar un parámetro β al cual elevar la similaridad de correlación de genes tal que la red de genes obtenida siga una distribución de tipo ley de potencias.
Buscamos el parametro β adecuado realizando un barrido para β desde 1 hasta 4 con pasos de 1 y desde 15 hasta 65 con pasos de 5, transformando los pesos de la red w_{ij}

mediante la ecuación 1.4 y realizando un ajuste lineal para el logaritmo de los pesos en función del logaritmo de la probabilidad. Tomamos el parámetro β que mejor ajustaba cada tratamiento en el sentido de R cuadrado y lo utilizamos para modificar los pesos w_{ij}. Por ejemplo, para el tratamiento 'Frío', obtuvimos $\beta = 4$, mientras que para 'Calor' obtuvimos $\beta = 55$. Una vez obtenidos los parámetros para la métrica mixta, desarrollamos un método heurístico para poder aplicar esta nueva métrica mixta.

1.2. Método heurístico

El método heurístico desarrollado consiste en tomar cada grupo de una partición realizada previamente con corte de árbol dinámico e intentar particionarlo de tres formas distintas.
La primera forma consiste en aplicar sobre cada grupo nuevamente corte de árbol dinámico, utilizando la métrica mixta en la confección del dendrograma, penalizando (incentivando) las conexiones más importantes, que son las que aparecen en la red de k primeros vecinos mutuas con alto (bajo) stress, método que llamaremos *lkta.dtc*. Por otro lado, la segunda y la tercera formas, que llamaremos *lkta.infomap* y *lkta.cnm*, consisten en obtener comunidades mediante infomap y cnm, respectivamente, en la red de los genes del grupo con los pesos modificados por la métrica mixta. Una vez obtenida una partición del grupo, calculamos el índice BHI de los subgrupos y volvemos a unir aquellos subgrupos que se encuentren por debajo de una desviación estandar del control nulo 1 presentado en la sección **??** mediante un algoritmo voraz o greedy, que en cada paso busca los dos subgrupos tales que el BHI de los dos juntos sea superior al BHI de cada uno por separado. Cuando el algoritmo no consigue unir dos subgrupos para mejorar el BHI, se detiene.
Finalmente, todos los subgrupos que todavía quedan por debajo de una desviación estandar del control nulo 1 son unidos entre sí en un único subgrupo.
Utilizaremos un cuarto método, llamado *insideX*, a modo de control. El mismo consiste en volver a particionar cada grupo usando corte de árbol dinámico con *deepsplit* $= 4$, pero sin cambiar la métrica por la métrica mixta. Esto nos permitirá controlar la efectividad de la métrica mixta para encontrar mayor resolución en las particiones.
Para cuantificar el cambio en la información biológica que brinda la nueva partición, elegimos tomar el valor medio del BHI de los subgrupos en comparación con el BHI del grupo original, $\langle BHI \rangle$, y el valor medio del BHI de los subgrupos que superan el control nulo, $\langle BHI \rangle_+$. A modo de ejemplo, presentamos en las figuras 1.2 y 1.3 los resultados de esta heurística aplicada a los grupos 2 y 9 del tratamiento 'Frío'. En las mismas, la curva roja indica el valor medio para la distribución del BHI del control nulo 1, la verde una desviación estandar por sobre la media y la negra, una desviación estandar por debajo. El punto lleno representa el grupo original y su BHI. En verde si su BHI supera el control nulo y en rojo si no. Los puntos vacíos en gris representan

los subgrupos nuevos. En la leyenda, el primer número es $\langle BHI \rangle$ y el segundo, entre paréntesis, es $\langle BHI \rangle_+$.

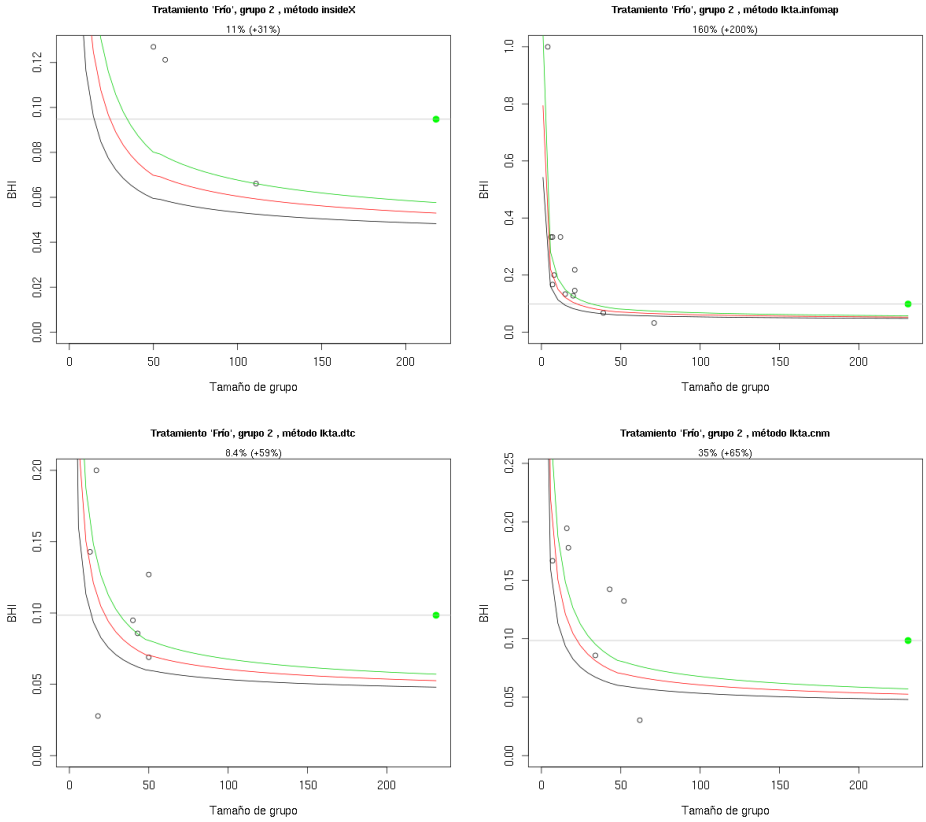

Figura 1.2: Métodos con métrica mixta y control para el grupo 2 del tratamiento 'Frío'.

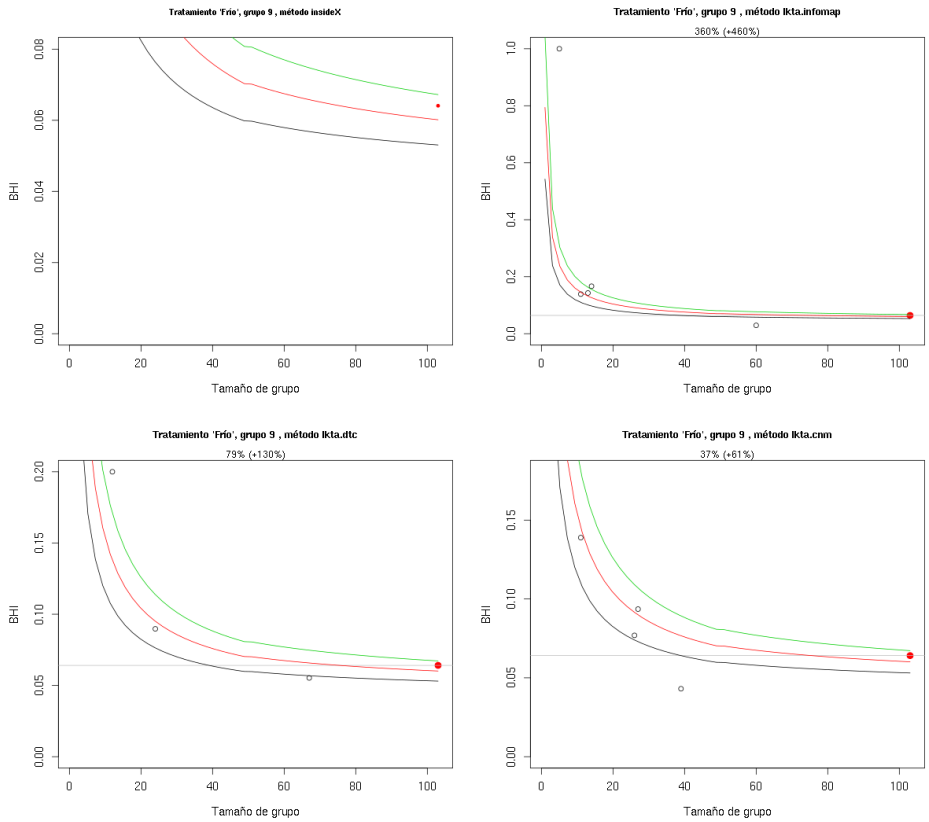

Figura 1.3: Métodos con métrica mixta y control para el grupo 9 del tratamiento 'Frío'.

Se observa que para el grupo 2, un grupo con un BHI por sobre el control nulo y el segundo grupo más grande de la partición, el método *insideX* logró encontrar heterogeneidades y partir el grupo, obteniendo tres subgrupos de más de 50 genes cada uno y una mejora del 11 % en $\langle BHI \rangle$, y una del 31 % para $\langle BHI \rangle_+$. Por otro lado, los tres métodos *lkta* lograron superar la mejora obtenida por *insideX* en $\langle BHI \rangle_+$, y tanto *lkta.infomap* como *lkta.cnm* lograron mejorar además el obtenido en $\langle BHI \rangle$, siendo *lkta.infomap* el que mejores puntajes logró, llegando a obtener incluso un grupo con un BHI de 0,97.
Para el grupo 9, un grupo con un BHI por debajo del control nulo y relativamente pequeño, el método *insideX* no logró encontrar heterogeneidades dentro del grupo y por lo tanto no pudo particionarlo, sin lograr ningún tipo de mejora en el mismo. Sin embargo, todos los métodos *lkta* lograron particionar el grupo y mejorar ambos índices, siendo nuevamente *lkta.infomap* el que logró la mejora más significativa, con un 360 % de incremento en $\langle BHI \rangle$ y un 460 % de incremento en $\langle BHI \rangle_+$.
Finalmente, realizamos este análisis para todos los tratamientos y todos los grupos en cada tratamiento. Para cada método, graficamos el $\langle BHI \rangle$ de los subgrupos en función del BHI del grupo original en la figura 1.4. En rojo, graficamos una recta de

pendiente unitaria. Se observa que todos los métodos que utilizan la métrica mixta consiguieron grupos con BHI superiores a los que tenían originalmente. En particular, para lkta.infomap el 92 % de los grupos fue dividido en subgrupos que poseen un BHI promedio superior al de su grupo original, mientras que lo mismo sucede en un 91 % de los grupos para lkta.cnm y un 87 % para lkta.dtc. Finalmente, para el control insideX, solo el 10 % de los grupos mejoró su BHI promedio al ser particionado nuevamente.

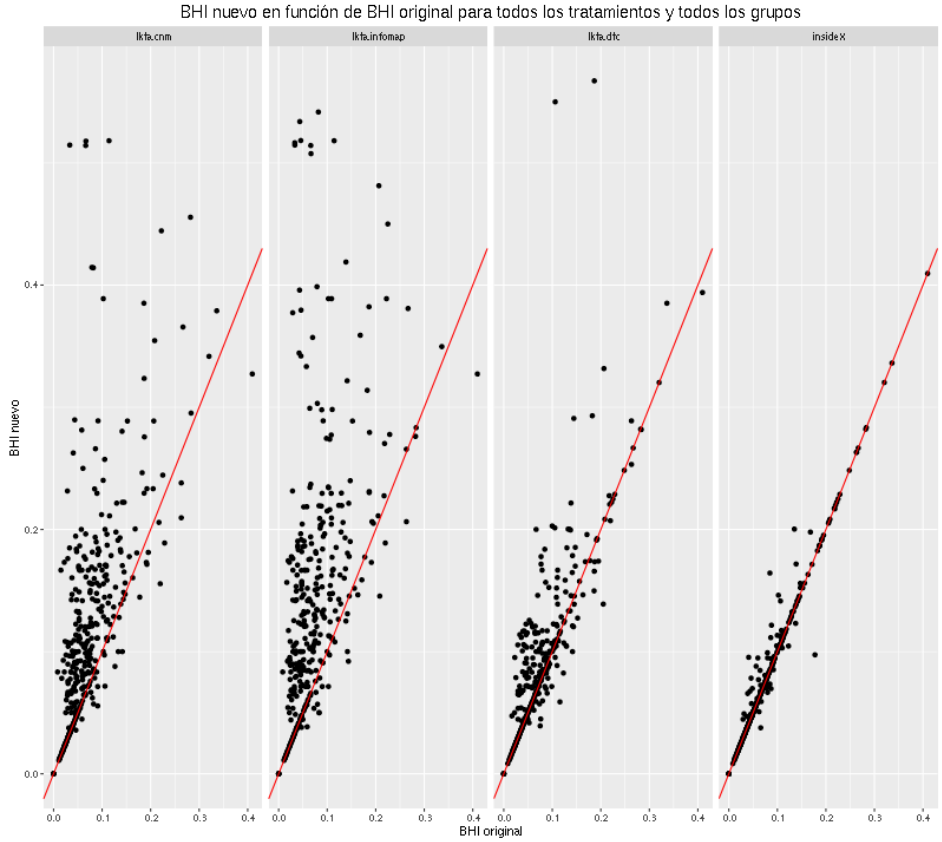

Figura 1.4: BHI nuevo en función de BHI original para todos los tratamientos y todos los grupos.

1.3. Resultados

Una vez obtenidos los nuevos grupos mediante los distintos métodos, buscaremos ganar conocimiento sobre el significado biológico de cada grupo y en particular estudiaremos si los genes dentro de un mismo grupo participan de algún proceso biologico común. Para ello, recurriremos al análisis de sobreexpresión de nodos GO.
El mismo consiste en tomar la ontología BP o CC y un grupo y analizar si existen nodos estadísticamente sobreexpresados en el grupo, utilizando una prueba de Fisher exacta.

Notemos que como la prueba debe repetirse para cada nodo de la ontología, estamos frente a una prueba de múltiples hipótesis, y si por ejemplo, la ontología tuviera 1000 nodos, tomar un p-valor de 1 % implicaría que 10 nodos estarían sobreexpresados por azar. Para reducir la tasa de falsos descubrimientos, utilizaremos una corrección introducida por Benjamini y Hochoberg en [3].
contar cual es el pvalue que usamos
Realizamos esta prueba en los grupos 2, 6 y 13 del tratamiento 'Frío'. En la figura 1.5 se observan los grupos originales y los subgrupos obtenidos por medio de cada método junto con el control nulo 1 y los correspondientes valores de BHI, mientras que en las figuras 1.6, 1.7 y 1.8 se visualizan los resultados de la prueba exacta de Fisher para cada uno de estos grupos.

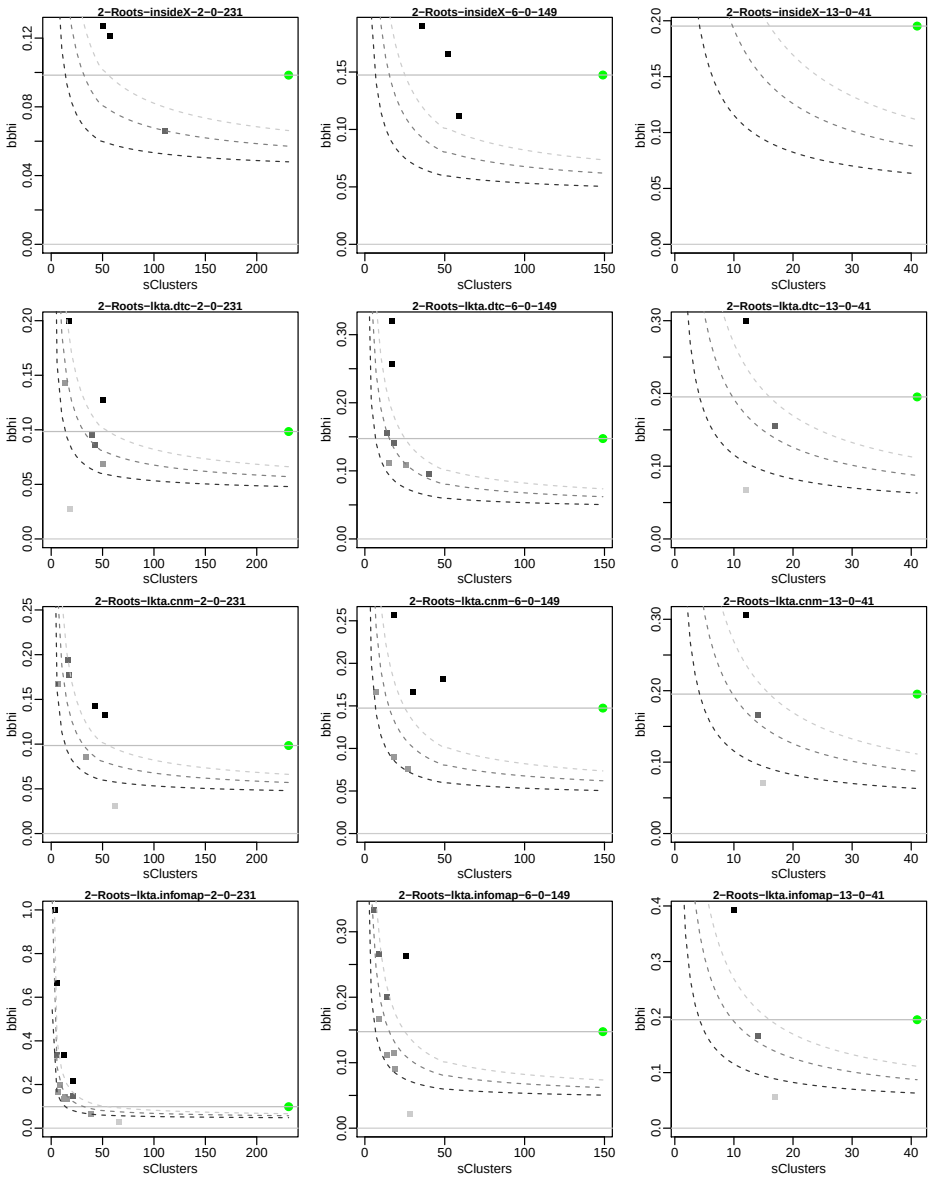

Figura 1.5: Métodos con métrica mixta y control para los grupos 2, 6, y 13 del tratamiento 'Frío'.

El cuadro principal representa el grupo original y los subgrupos para cada uno de los métodos, junto con los nodos más informativos que resultaron sobreexpresados en el test de Fisher y el proceso biológico al que están asociados. Los mismos se encuentran ordenados mediante un dendrograma por contenido de información y los casilleros coloreados corresponden al valor del p-valor obtenido para cada nodo, con el código de colores en el cuadro de arriba.

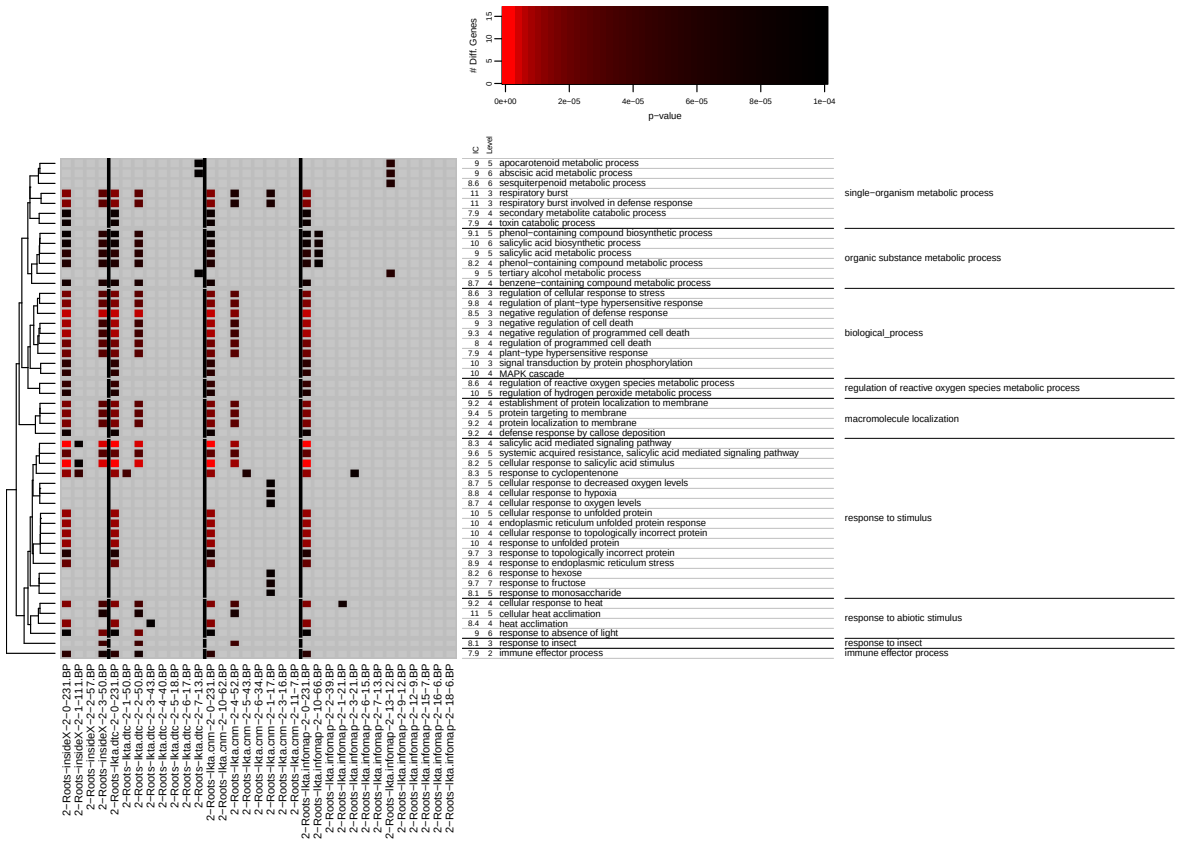

Figura 1.6: Resultado del test de sobreexpresión de nodos GO para el grupo 2 del tratamiento 'Frío' para los métodos con métrica mixta y control.

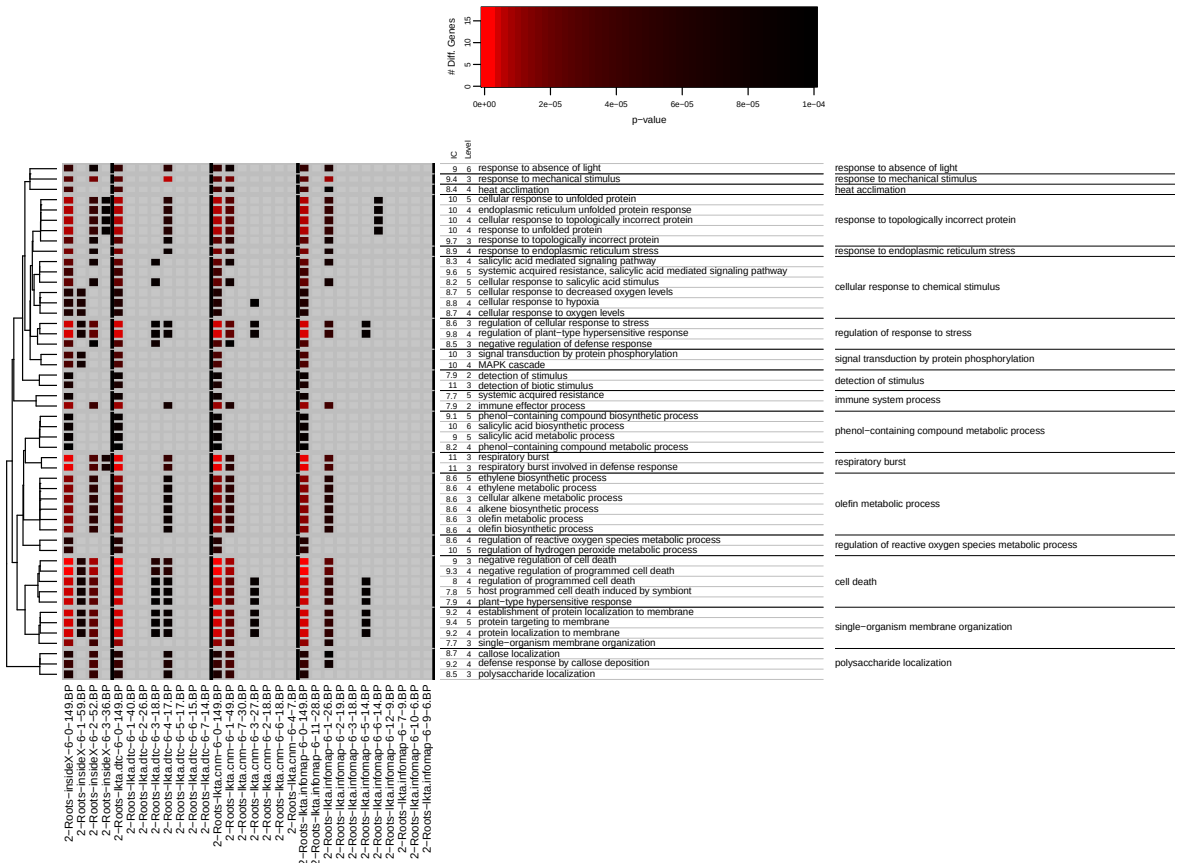

Figura 1.7: Resultado del test de sobreexpresión de nodos GO para el grupo 6 del tratamiento 'Frío' para los métodos con métrica mixta y control.

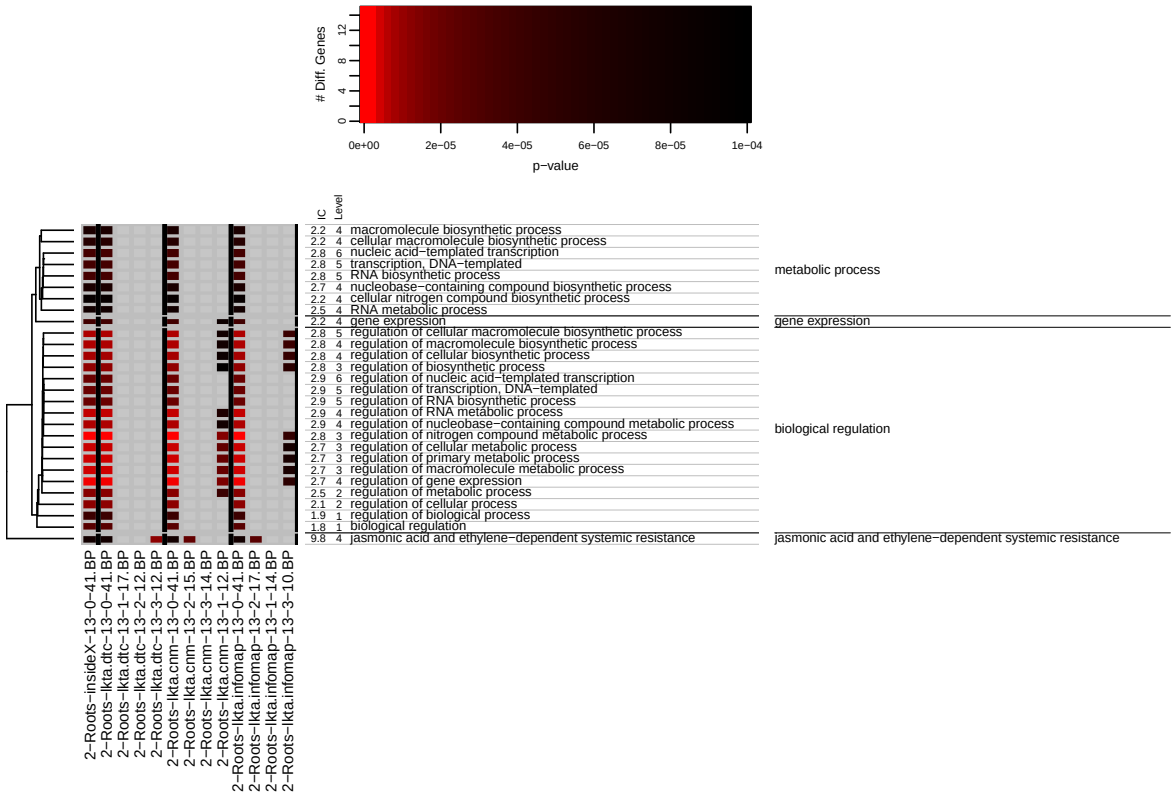

Figura 1.8: Resultado del test de sobreexpresión de nodos GO para el grupo 13 del tratamiento 'Frío' para los métodos con métrica mixta y control.

Para el grupo 2, el segundo grupo más grande de la partición, de 231 genes, y el método insideX, se observa que uno de los dos subgrupos con ganancia en BHI, el de 50 genes, explica casi en su totalidad los procesos biológicos del grupo original, mientras que no se encontró en el otro subgrupo con ganancia nodos sobreexpresados. Por otro lado, para lkta.dtc y lkta.cnmse obtuvo un subgrupo de 50 genes con ganancia en BHI que explica los procesos del grupo original mientras que el restro de los subgrupos con ganancia no realiza aportes significativos al conocimiento biológico. Finalmente, lkta.infomap, el de mayor ganancia en el índice BHI, ninguno de los subgrupos logra explicar los nodos sobreexpresados.
Para el grupo 6, un grupo de tamaño intermedio, con 149 genes, el método insideX consigue nuevamente dos subgrupos con ganancia en BHI, de 59 y 52 genes respectiva-

mente, que explican en conjunto la mayor parte de los nodos sobreexpresados del grupo original. Para lkta.dtc se obtienen dos subgrupos informativos de tamaño pequeño, de 18 y 17 genes respectivamente. Para lkta.cnm y lkta.infomap se obtuvieron un subgrupo informativo de 49 y otro de 26 genes respectivamente.
Finalmente, para el grupo 13, un grupo pequeño de 41 genes, el método insideX no logró producir subgrupos, mientras que solo para lkta.cnm y lkta.infomap se consiguieron grupos explicativos para un grupo de 12 y otro de 10 genes respectivamente.
Por un lado, observamos que en varios casos una gran parte de los nodos sobreexpresados del grupo original podían ser explicados por uno o dos subgrupos de tamaño reducido, lo que implica que existen efectivamente heterogeneidades en los grupos originales que no podía detectar el método de agrupamiento inicial $ds1$. Por otro lado, las estructuras detectadas fueron de tamaños intermedios, de alrededor de 50 o 60 genes. Esto da un indicio de que esta debería ser la resolución o escala a la cual deberíamos trabajar con los métodos de agrupamiento. Finalmente, este análisis nos permite ver que si bien lkta.infomap genera las particiones de grupos con mayor ganancia en el índice BHI, los subgrupos obtenidos no necesariamente brindan información biológica. Esto podría deberse al tamaño tan reducido de los subgrupos, que no facilita su interpretación en términos de GO.

Bibliografía

[1] A. J. Berenstein. *Técnicas de Mecánica estadística para la detección de correlaciones en perfiles de expresión génica.* Tésis de Licenciatura en Ciencias Físicas (2010).

[2] S. Horvat. *A general framework for weighted gene co-expression network analysis.* Stat. Appl. Genet. Mol. Biol. (2005).

[3] Y. Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* **57** (1995) 289.