Análisis y Detección de Correlaciones en Relevamientos Transcripcionales de Gran Escala

Andrés Rabinovich
Director: Dr. Ariel Chernomoretz

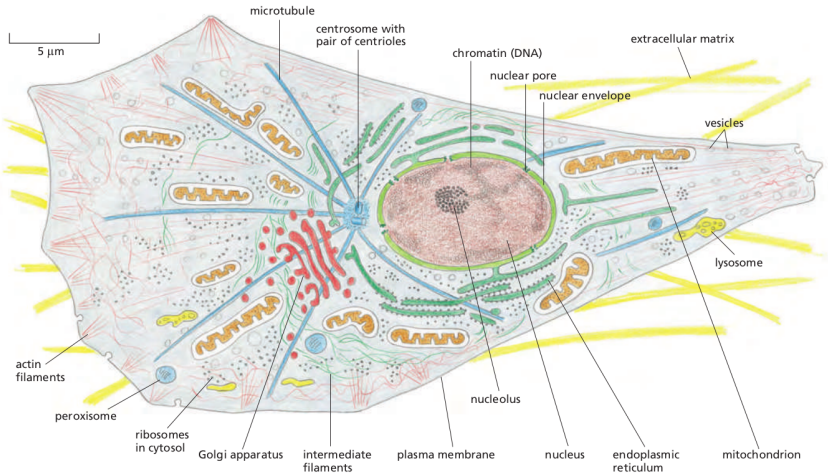
Departamento de Física
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

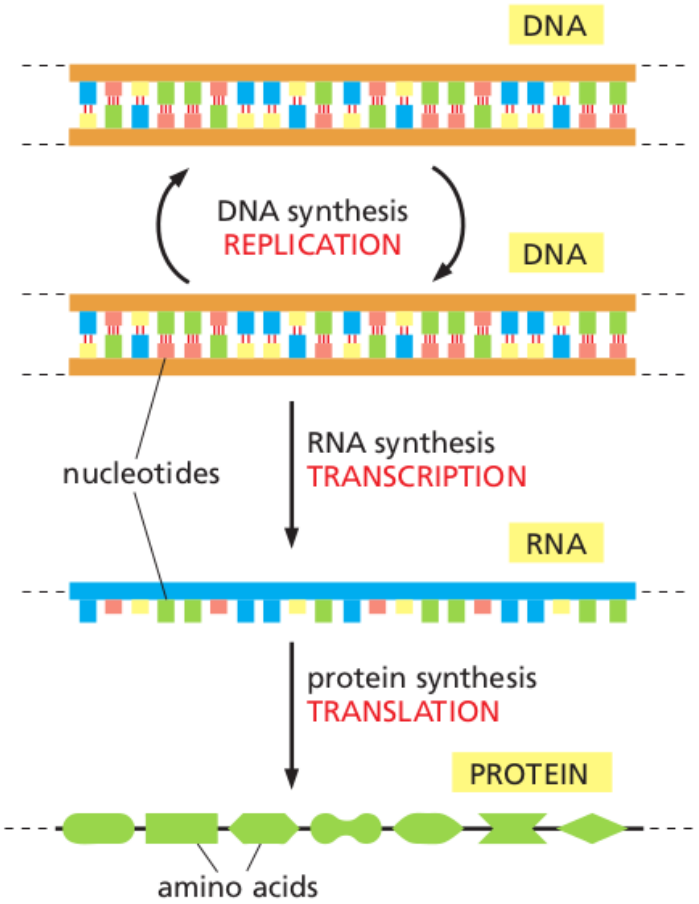
Marzo 2016.

Contenido

Transcripción y traducción (dogma central de la biología molecular)

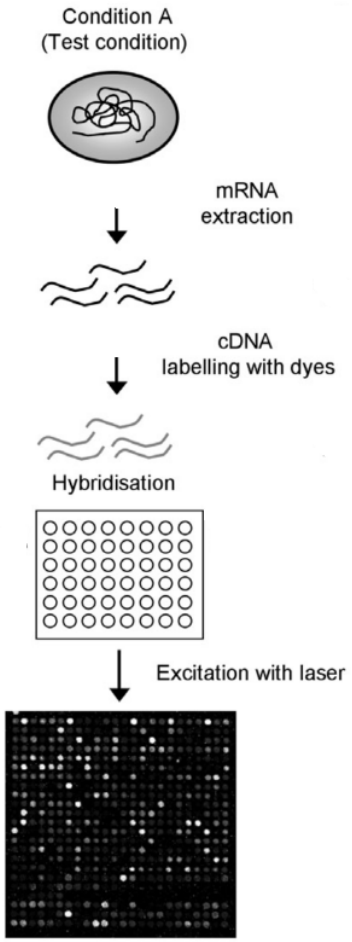
Células, ADN, ARNm, proteínas y otras yerbas...





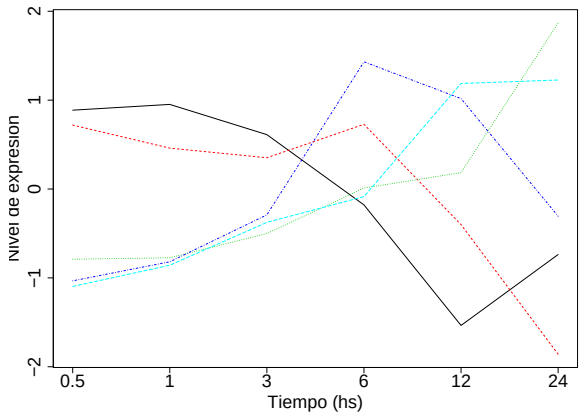
Cambios transcripcionales en respuesta a estrés abiótico en *A. thaliana*

A. Thaliana



Datos de estrés abiótico:

- 11 tratamientos: frío, calor, osmótico, salinidad, sequía, genotoxicidad, oxidación, UV, herida, recuperación y control.
- $\approx$ 22000 genes.
- Nos quedaremos con un subconjunto de $\approx$ 6000 genes que son los que se movieron en algún tratamiento.
- entre 4 y 8 mediciones temporales por gen y por tratamiento.



Detección de correlaciones

Queremos inferir estrategias del organismo frente a los tratamientos.

Lo vamos a hacer usando métodos de agrupamiento o "clustering" para encontrar relaciones y estructura en esta gran cantidad de datos.

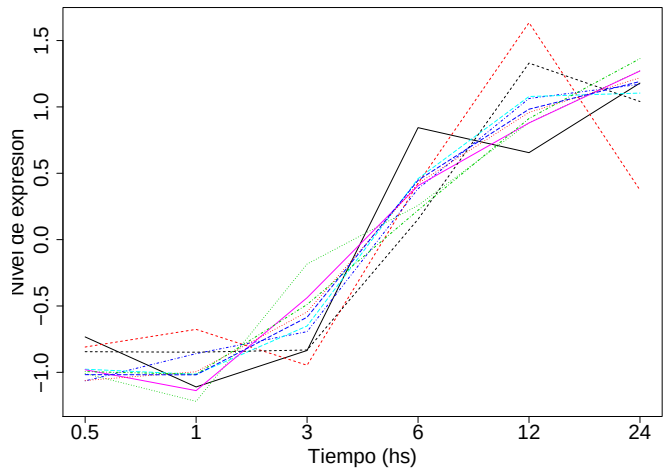
- Son métodos no supervisados.
- Consisten en agrupar elementos "similares entre si".
- Permiten el descubrimiento de patrones en los datos.
- Posibilitan obtener conclusiones sobre los datos.

A modo de ejemplo

El conjunto: $\{-5, -3, -2, 2, 3\}$

Agrupado por módulo: $\{-5\}$, $\{-3, 3\}$ y $\{-2, 2\}$

Agrupado por signo: $\{-5, -3, -2\}$ y $\{2, 3\}$

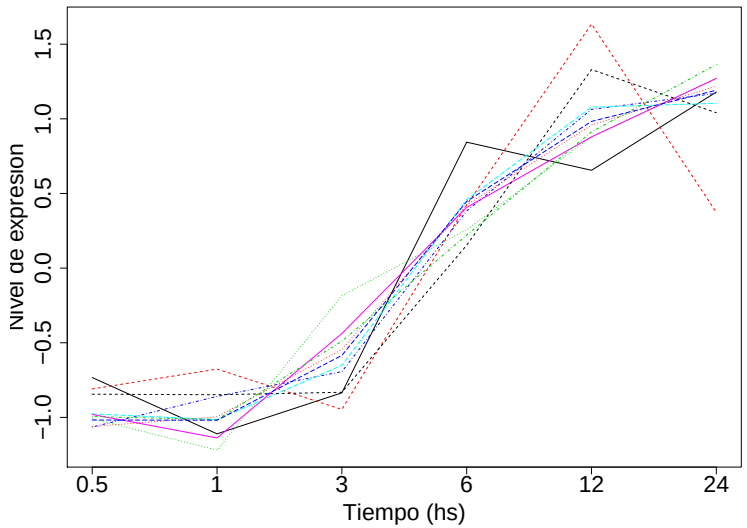


Medidas de similaridad y distancia

Distancia basada en el coeficiente de correlación de Pearson:

$$r(\vec{x},\vec{y}) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \tag{1}$$

$$d_{ccp}(\vec{x},\vec{y}) = 1 - r(\vec{x},\vec{y}) \tag{2}$$



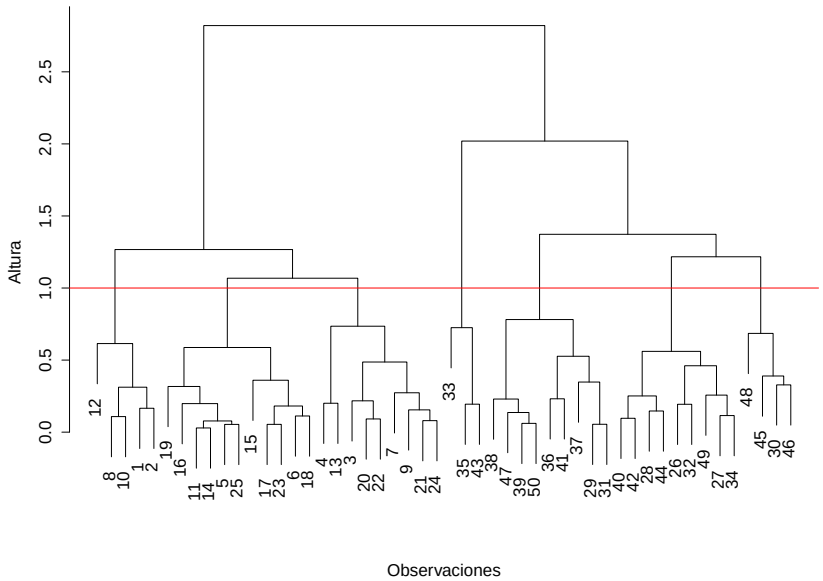
Métodos de agrupamiento

Método k-means

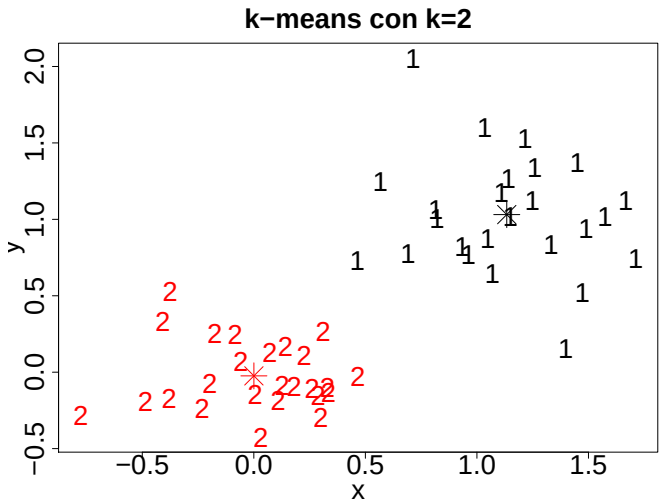
- Busca estructuras compactas.
- La cantidad k de grupos debe ser fijada a priori.
- Muy rápida ejecución.

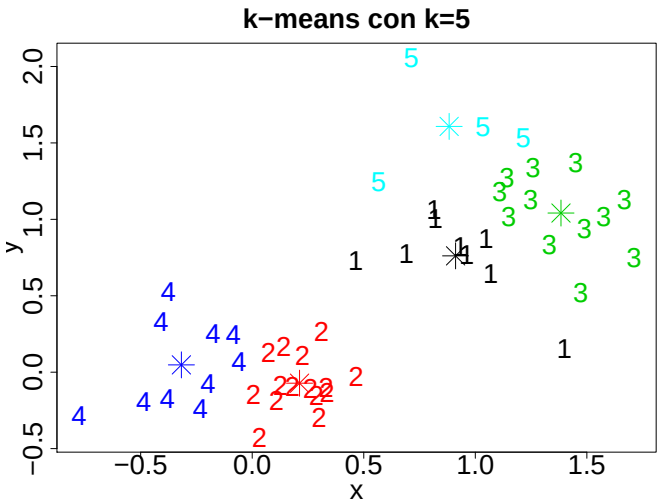
Método corte de árbol dinámico

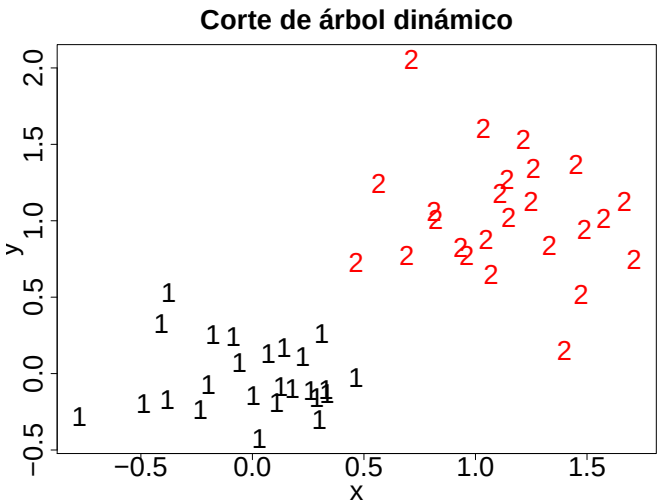
- Agrupamiento jerárquico.
- El agrupamiento puede representarse mediante un dendrograma.
- Utiliza la distancia de correlación.
- La cantidad óptima de grupos k es decidida por el método.



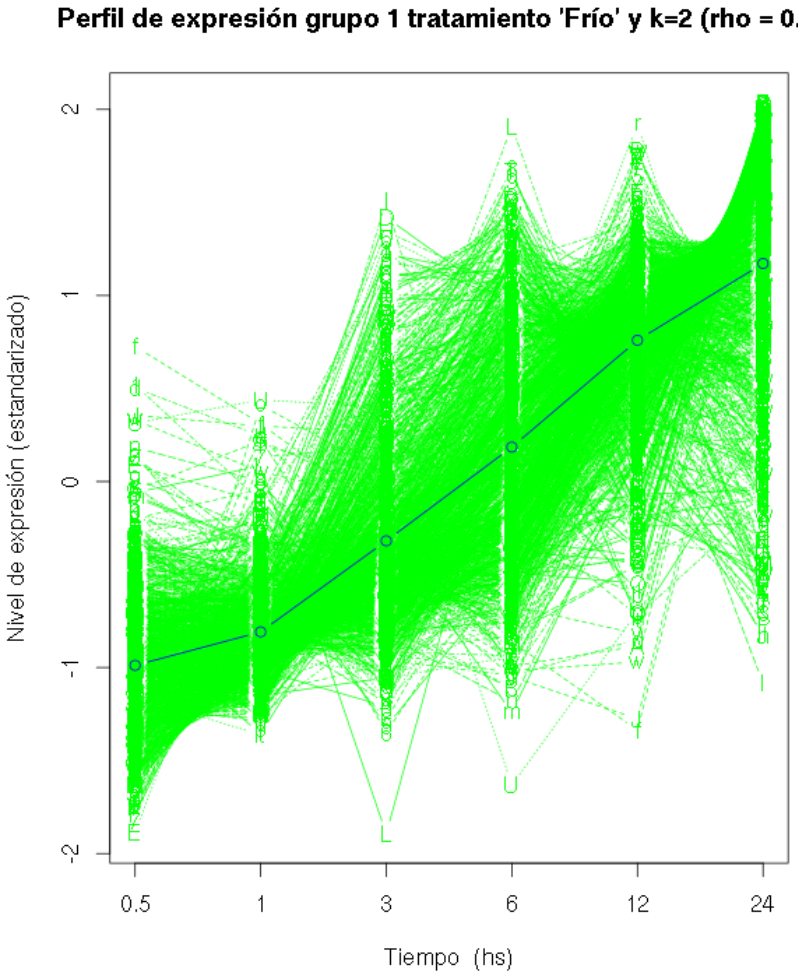
Métodos de agrupamiento - ejemplos

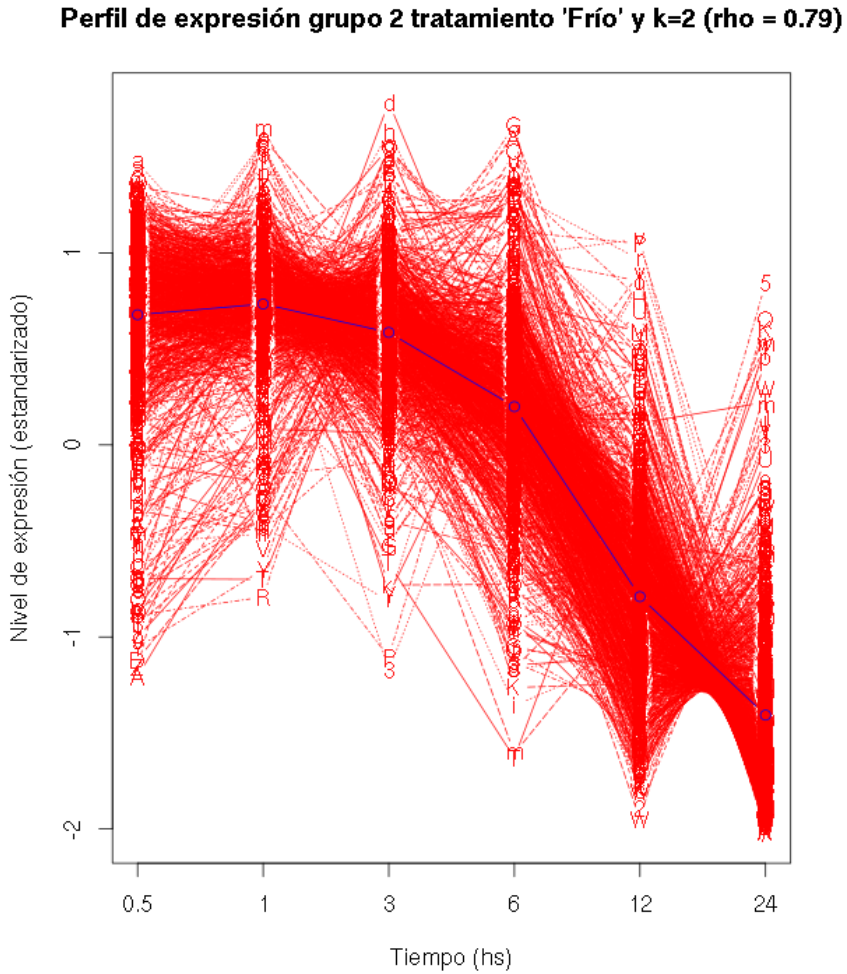






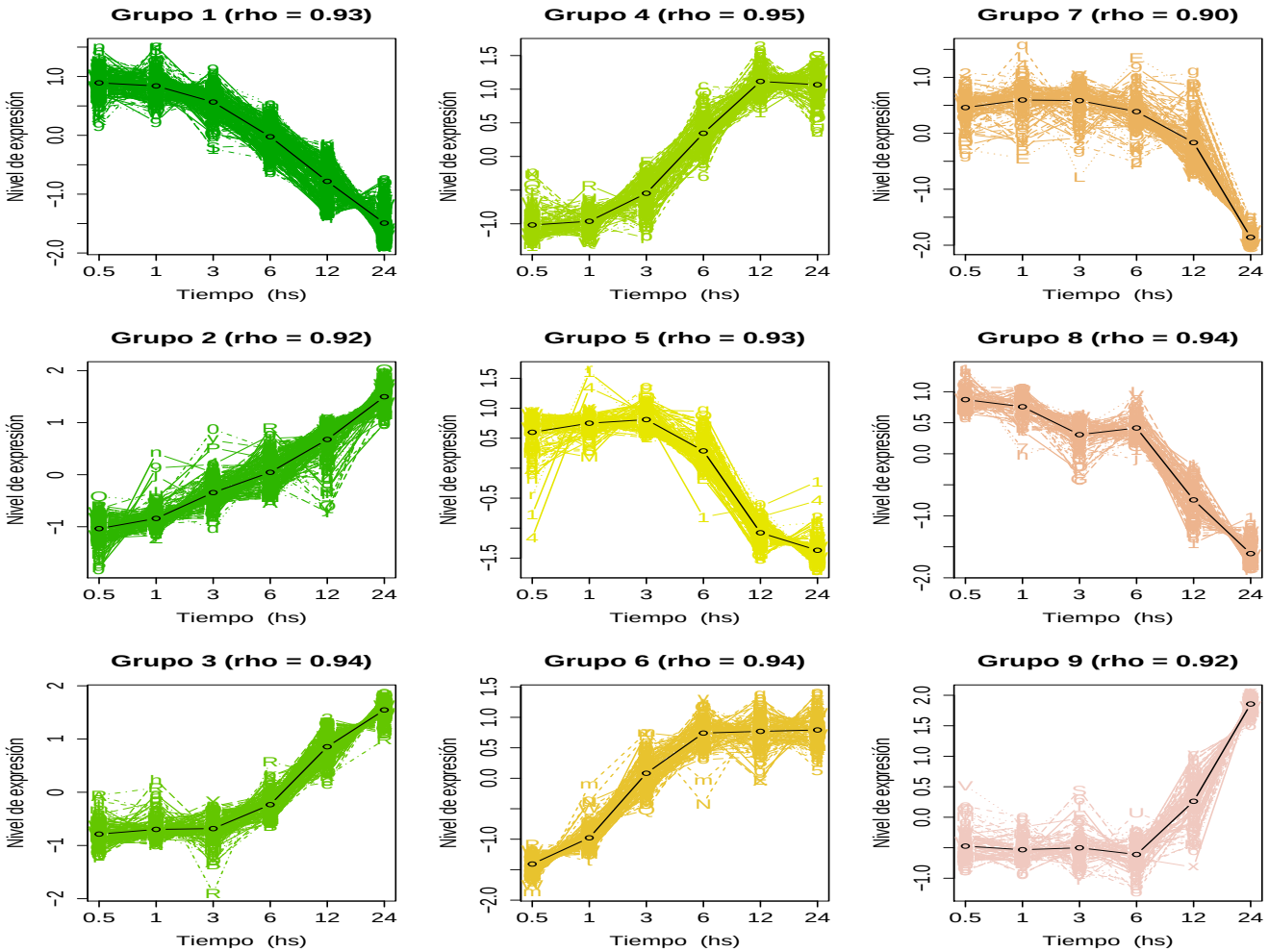
Perfiles tratamiento "Frío" con k-means





Perfiles tratamiento "Frío" con corte de árbol dinámico

A modo de ejemplo, los nueve perfiles más grandes de una partición de tratamiento "Frío" y DS1.



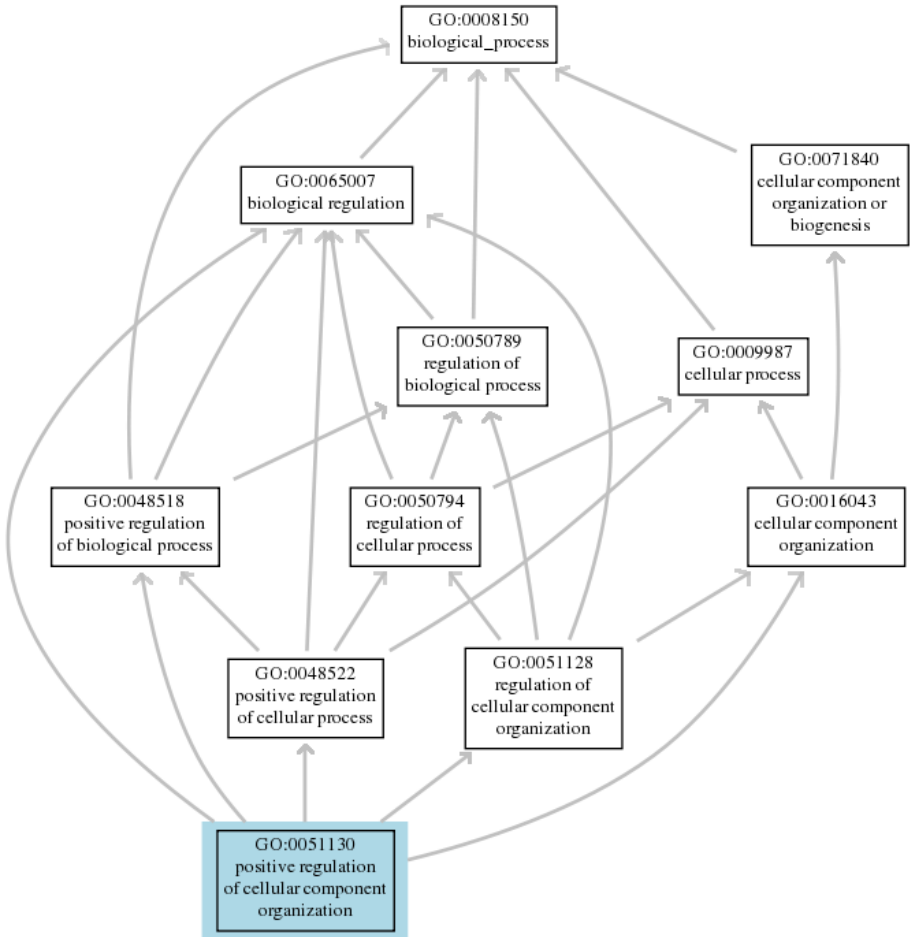
Granularidad de las particiones

Granularidad y resolución de los métodos

- Una partición A es más fina que una partición B si cada grupo de A está contenido en un grupo de B.
- Tenemos tres formas de realizar particiones de nuestros datos.
- DS4 genera particiones más finas que DS1 y este a su vez que k-means.
- Tenemos distintas maneras de encontrar estructura en nuestros datos y las distintas heterogeneidades aparecerán a distintas escalas.
- Deberemos buscar la escala óptima a la que trabajar con este conjunto de datos y para eso vamos a pasar a un espacio de conocimiento biológico.

Ontología génica (GO)

- Provee un vocabulario controlado de términos.
- Permite comparar y clasificar entidades biológicas.
- Tres ontologías: procesos biológicos (BP), componentes celulares (CC) y funciones moleculares (MF).
- Estructura de grafo acíclico dirigido (DAG).
- Cada nodo representa un témino que describe alguna función.
- Los nodos se unen entre si por medio de relaciones “es un” o “es parte de”.



Un gen descrito por un término está “anotado” en ese término.

Observables

Buscamos cuantificar la congruencia biológica de las particiones halladas

Densidad de interacción:

$$ID(GO_j) = \frac{NE(GO_j)}{N(GO_j)} \qquad (3)$$

Con $NE(GO_j)$ la cantidad de pares de genes anotados en GO_j que se encuentran juntos en un mismo grupo transcripcional C_x y $N(GO_j)$ la cantidad de pares de genes anotados en GO_j.

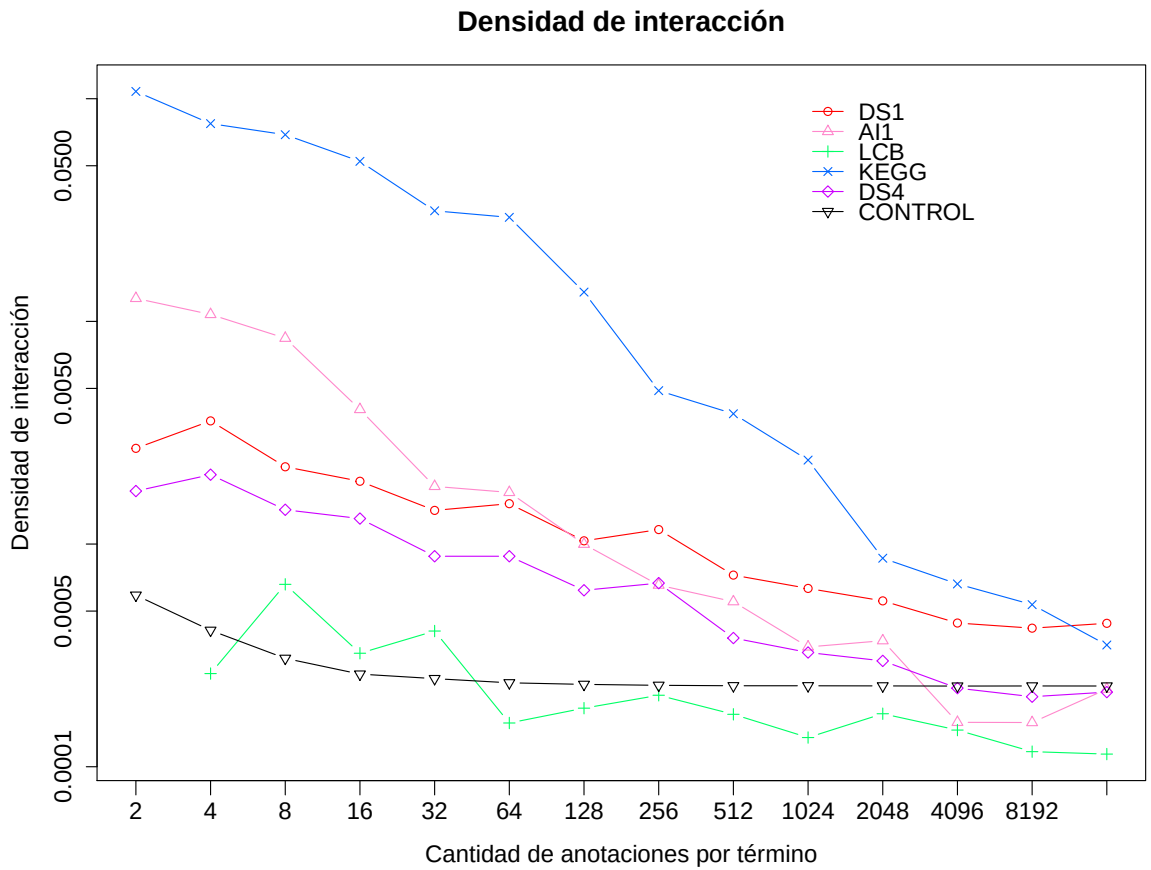
Indice de homogeneidad biológica:

$$BHI_j = \frac{1}{n_j(n_j-1)} \sum_{x \neq y \in D_j} I(C(x) = C(y)) \qquad (4)$$

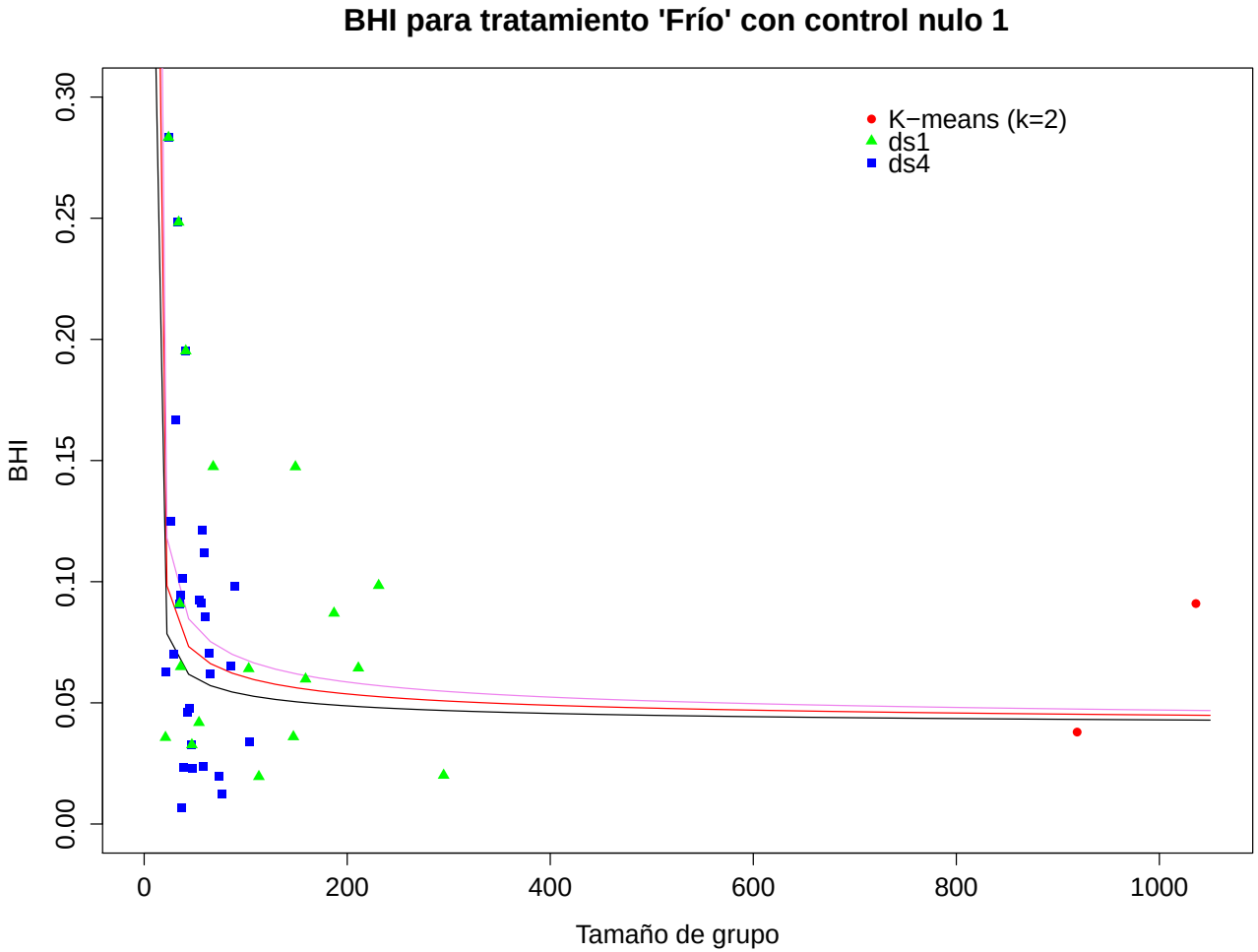
Con n_j la cantidad de genes anotados en el grupo D_j.
La función indicadora $I(C(x) = C(y))$ que toma el valor 1 si hay al menos una clase en donde ambos genes estén anotados, y 0 en caso contrario.

Densidad de interacción

1. Términos mas específicos presentan mayor ID en una relación decreciente.
2. DS1 presenta mayor congruencia biológica que DS4. Indicio acerca de la escala apropiada.
3. Ambos presentan mayor congruencia biológica que control nulo.
4. Los agrupamientos inducidos por otra información presentan mayor congruencia que los inducidos por expresión.



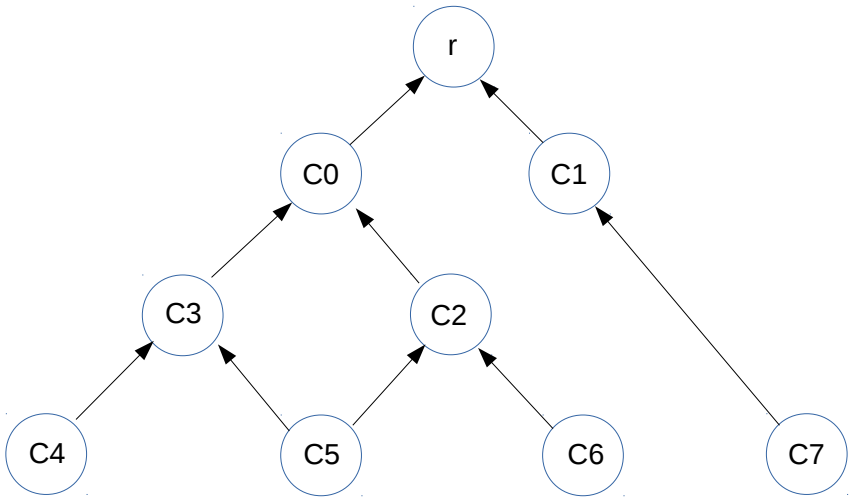
Indice de homogeneidad biológica



Grupos altamente coherentes pero de baja calidad de BHI. No tienen soporte biológico.

Similaridad entre genes en GO

Podemos definir similaridades entre genes en el espacio GO



Utilizando la similaridad entre términos:

$$Sim_{res}(c_i, c_j) = \max_{c \in S(c_i, c_j)} (-log_2[P(c)]) = IC(MICA[c_i, c_j]) \qquad (5)$$

KTA global

La noción de similaridad de a pares en cada espacio esta dada en términos de una función k llamada kernel tal que

$$K = K_{ij} = k(x_i, x_j) \tag{6}$$

El KTA de un kernel k_1 con respecto a un kernel k_2 del conjunto C

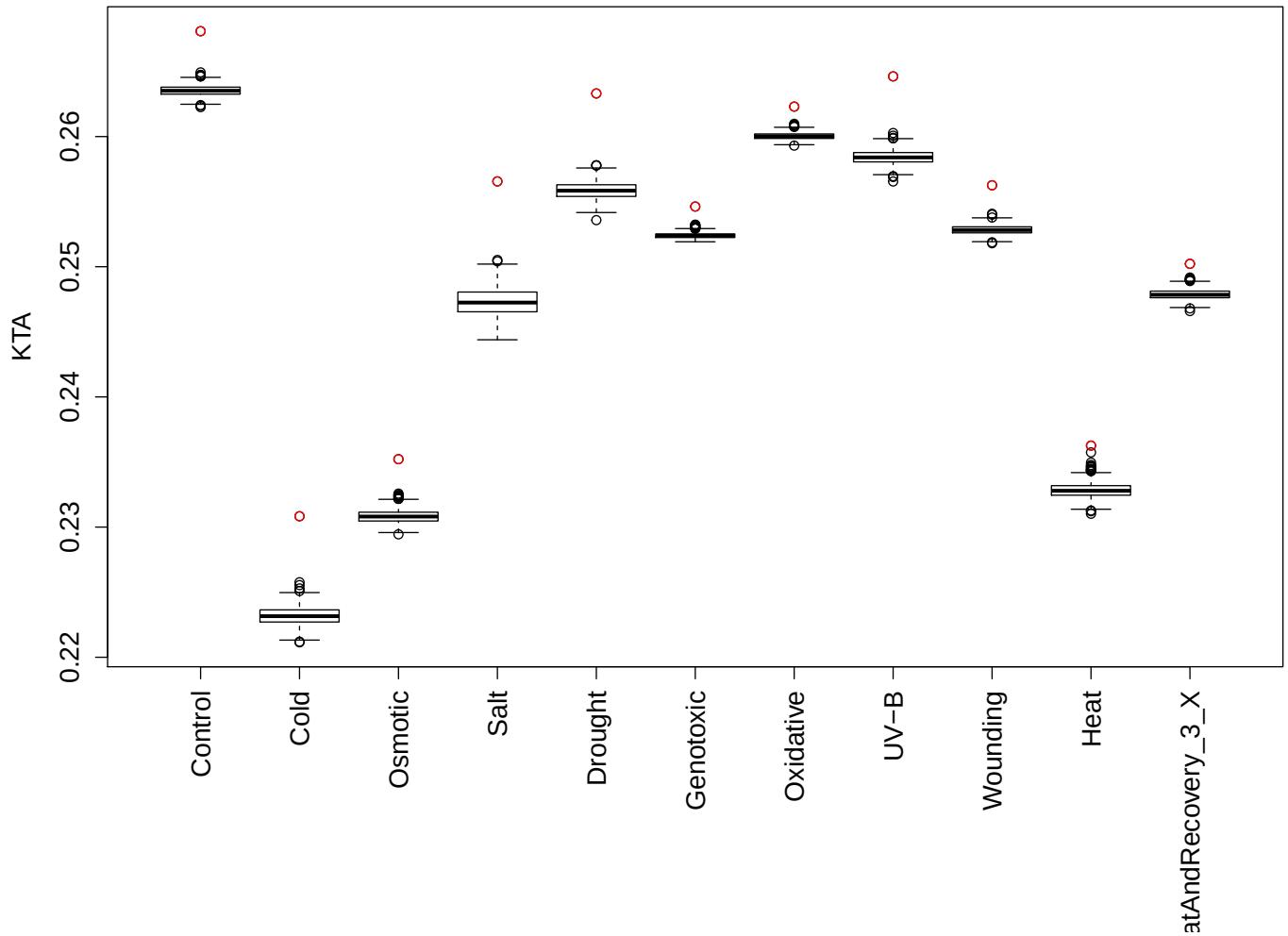
cuantifica la similaridad entre dos espacios y se define como:

$$\hat{A}(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{7}$$

con $\langle K_1, K_1 \rangle_F = \sum_{i,j=1}^{m} K_1(x_i, x_j) K_2(x_i, x_j)$ es el producto interno de Frobenius.
Intiutivamente, si $\langle K_1, K_1 \rangle$ es grande, ambos kernels son coherentes.

KTA global

KTA global entre expresión y ontología BPB con control nulo

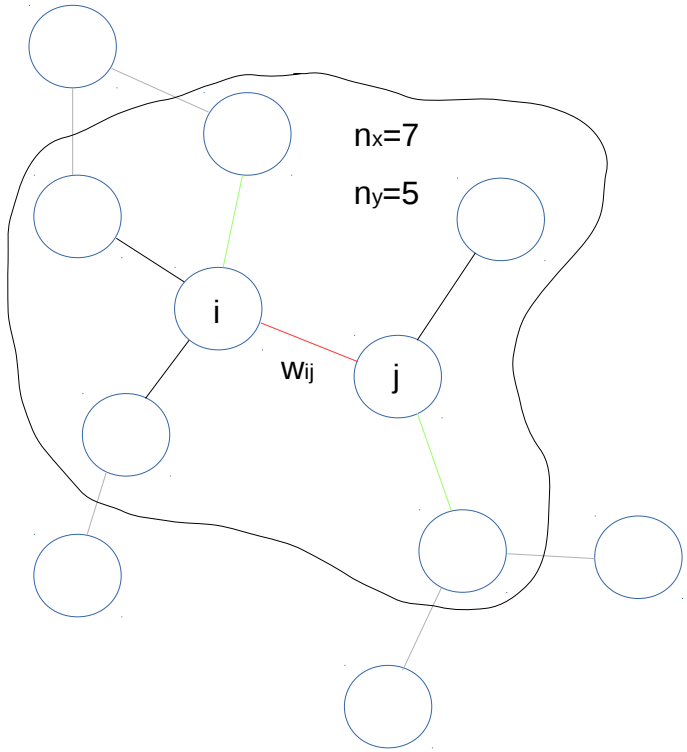


Red 30 primeros vecinos mutuos - vecindades locales

Queremos detectar zonas de alta coherencia.

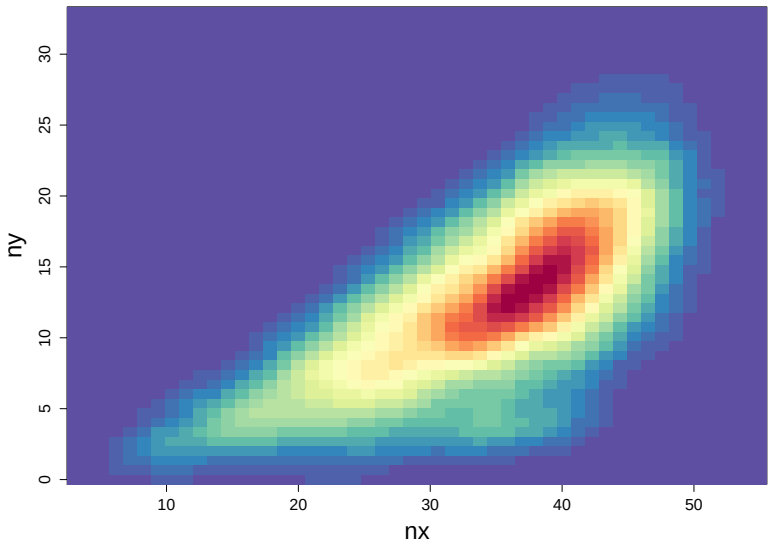
Generamos una red de 30 primeros vecinos mutuos y vamos a ver arista por arista, una localidad definida por los primeros vecinos:

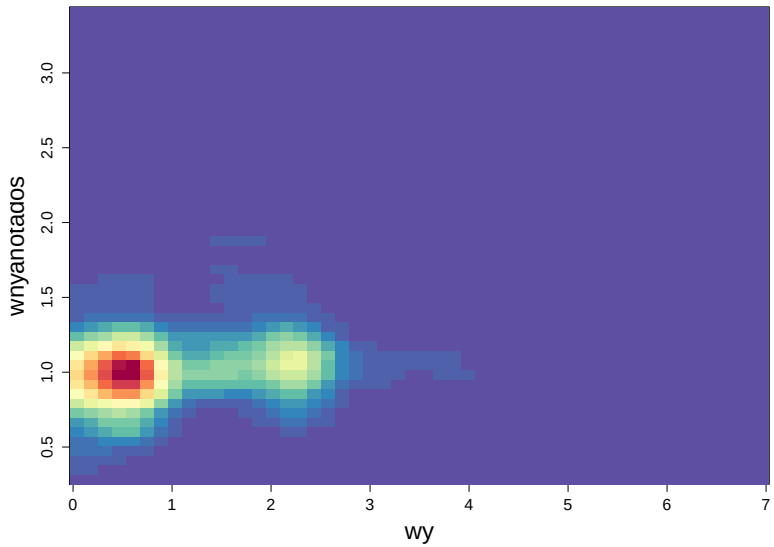
- n_x nodos.
- n_y nodos anotados.
- wyn promedio de pesos de aristas en GO.
- $wyn_{anotados}$ promedio de pesos de aristas en GO con nodos anotados.

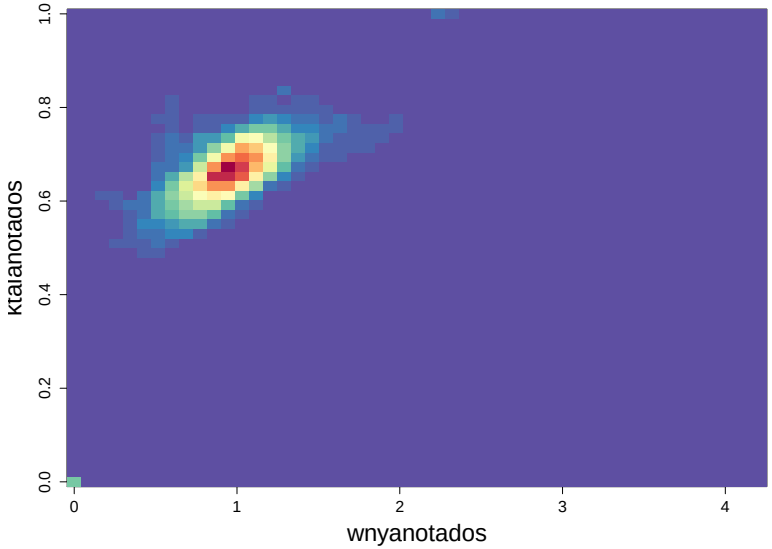


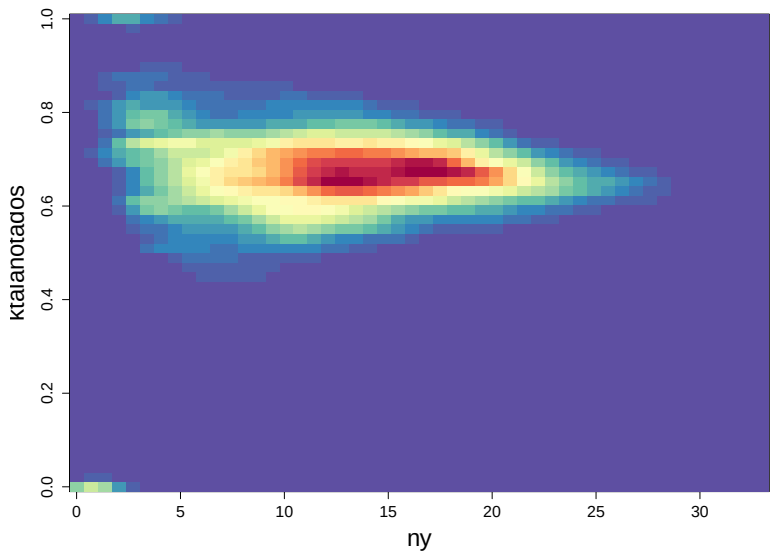
A modo de ejemplo, la red para tratamiento “Frío” consta de 1951 nodos y 18436 aristas.

Caracterización de vecindades locales tratamiento “Frío”









Métrica mixta

Dada una arista, el peso de una arista y el promedio de pesos, tenemos una manera de decir cuando una vecindad es o no biologicamente coherente.

Vamos a usar esto para encontrar grupos transcripcionales teniendo en cuenta las coherencias biológicas locales modificando los pesos:

$$w_{ij} = simcor_{ij}^{\beta * stress_{ij}} \tag{8}$$

Donde:

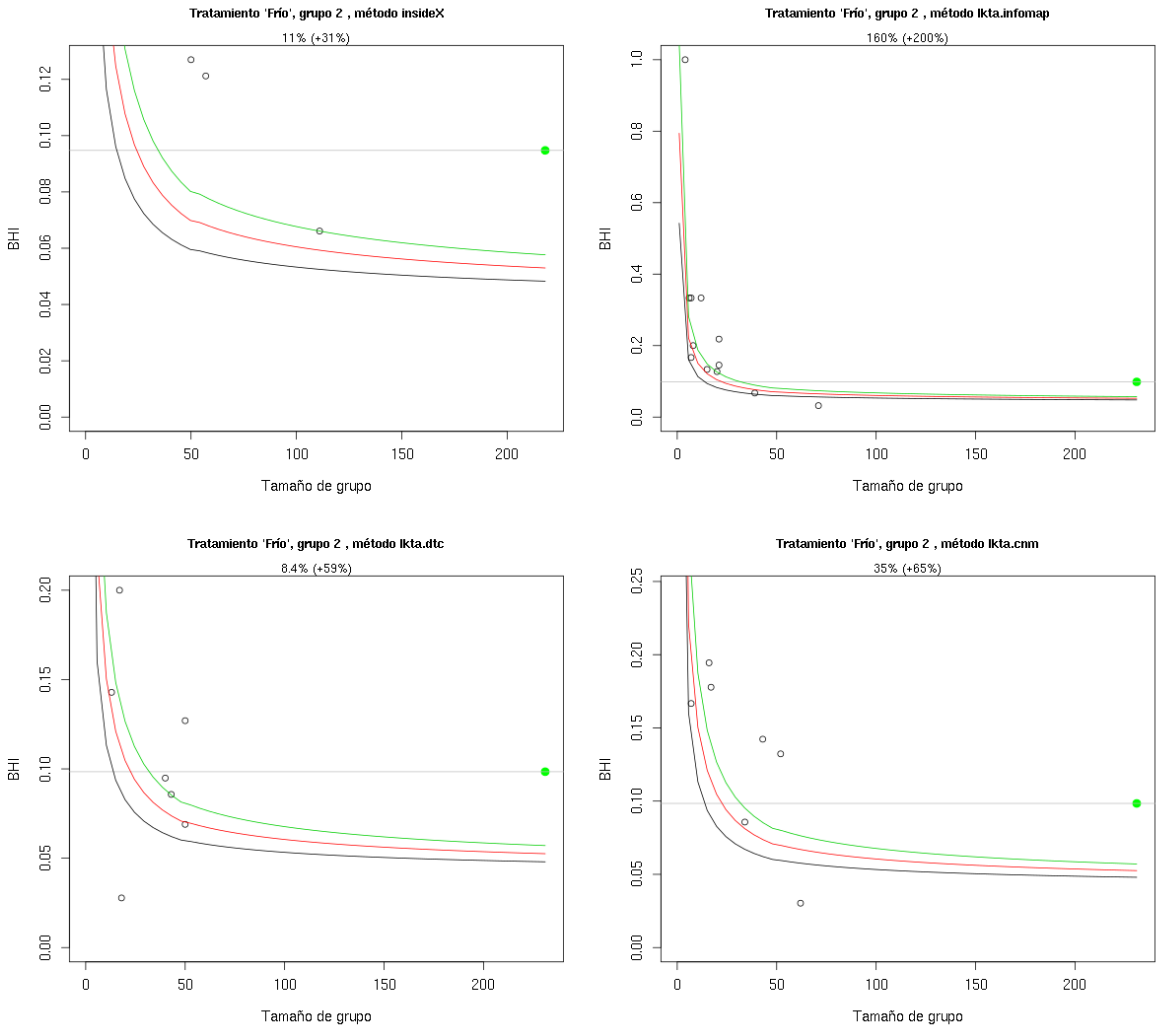
$$stress_{ij} = \frac{KTA_{fondo}}{KTAl_{ij}} \tag{9}$$

Típicamente el *stress* oscila entre 0,8 y 1,2.

β es un parámetro que permite aumentar aún más la homogeneidad de la red.

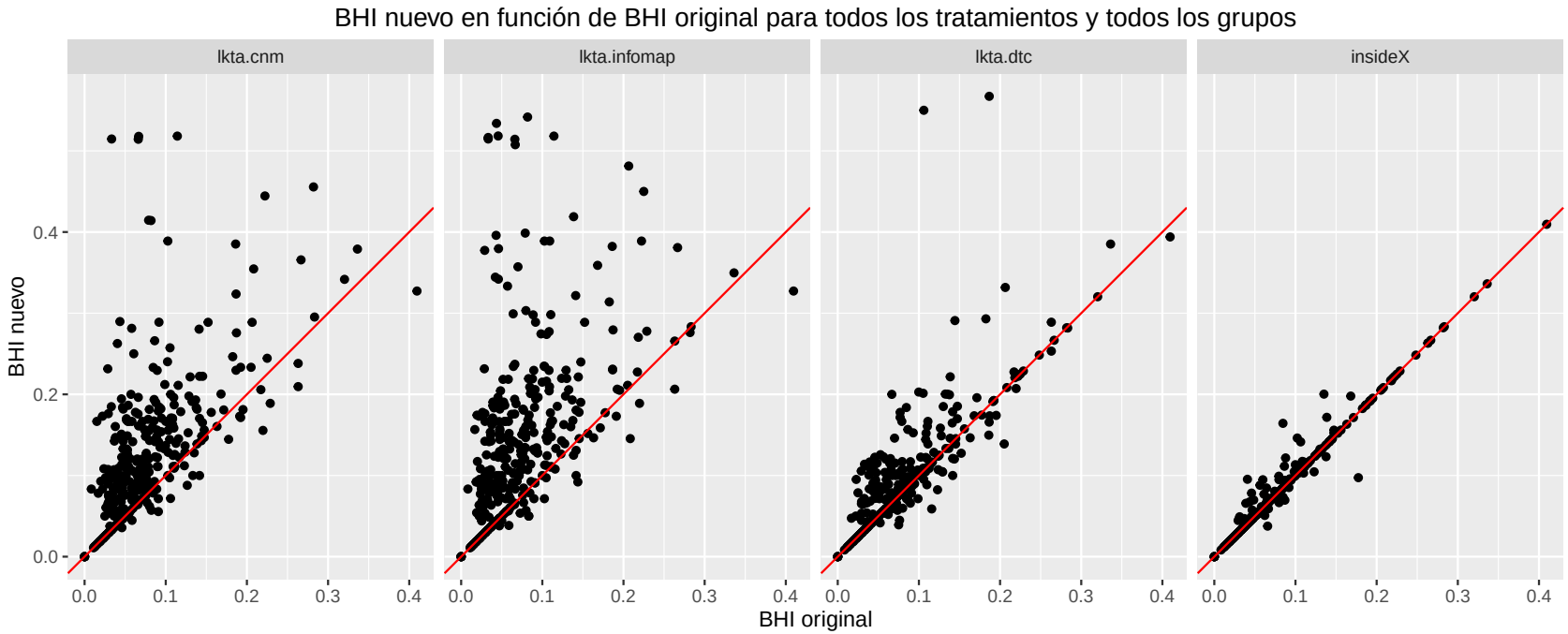
Métodos heurísticos

Buscamos subestructura en los grupos a partir de la métrica mixta

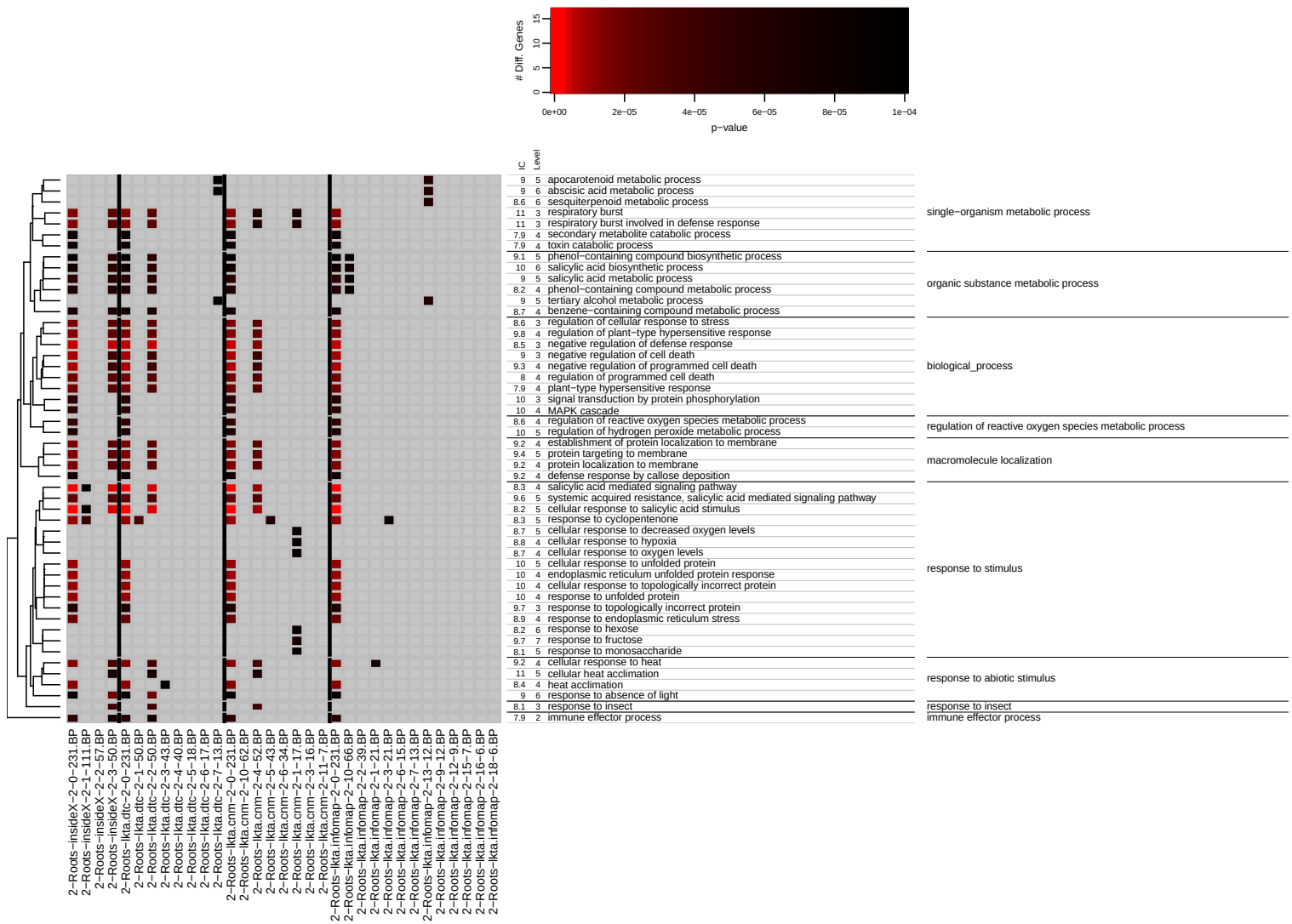


Métodos heurísticos - caracterización de particiones

Caracterizamos los nuevos subgrupos hallados



Interpretación biológica



Conclusiones y perspectivas

- Mediante técnicas de agrupamiento de datos fue posible encontrar grupos de genes con perfiles de expresión altamente correlacionados.
- Distintos métodos darán distintas particiones en función de la resolución que logran.
- Mediante una métrica mixta fue posible encontrar particiones con alta homogeneidad biológica y con alta correlación transcripcional.
- Utilizamos la ontología GO para dar una interpretación biológica a los grupos obtenidos y encontramos que en general, la granularidad óptima de los grupos fue de ≈ 50 genes.
- Estas técnicas podrían funcionar como punto de partida para inferir funciones biológicas de genes de los que se tiene poco conocimiento.
- Sería interesante en un futuro agregar la información contenida en otros espacios de conocimiento biológico, como ser vías metabólicas o redes de interacción de proteínas.

Agradecimientos

¡Muchas gracias!
FOTO DEL GRUPO