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

Detección de correlaciones

Queremos encontrar relaciones entre grandes cantidades de datos.

Lo vamos a hacer usando métodos de agrupamiento o "clustering".

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

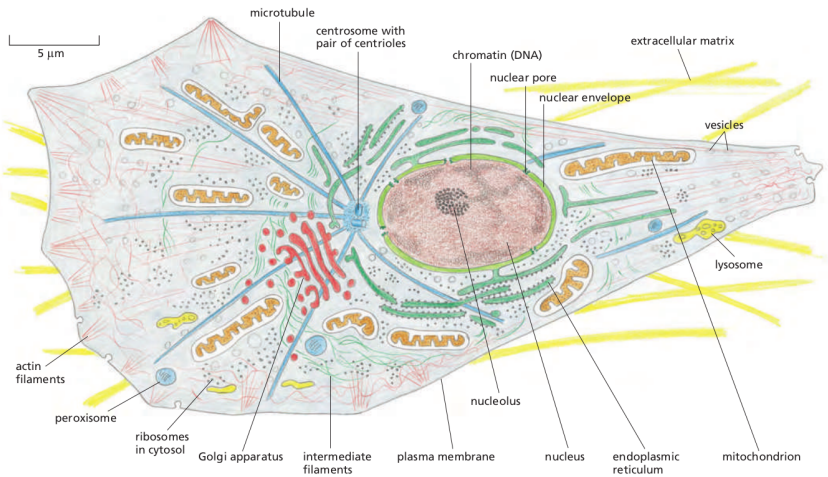
A modo de ejemplo

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

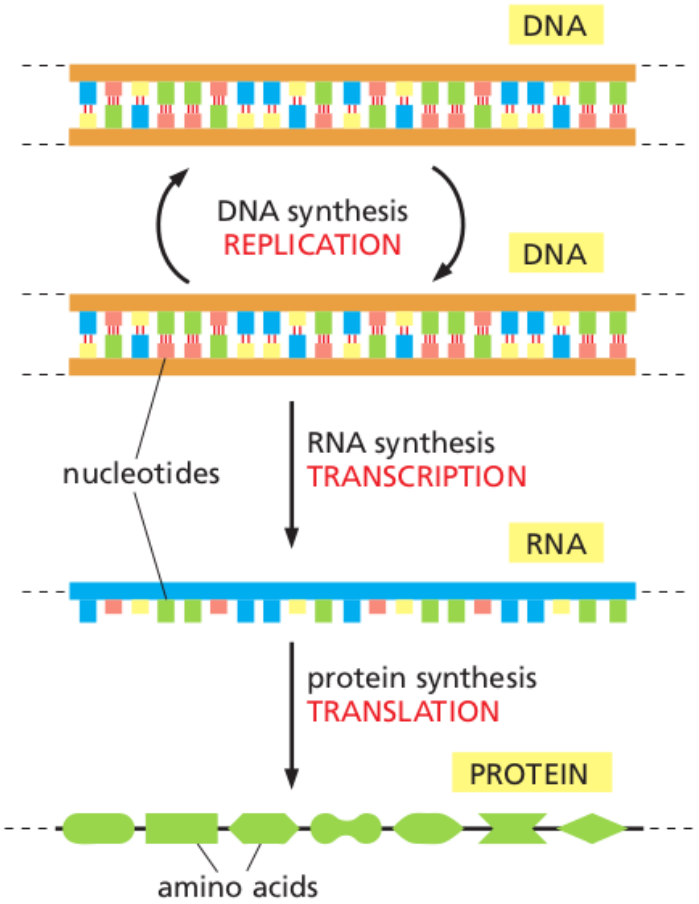
Agrupado por módulo: $\{-5\}$, $\{-3, 3\}$ y $\{-2, 2\}$

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

Transcripción y traducción (dogma central de la biología molecular)



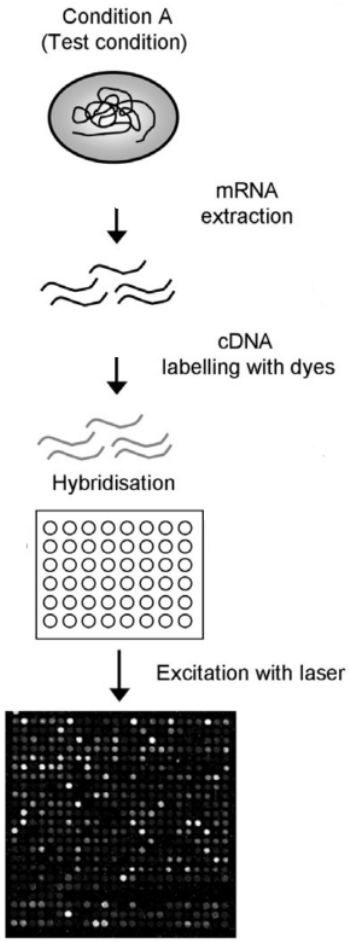
(a) Célula eucariota



(b) Dogma central de la biología molecular

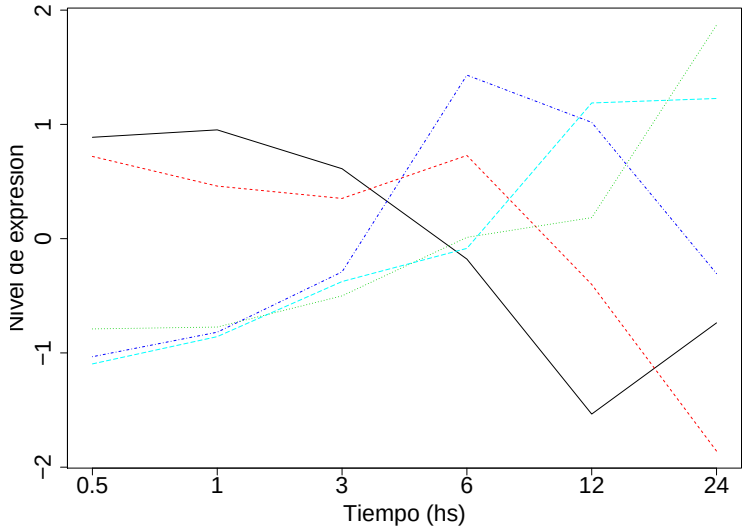
Cambios transcripcionales en respuesta a estrés abiótico en *A. thaliana*

A. Thaliana



Datos de estrés abiótico:

- 11 tratamientos
- $\approx$ 22000 genes
- entre 4 y 8 mediciones temporales por gen y por tratamiento

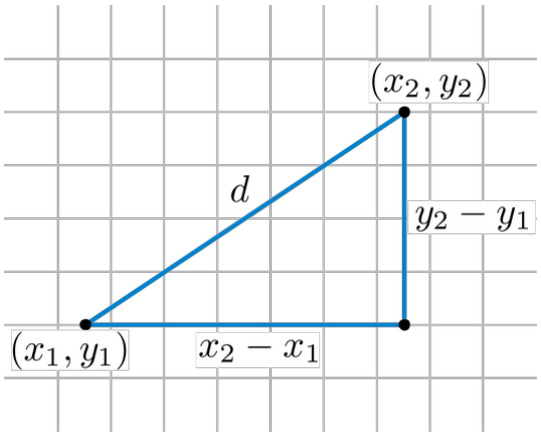


Medidas de similaridad y distancia

Necesitamos definir que significa que dos datos sean "similares"

Distancia euclidiana en espacio de alta dimensionalidad:

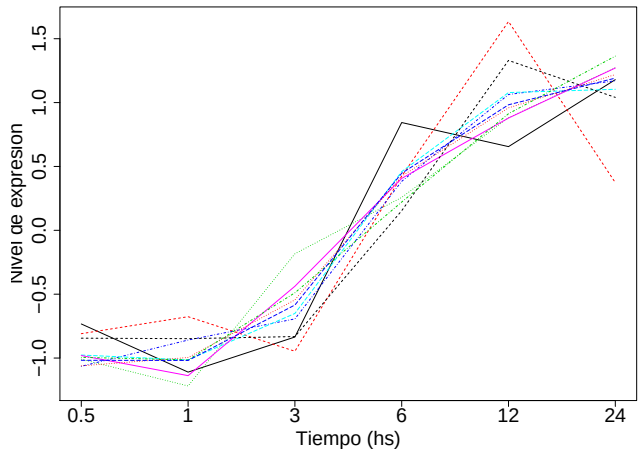
$$d_{euc}(\vec{x},\vec{y}) = [\sum_{i=1}^{n}(x_i - y_i)^2]^{\frac{1}{2}} \quad (1)$$



Distancia basada en el coeficiente de correlación de Pearson:

$$r(\vec{x},\vec{y}) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \quad (2)$$

$$d_{ccp}(\vec{x},\vec{y}) = 1 - r(\vec{x},\vec{y}) \quad (3)$$



Método de agrupamiento k-means

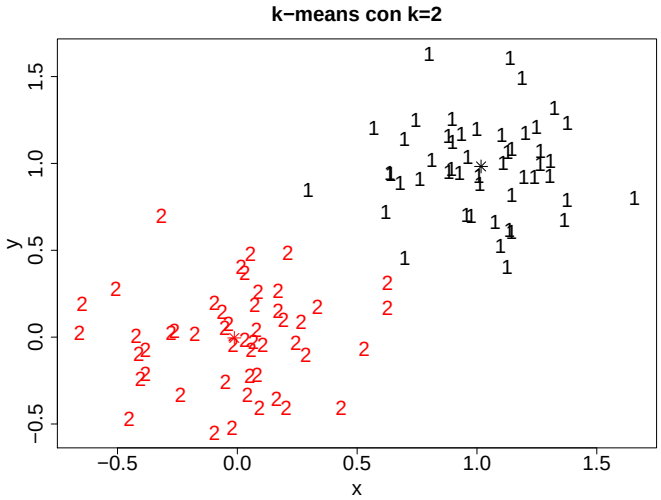
- Agrupamiento no jerárquico.
- Cada observación pertenece al grupo con la media más cercana.
- La cantidad k de grupos debe ser fijada a priori.
- Utiliza la distancia euclidiana.

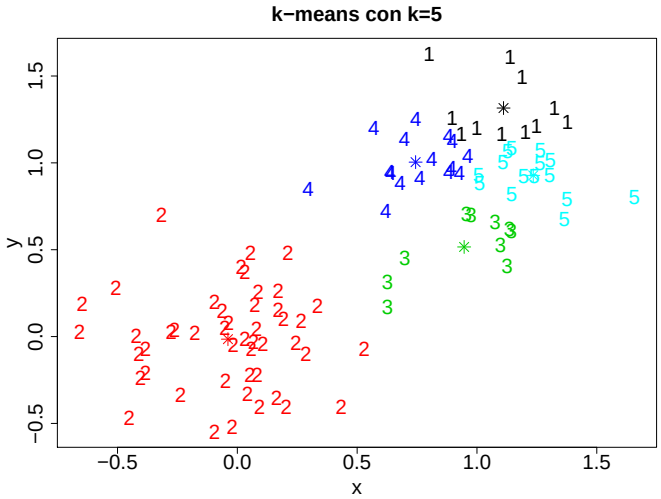
Para datos estandarizados:

$$\tilde{x_i} = \frac{x_i - \bar{x}}{s_x} \qquad (4)$$

la distancia euclidiana se relaciona con la correlación como:

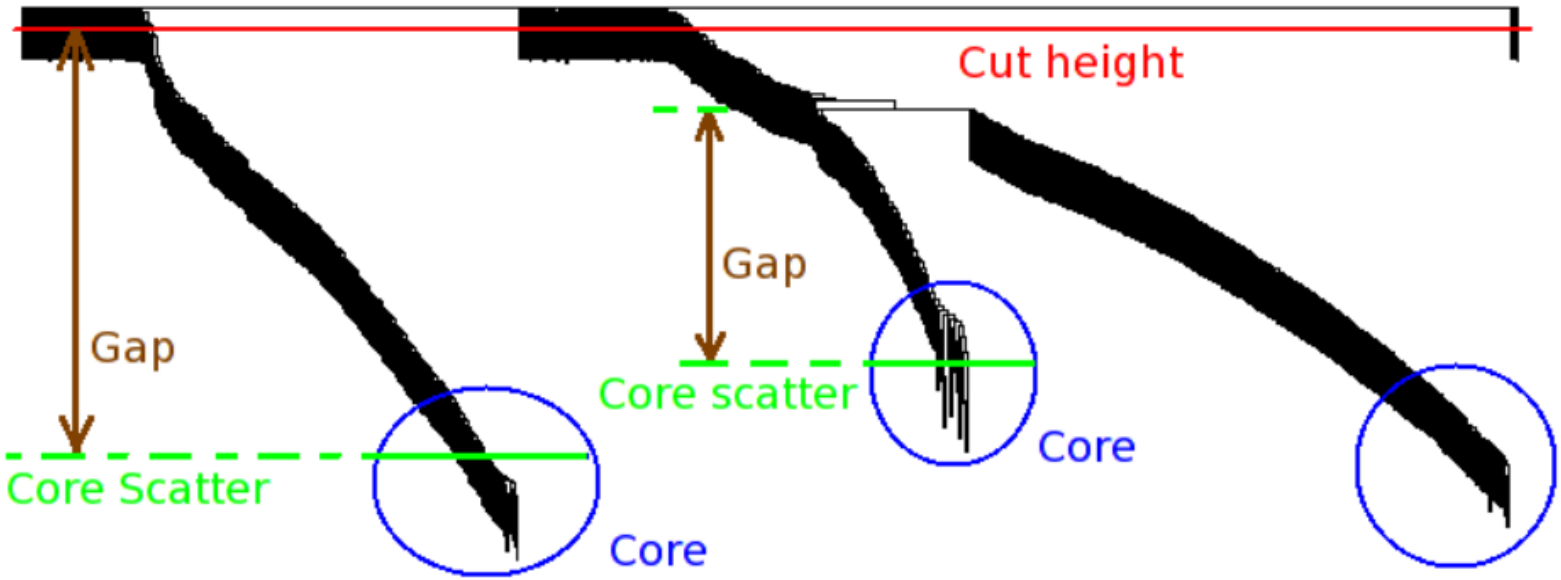
$$d(\vec{x}, \vec{y}) = \sqrt{2(n-1)(1 - r(\vec{x}, \vec{y}))} \qquad (5)$$



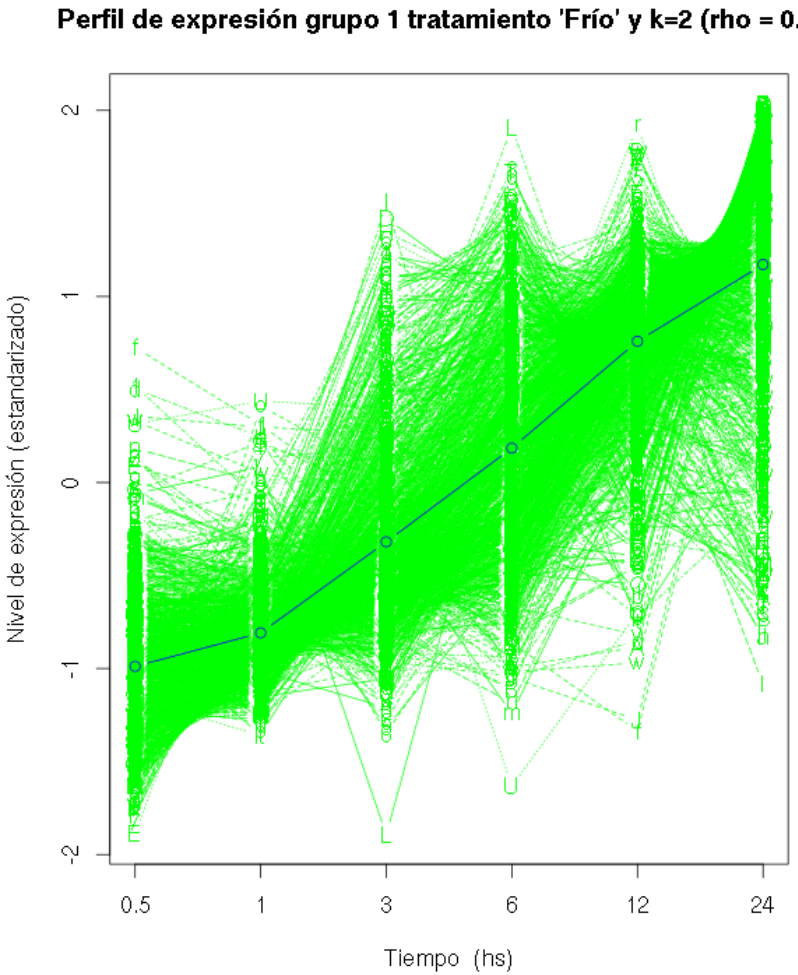


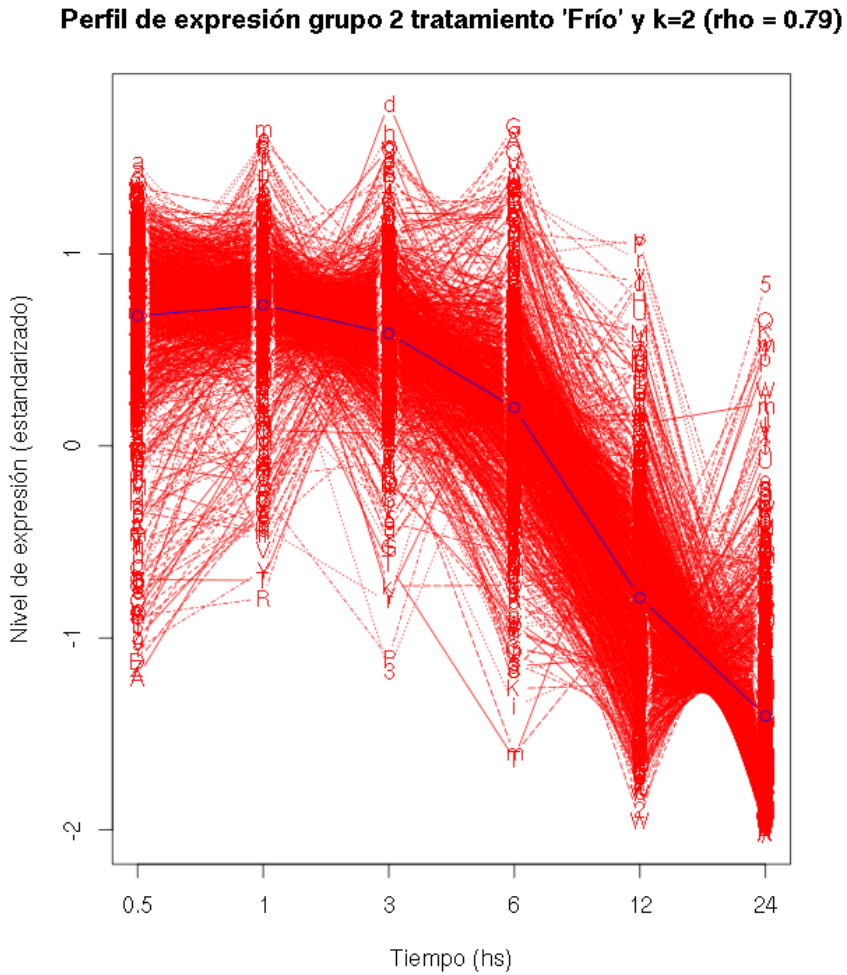
Método de agrupamiento corte de árbol dinámico

- Agrupamiento jerárquico.
- Utiliza la distancia de correlación.
- Se puede "sintonizar" la resolución del método.
- DS1 particiones gruesas, con pocos grupos bien definidos.
- DS4 particiones finas, con muchos grupos más dispersos.



Perfiles tratamiento "Frío" con k-means

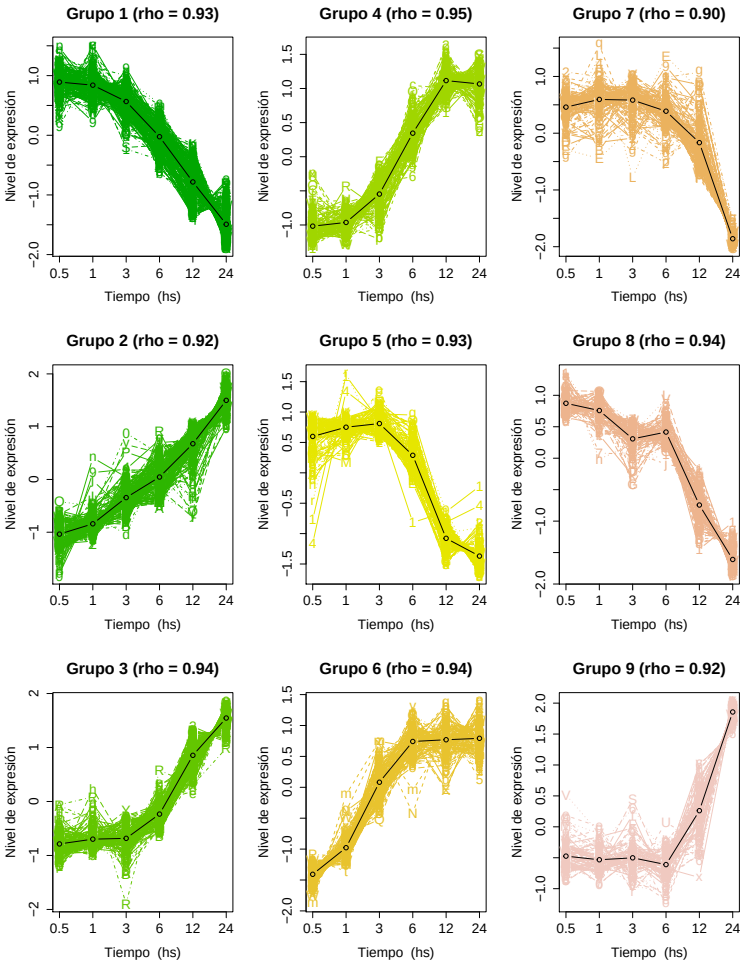




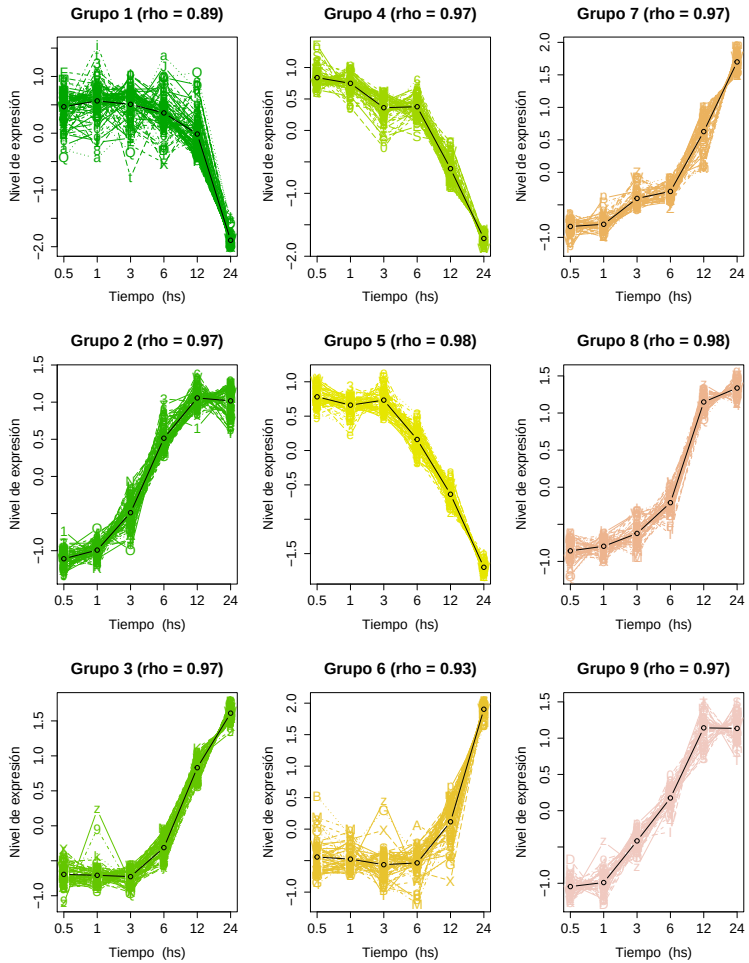
Perfiles tratamiento "Frío" con corte de árbol dinámico

A modo de ejemplo, los nueve perfiles más grandes de cada partición

DS1 (particiones más gruesas)



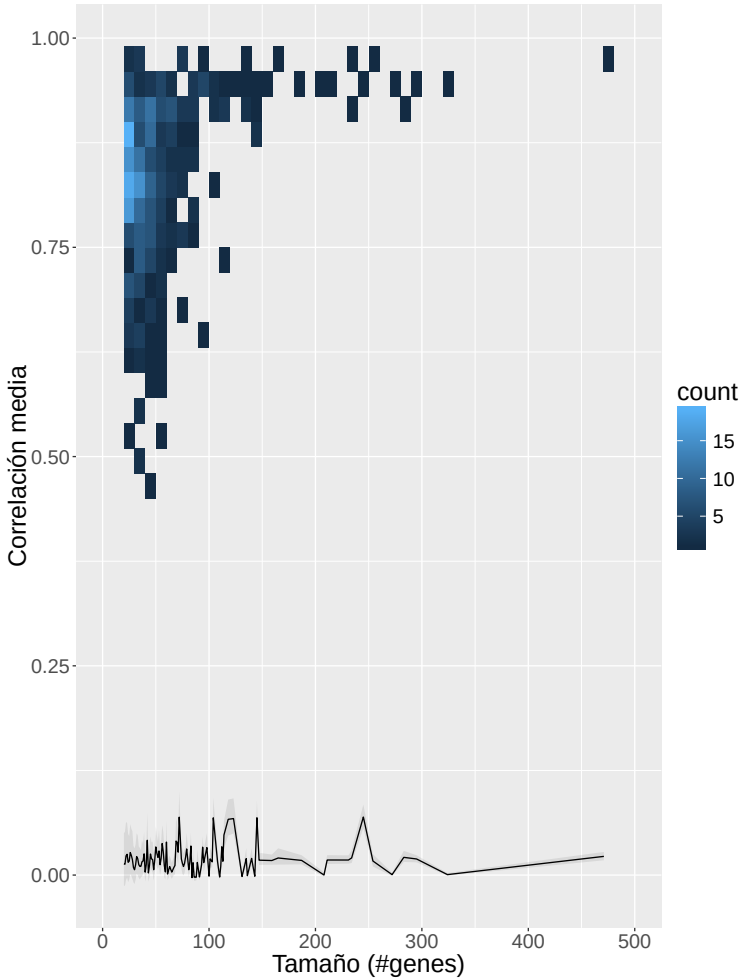
DS4 (particiones más finas)



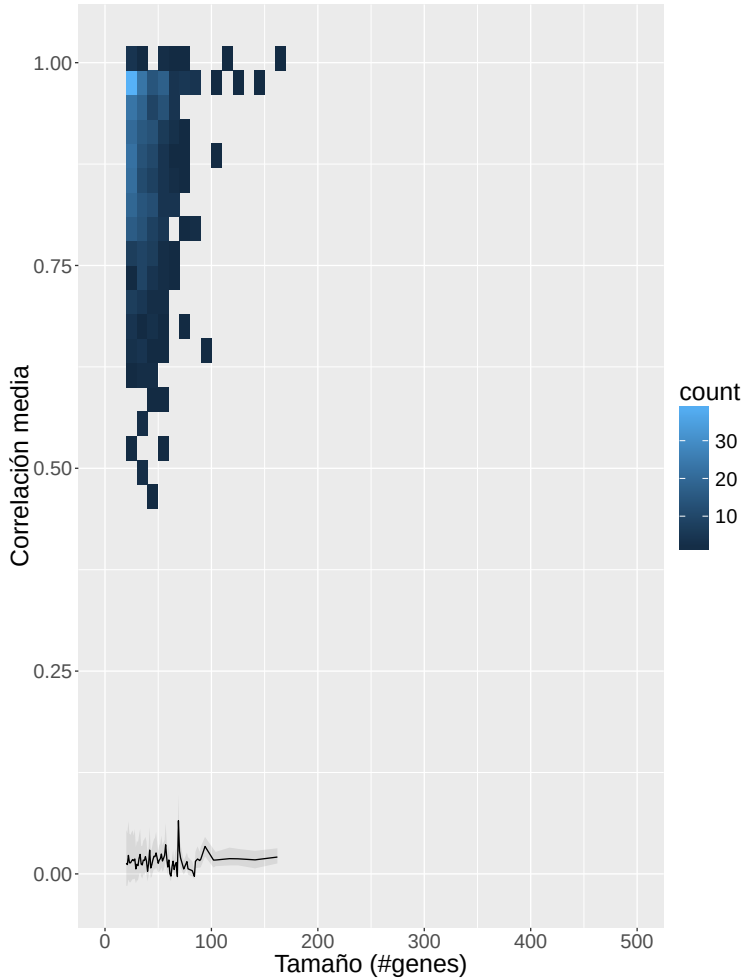
Caracterización de particiones corte de árbol dinámico

Correlación media por tamaño de grupo

DS1 (particiones más gruesas)

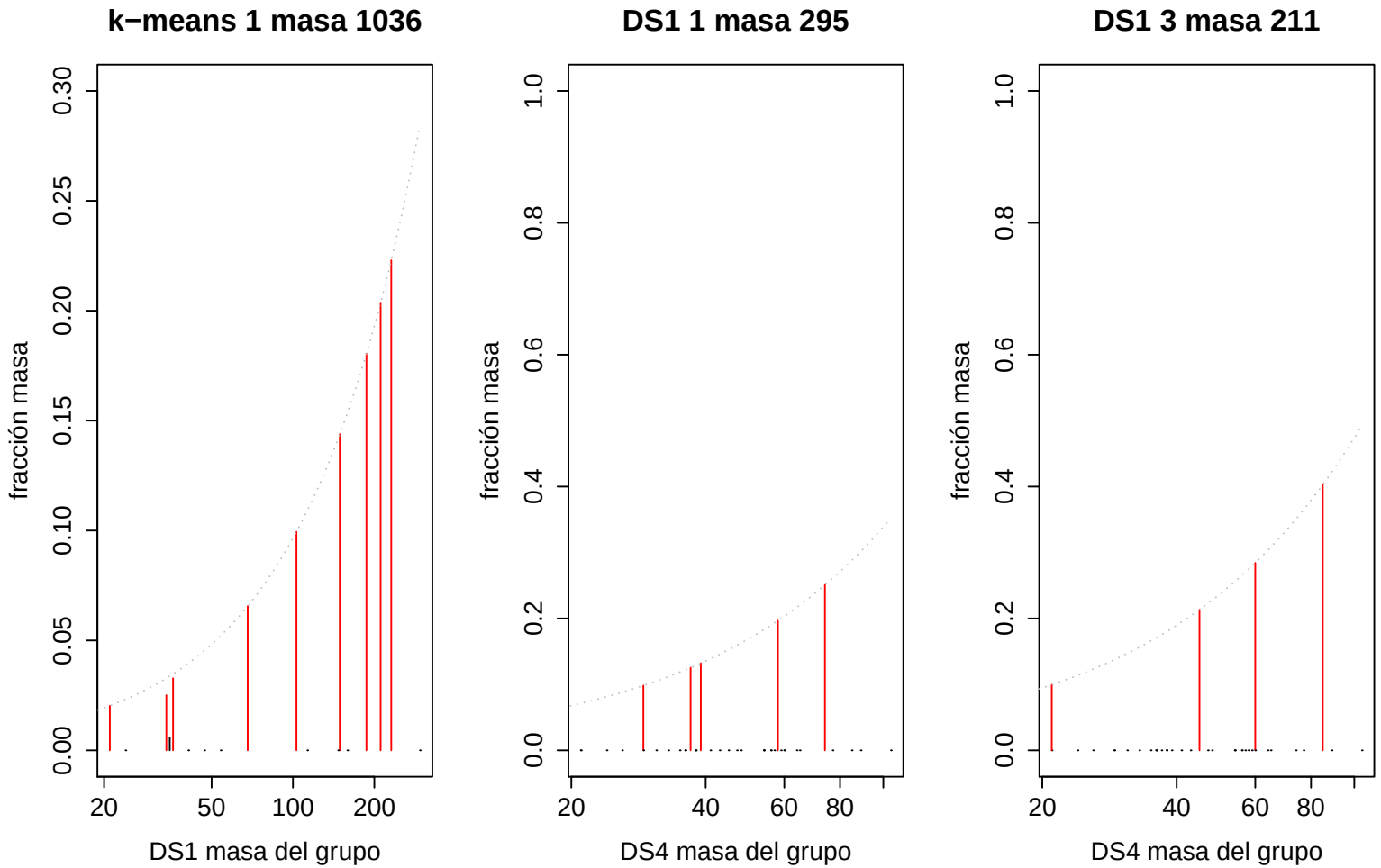


DS4 (particiones más finas)



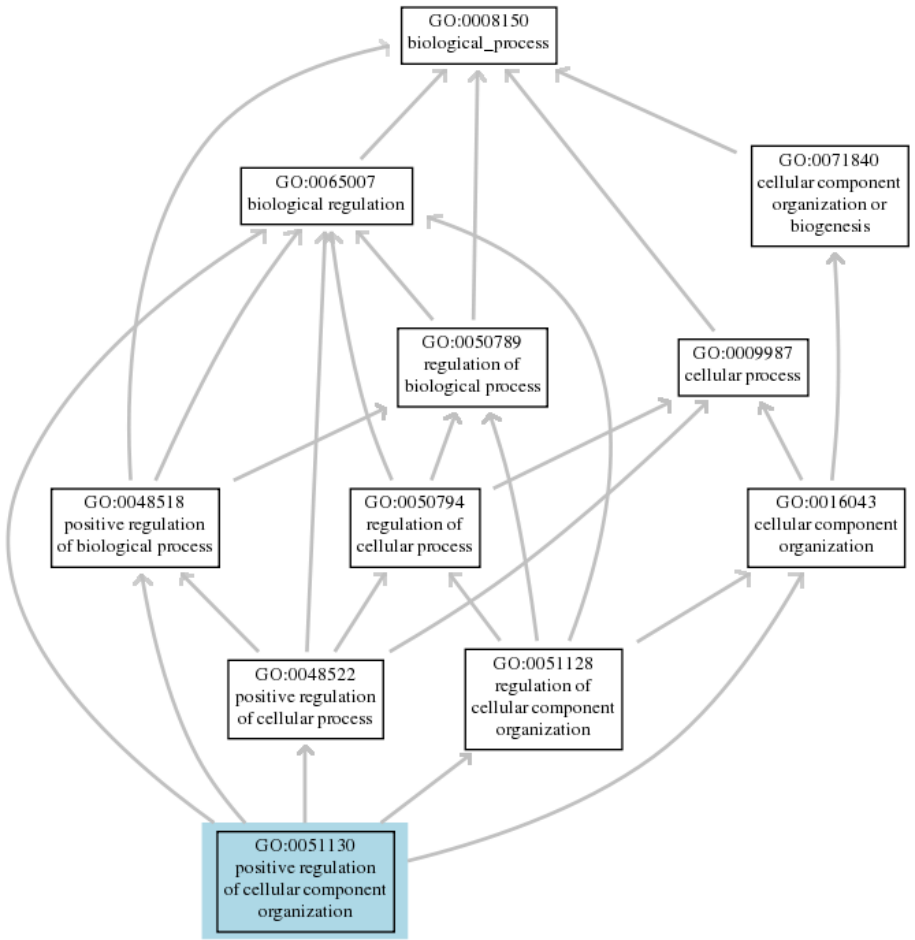
Caracterización de granularidad de las particiones halladas

Fracción de grupos en una partición más fina dentro de grupos de una partición más gruesa (tratamiento "Frío")



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:

$$ID(GO_j) = \frac{NE(GO_j)}{N(GO_j)} \qquad (6)$$

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.

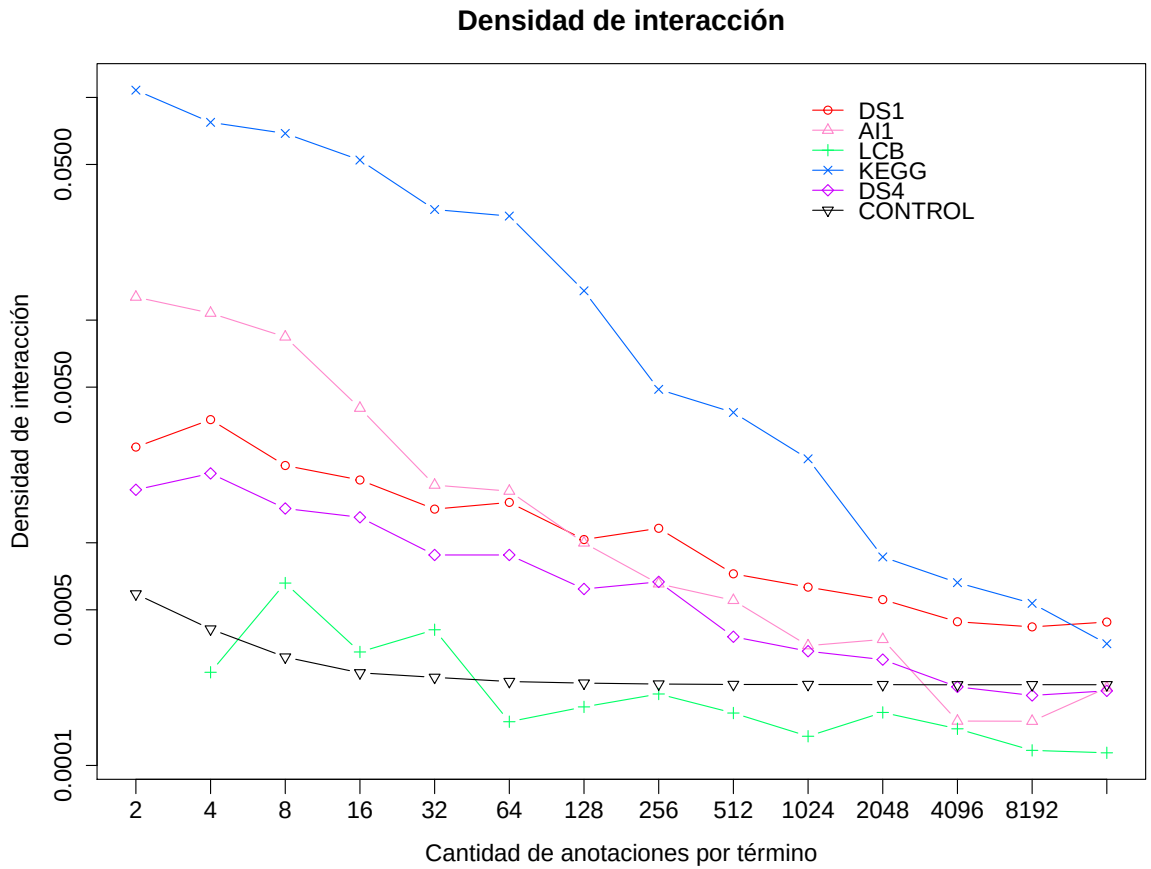
Indice de homogeneidad biológica:

$$BHI_j = \frac{1}{n_j(n_j-1)} \sum_{x \neq y \in D_j} I(C(x) = C(y)) \qquad (7)$$

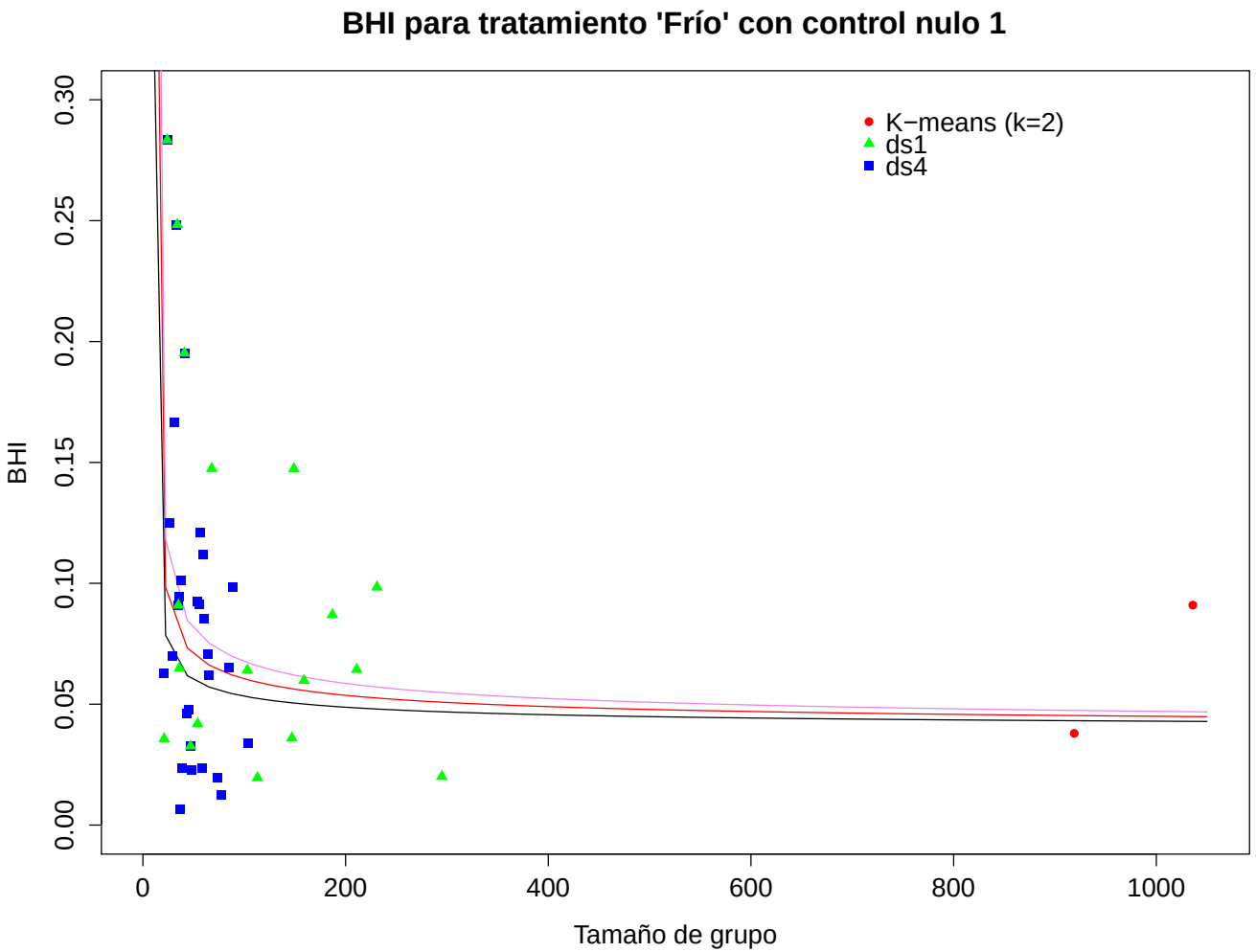
Con n_j la cantidad de genes anotados en el grupo D_j.
La función indicadora $I(C(x) = C(y))$ que toma el valor 1 si hay al menos una clase en donde ambos genes estén anotados, y 0 en caso contrario.

Densidad de interacción

1. Términos mas específicos presentan mayor ID.
2. Agrupamientos hasta 100 genes correlacionan con la información biológica embebida en la ontología.
3. Las estructuras en KEGG presentan mayor congruencia con las ontologías, seguidas por AI1 y expresión.
4. $ds1$ presenta mayor congruencia biológica que $ds4$. Indicio acerca de la escala de granularidad apropiada.



Indice de homogeneidad biológica



Particiones altamente coherentes pero de baja calidad de BHI.

Similaridad entre genes en GO

Definimos la similaridad entre genes en el espacio GO como:

$$Sim_{rcmax}(GO(g_1), GO(g_2)) = \text{máx}\{\frac{1}{N}\sum_i \max_{1\leq j\leq M} S_{ij}, \frac{1}{M}\sum_j \max_{1\leq i\leq N} S_{ij}\} \tag{8}$$

Donde:

$$S(g_1, g_2)_{ij} = Sim_{res}(GO(g_1^i), GO(g_2^j)), \forall i \in \{1, ..., N\} y \forall j \in \{1, ..., M\} \tag{9}$$

con:

$$Sim_{res}(c_i, c_j) = \max_{c\in S(c_i,c_j)}(-log_2[P(c)]) = IC(MICA[c_i, c_j]) \tag{10}$$

la similaridad entre términos.

KTA global

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

$$K = K_{ij} = k(x_i, x_j) \tag{11}$$

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

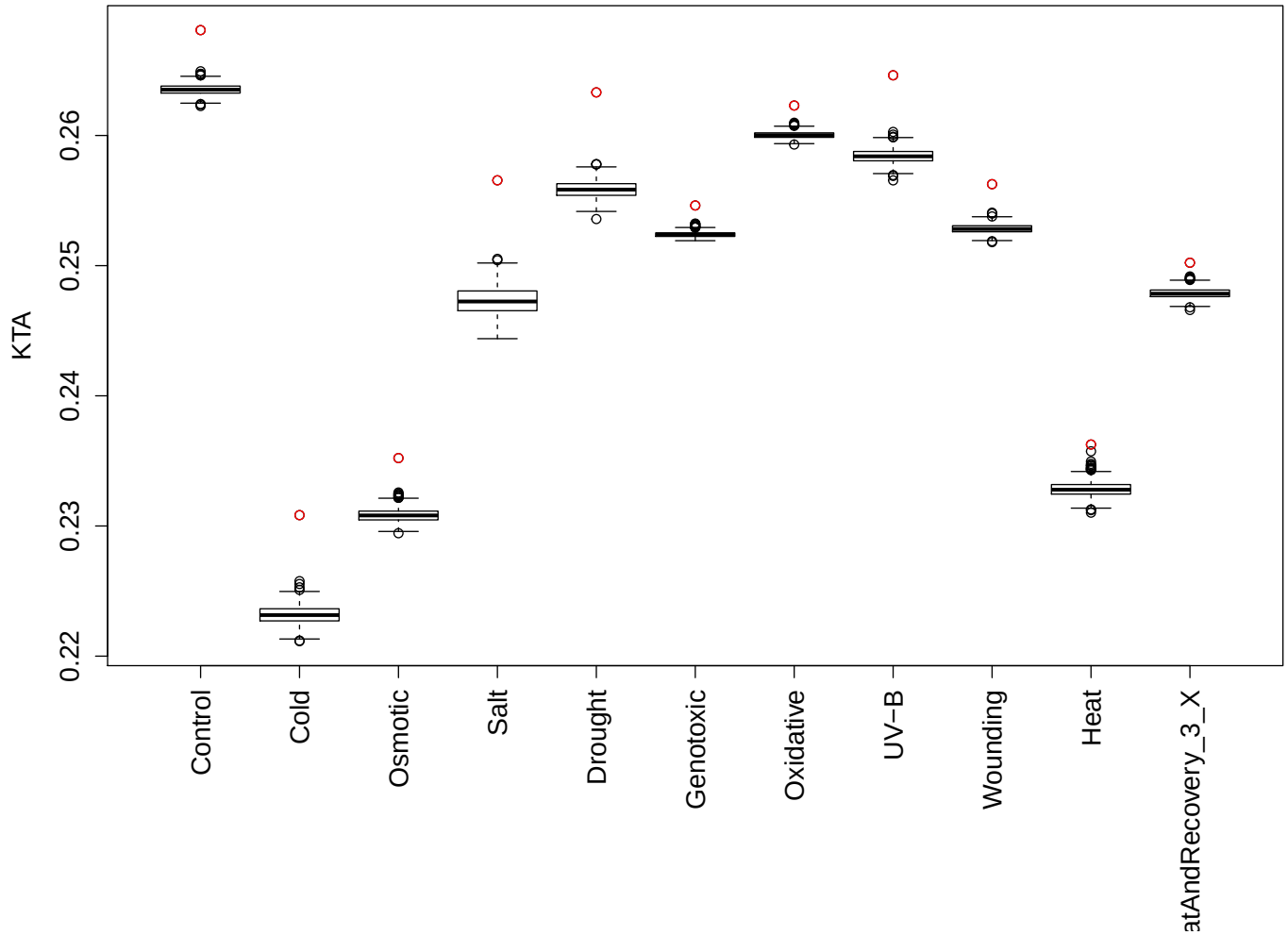
cuantifica la similaridad entre dos espacios y se define como:

$$\hat{A}(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{12}$$

con $\langle K_1, K_1 \rangle_F = \sum_{i,j=1}^{m} K1(x_i, x_j) K2(x_i, x_j)$ es el producto interno de Frobenius.

KTA global

KTA global entre expresión y ontología BP con control nulo



KTA local para modulación de heterogeneidades transcripcionales

Métrica mixta

Método heurístico

Interpretación biológica

Conclusiones y perspectivas