Análisis y Detección de Correlaciones en Relevamientos Transcripcionales de Gran Escala

Andrés Rabinovich
Director: Dr. Ariel Chernomoretz

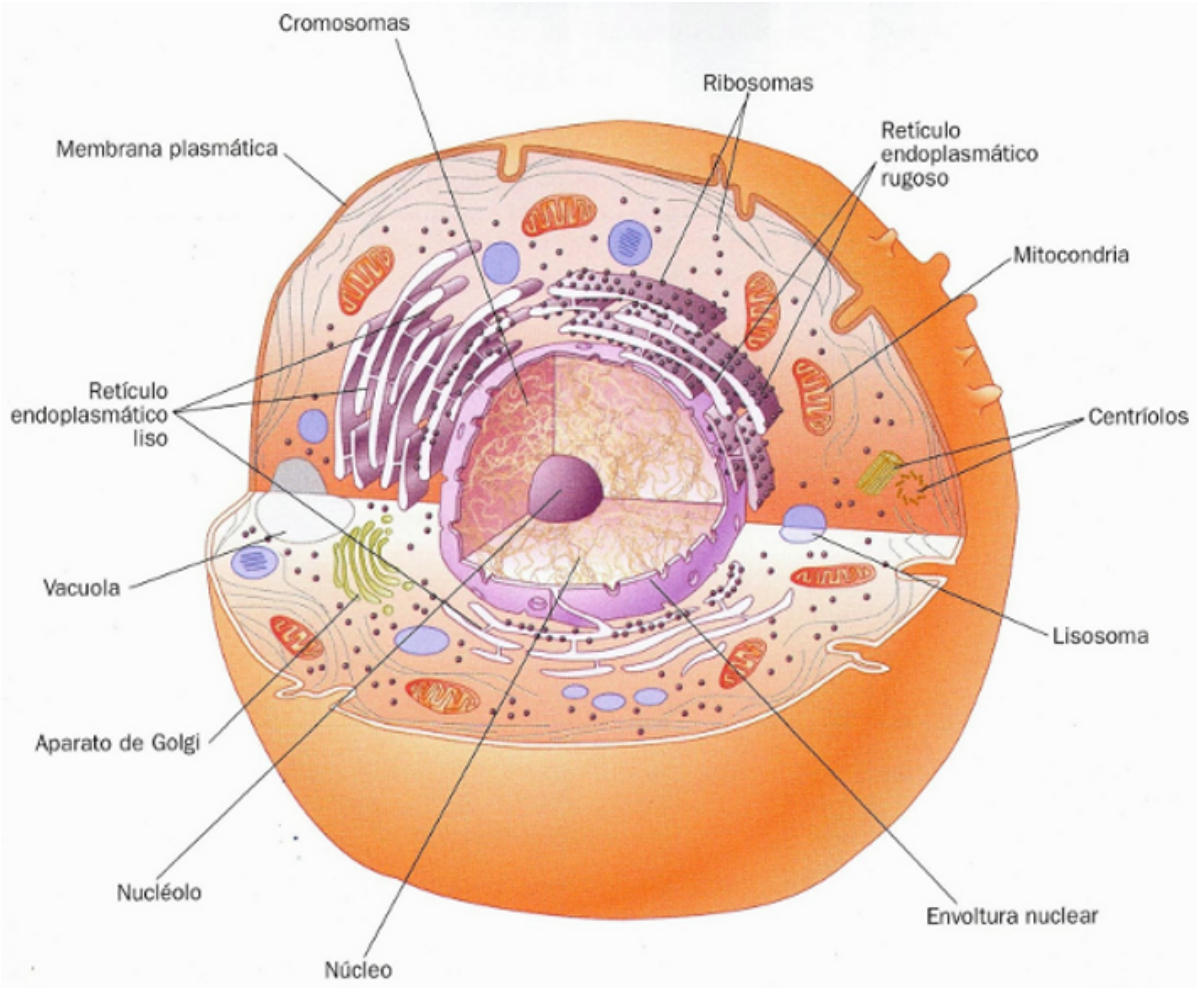
Departamento de Física
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

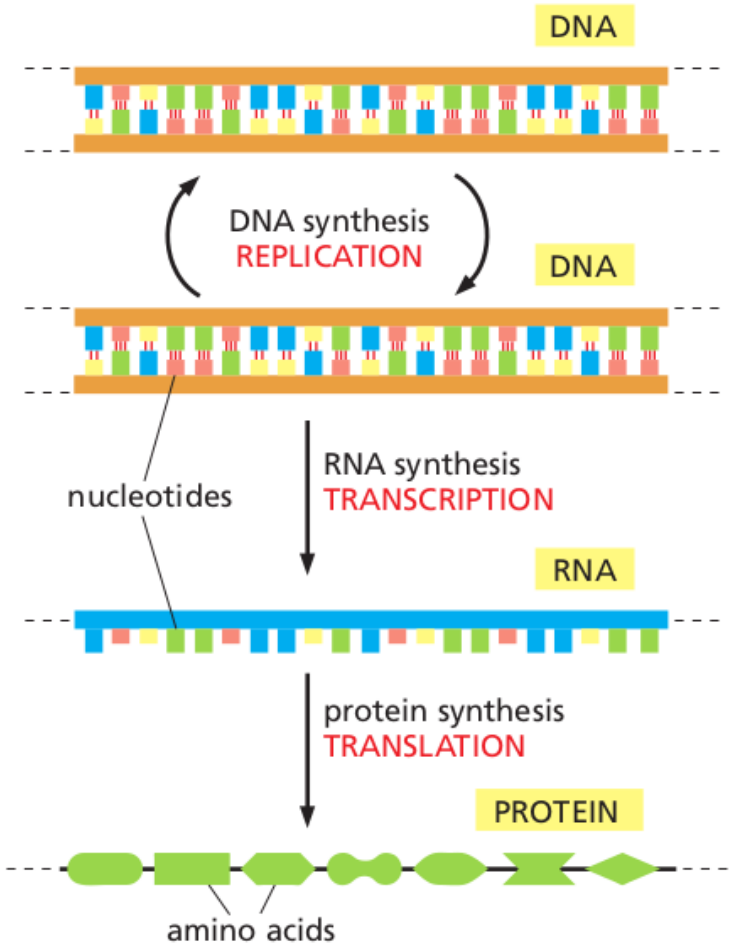
Marzo 2016.

Contenido

Transcripción y traducción (dogma central de la biología molecular)

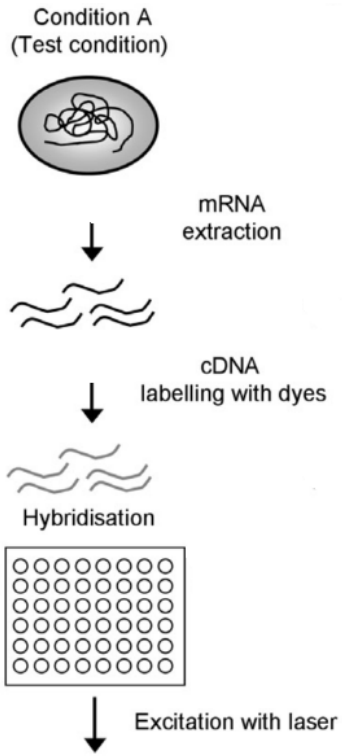
Células, ADN, ARNm, proteínas y otras yerbas...





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

A. Thaliana [1]



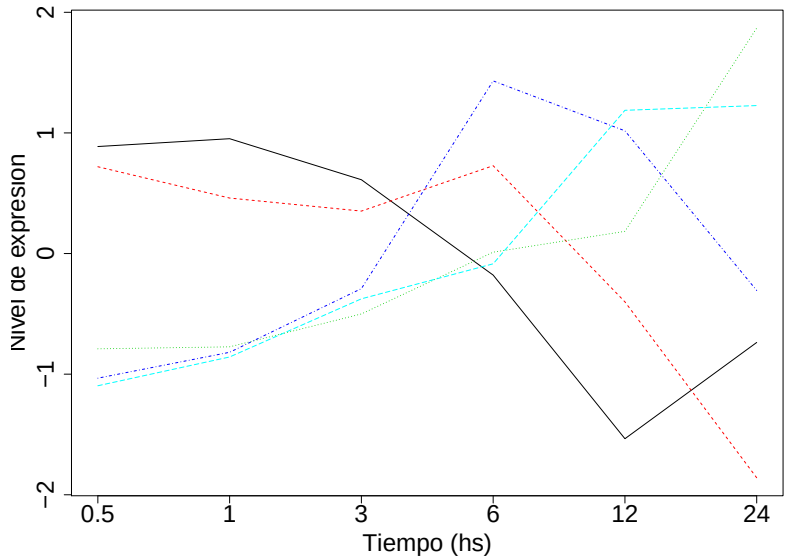
Gen	1 hs.	6 hs.	24 hs.
AT4G09000	0.2	0.8	0.4
AT1G78300	1.2	1.2	0.8
AT5G38480	0.5	0.7	1.2
.	.	.	.
.	.	.	.
.	.	.	.
AT1G35160	1.1	0.9	0.6

Micromatriz de ADN

[1] arabidopsis.org: AtGenExpress.

Datos de estrés abiótico:

- **10 tratamientos + control**: frío, calor, osmótico, salinidad, sequía, genotoxicidad, oxidación, UV, herida, recuperación.
- $\approx$ 22000 genes.
- $\approx$ 6000 genes se **movieron** en algún tratamiento.
- **Entre 4 y 8 mediciones** temporales por gen y por tratamiento.

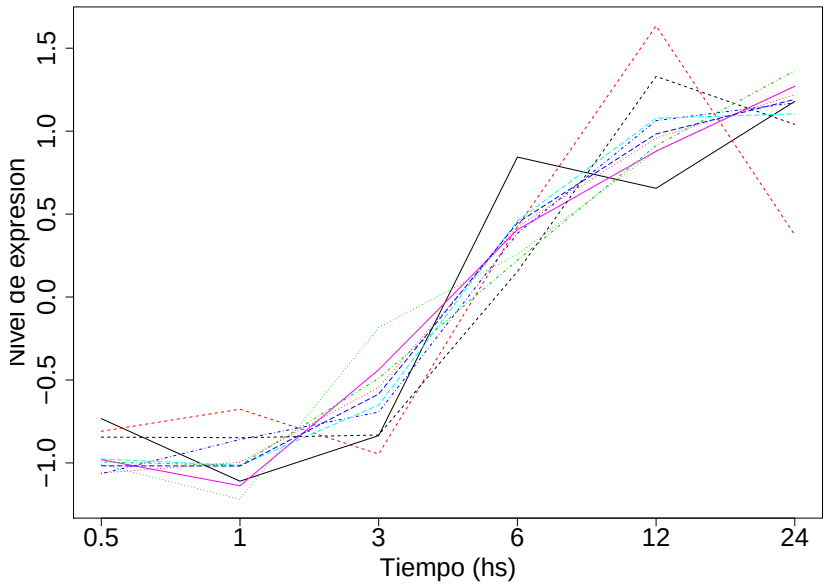


Detección de correlaciones

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 sí"**.
- Permiten el **descubrimiento de patrones** en los datos.
- Posibilitan obtener **conclusiones** sobre los datos.

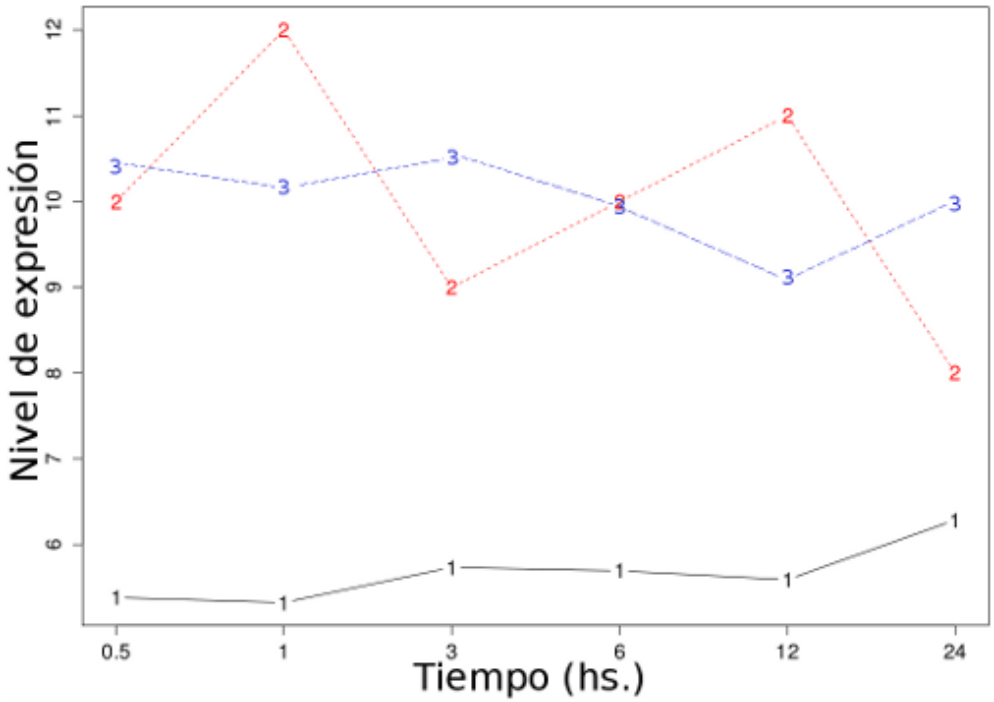


Usamos el coeficiente de correlación de Pearson

Distancia basada en el coeficiente de correlación de Pearson

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

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



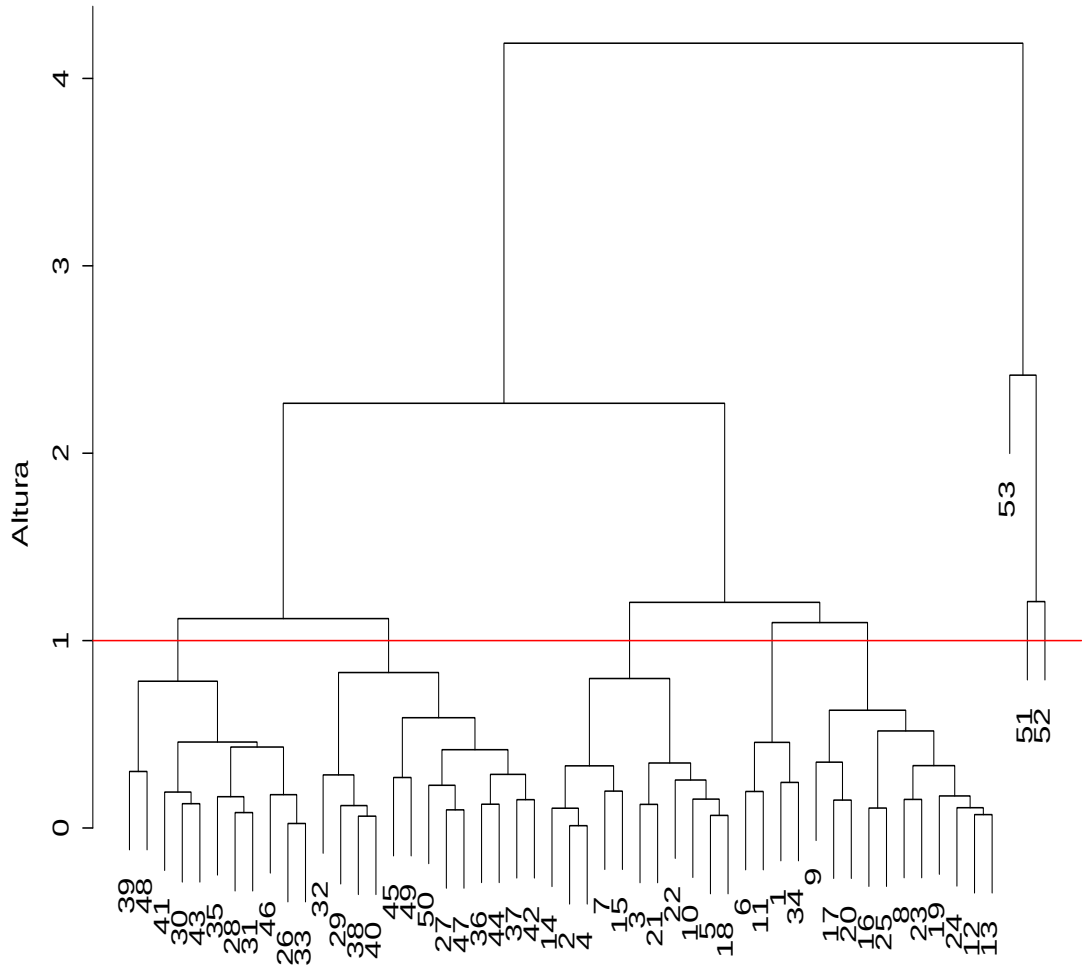
Métodos de agrupamiento

Método k-means

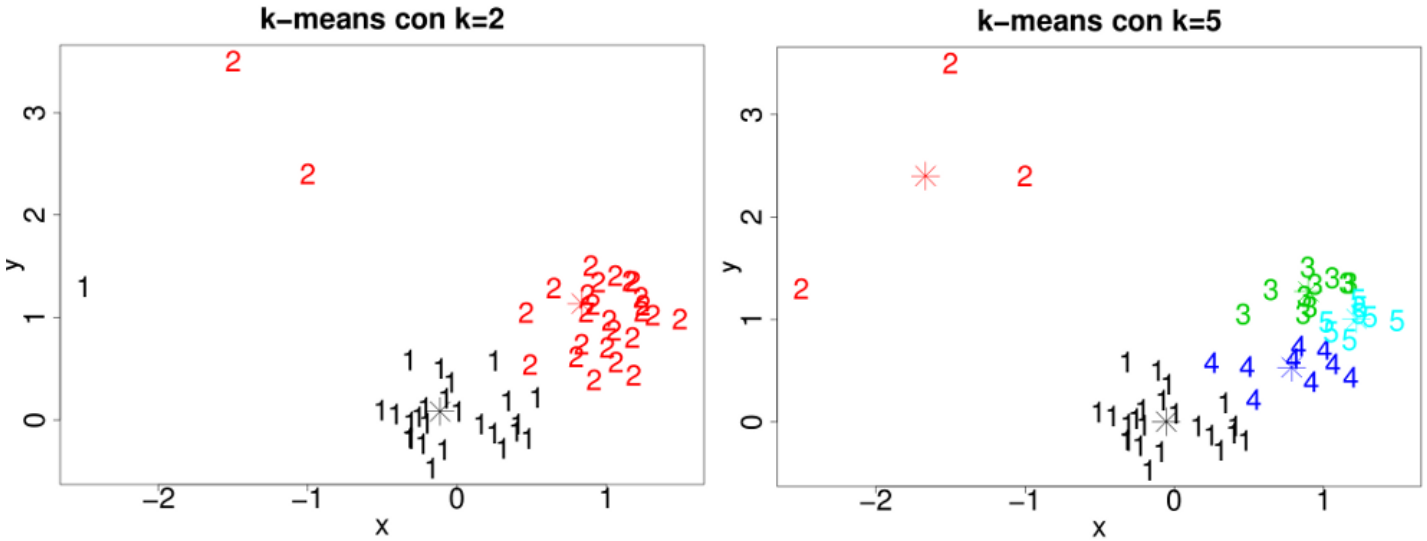
- Busca estructuras compactas.
- Muy rápida ejecución.
- La cantidad k de grupos debe ser fijada a priori.
- Existen figuras de mérito para decidir el k óptimo.

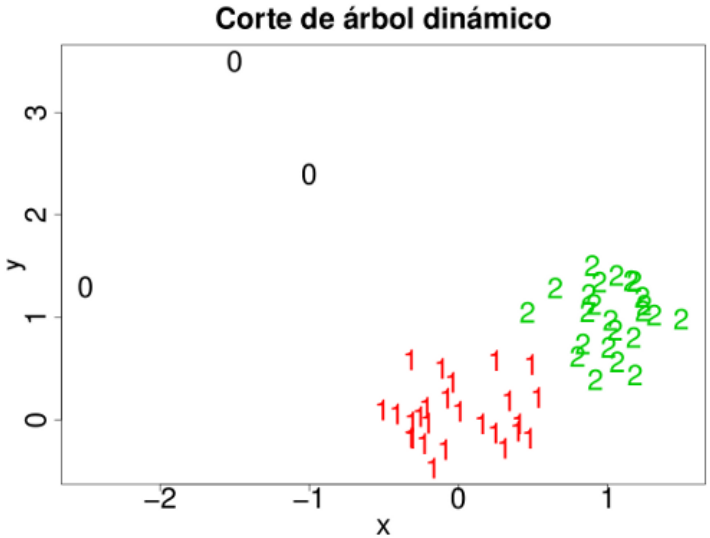
Método corte de árbol dinámico

- Agrupamiento jerárquico.
- Representación mediante **dendrograma**.
- Ajuste por *deepsplit* (usamos DS1 y DS4).

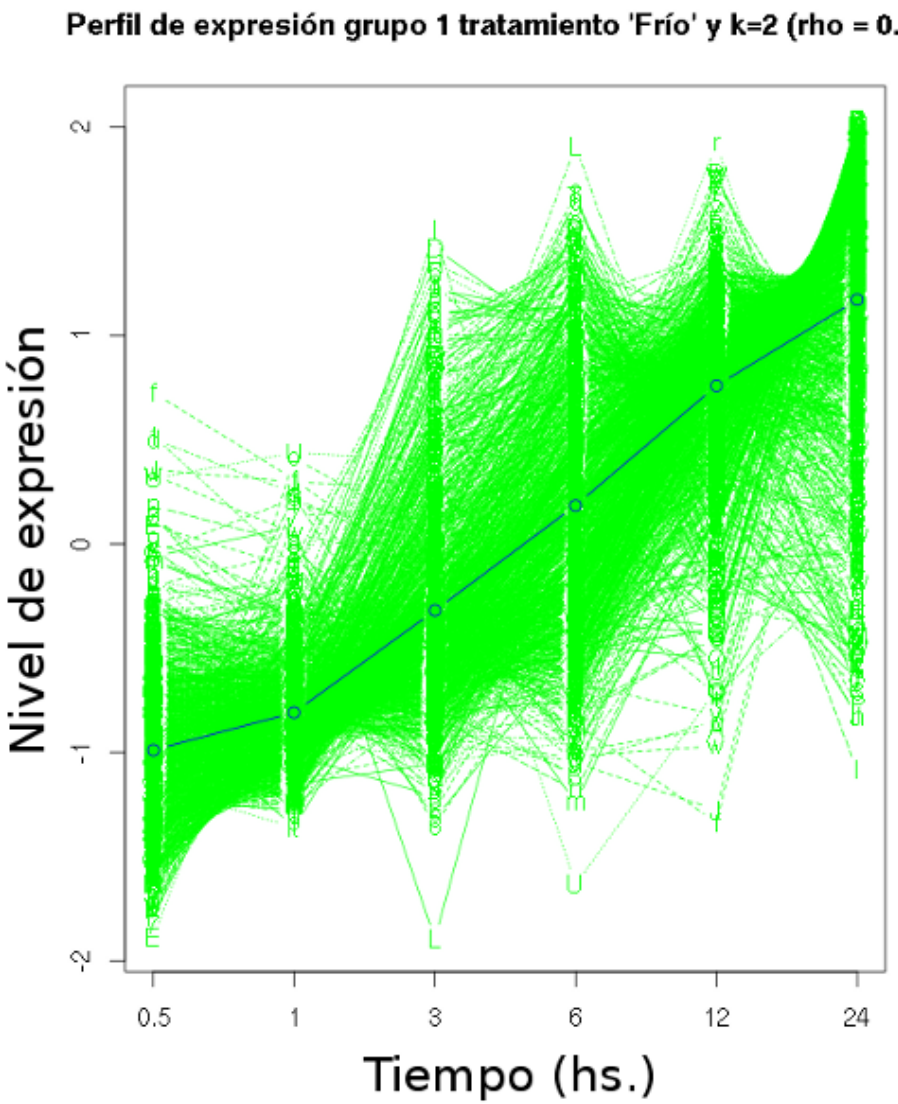


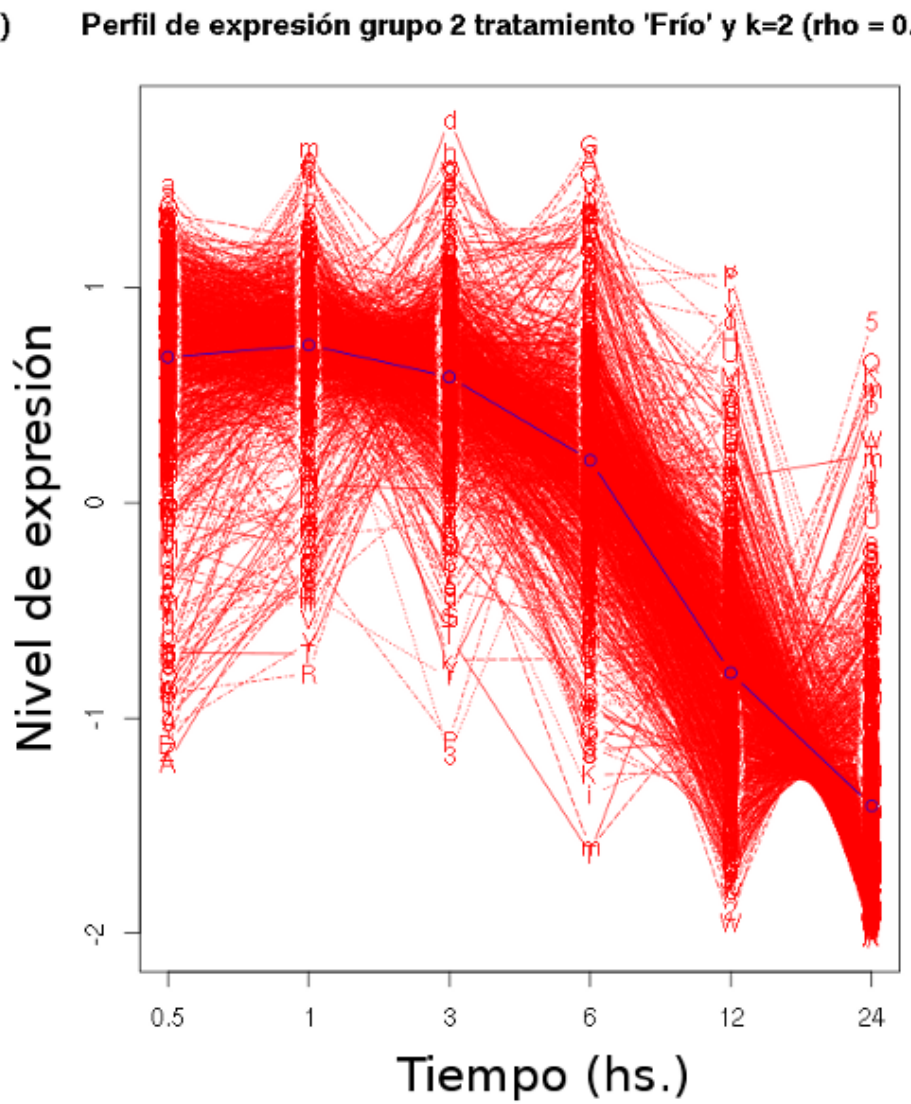
Ejemplo de agrupamientos





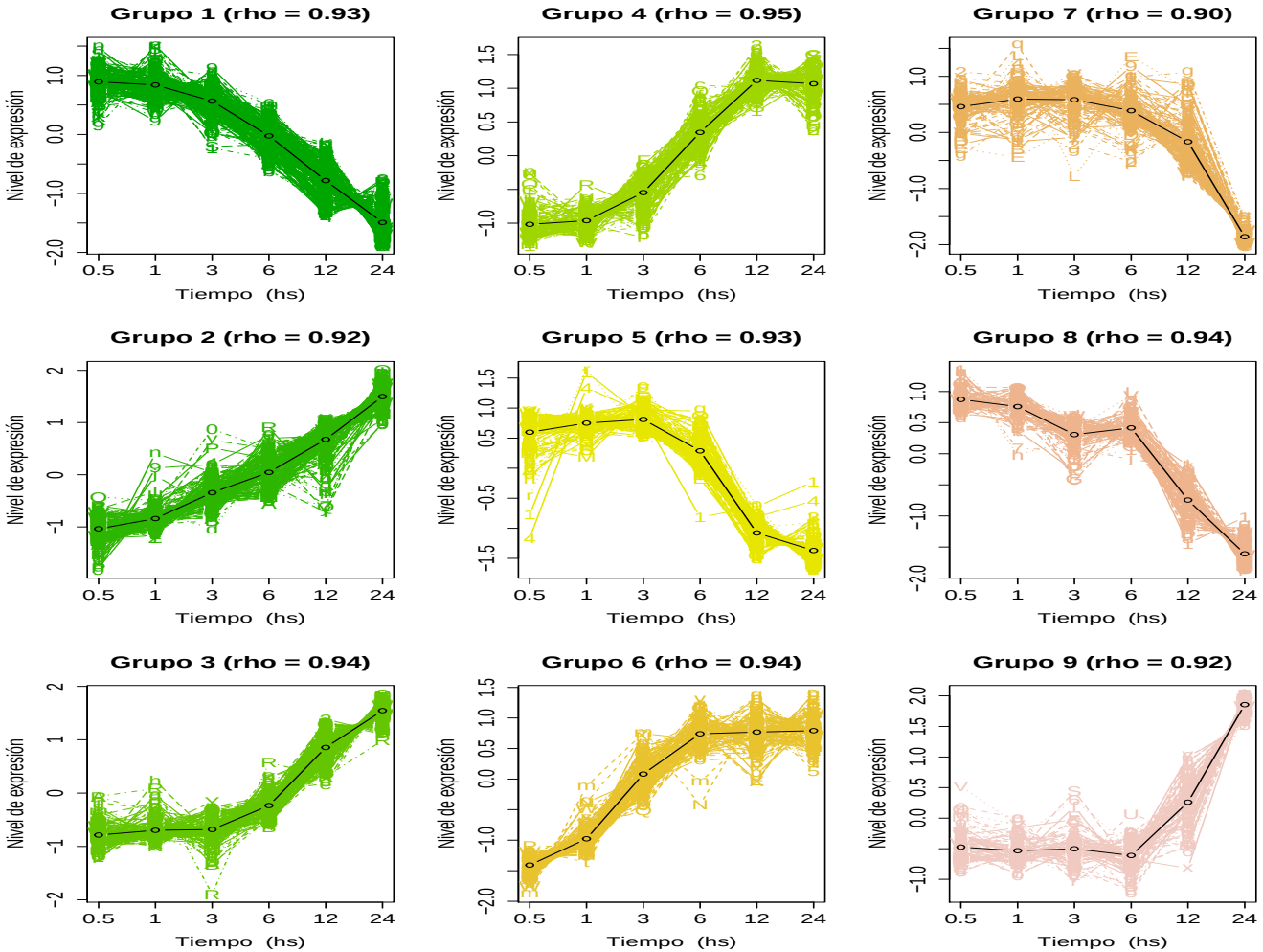
Grupos con k-means





Grupos con corte de árbol dinámico

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

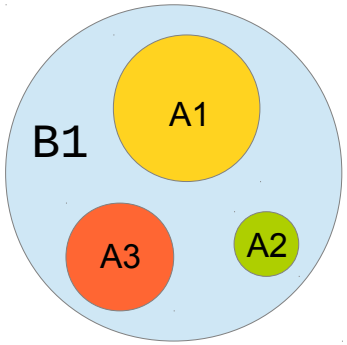


El problema de la escala

Granularidad y resolución de los métodos

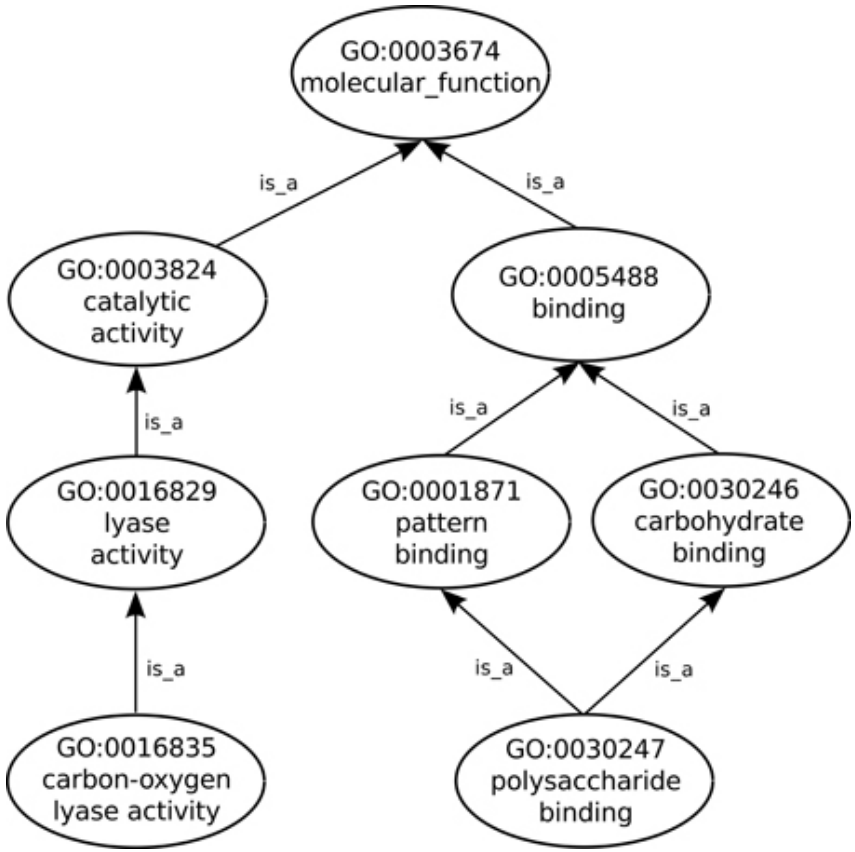
- Una partición A es **más fina** que una partición B si cada grupo de A está contenido en un grupo de B.
- Tenemos **tres formas de realizar particiones** de nuestros datos.
- DS4 genera particiones más finas que DS1 y este a su vez que k-means.
- Tenemos distintas maneras de encontrar estructura en nuestros datos y las **distintas heterogeneidades aparecerán a distintas escalas**.

Vamos a ver si existe una **escala óptima en un sentido biológico** a la que trabajar con éste conjunto de datos y para eso vamos a utilizar 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 término 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:

$$ID(GO_j) = \frac{NE(GO_j)}{N(GO_j)} \quad (1)$$

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.

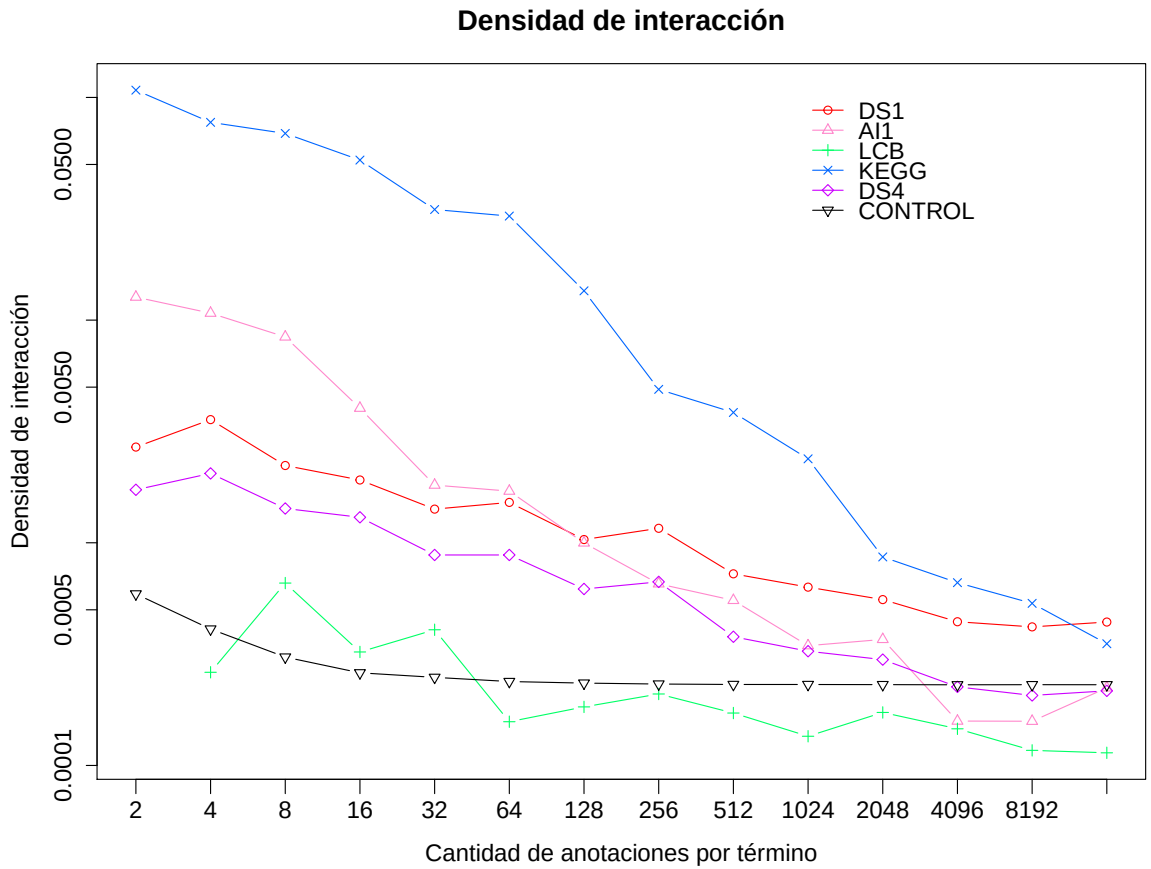
Indice de homogeneidad biológica:

$$BHI_j = \frac{1}{n_j(n_j-1)} \sum_{x \neq y \in D_j} I(C(x) = C(y)) \quad (2)$$

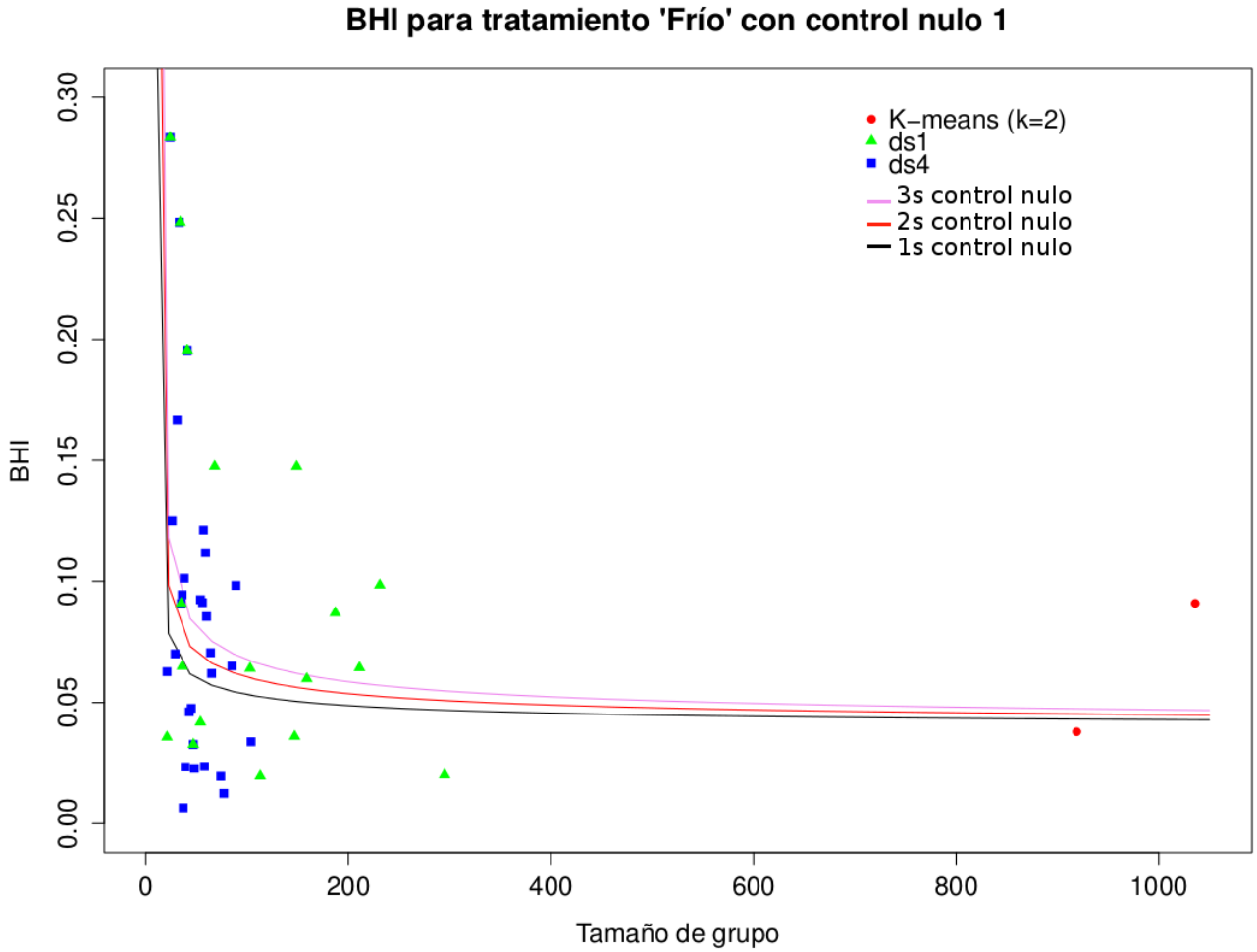
Con n_j la cantidad de genes anotados en el grupo D_j.
La función indicadora $I(C(x) = C(y))$ toma el valor 1 si hay al menos una clase en donde ambos genes estén anotados, y 0 en caso contrario.

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 el control nulo.
4. Los agrupamientos inducidos por otra información presentan mayor congruencia que los inducidos por expresión.



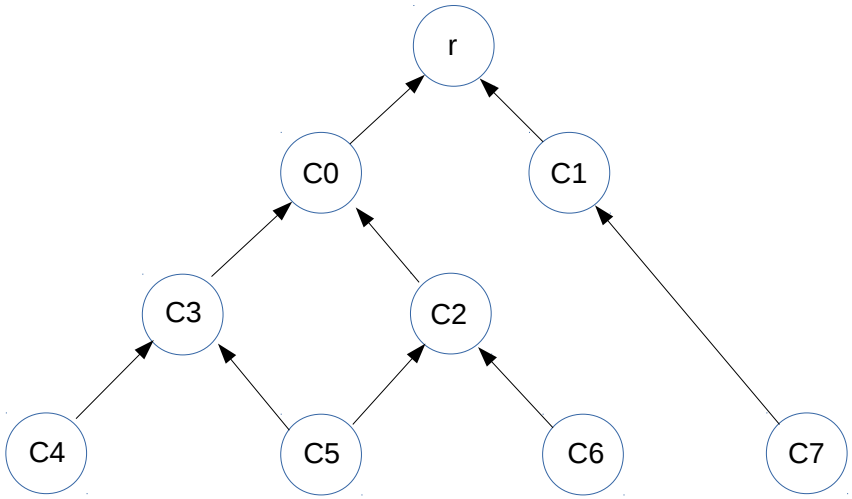
Indice de homogeneidad biológica



Grupos altamente coherentes pero de baja calidad de BHI. Bajo soporte biológico.

Similaridad entre genes en GO

Podemos definir similaridades entre genes en el espacio GO



Utilizando la similaridad entre términos:

$$Sim_{res}(c_i, c_j) = \max_{c \in S(c_i, c_j)} (-log_2[P(c)]) = IC(MICA[c_i, c_j]) \qquad (3)$$

$$Sim_{rcmax}(GO(g_1), GO(g_2)) = \max\{\frac{1}{N}\sum_i \max_{1 \leq j \leq M} S_{ij}, \frac{1}{M}\sum_j \max_{1 \leq i \leq N} S_{ij}\} \qquad (4)$$

KTA global

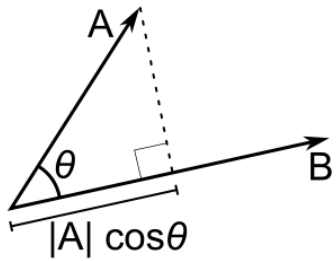
Matriz de similaridad de a pares:

$$K = K_{ij} = k(x_i, x_j) \tag{5}$$

El KTA se define como:

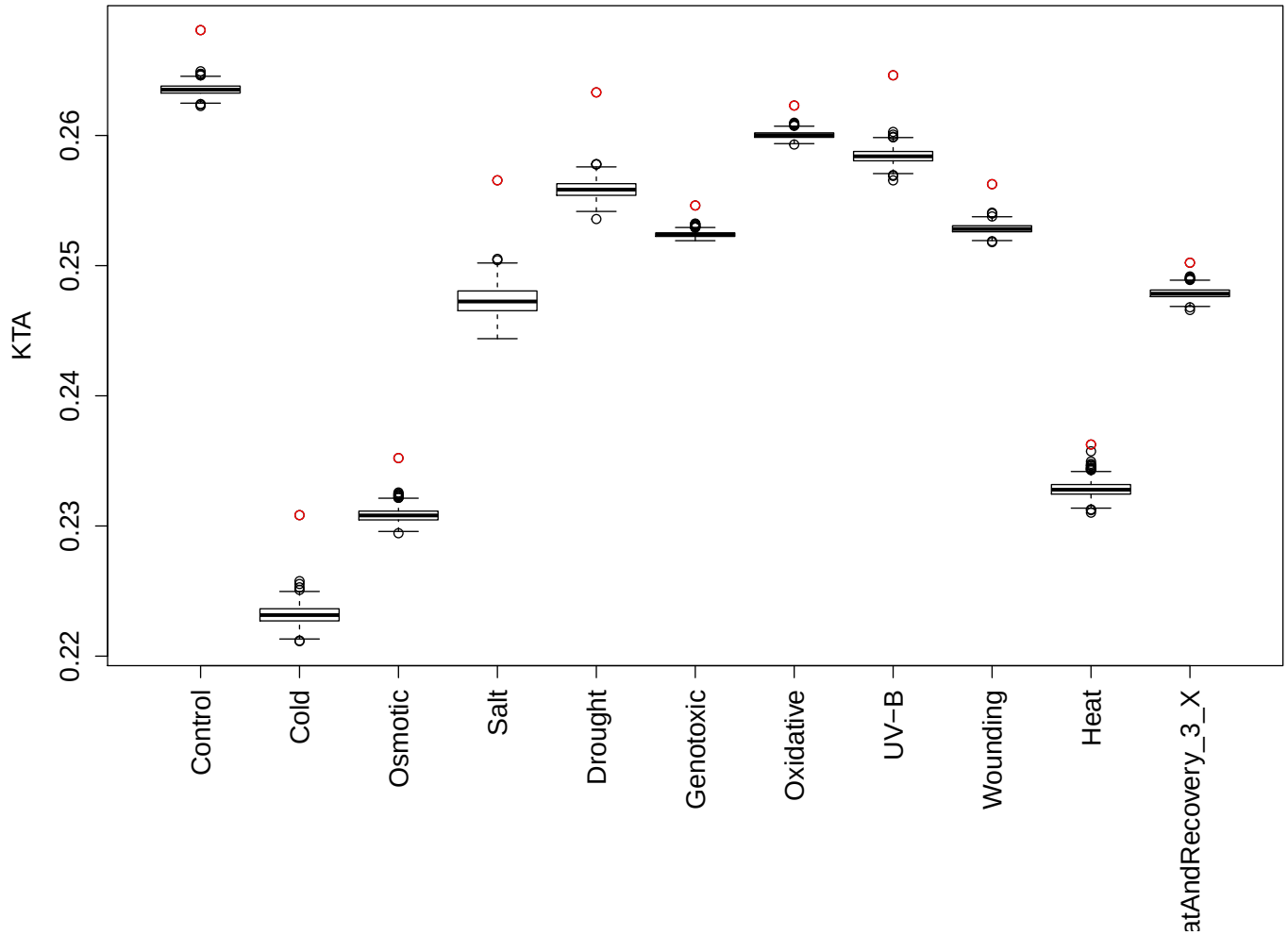
$$KTA(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{6}$$

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



KTA global

KTA global entre expresión y ontología BP con control nulo

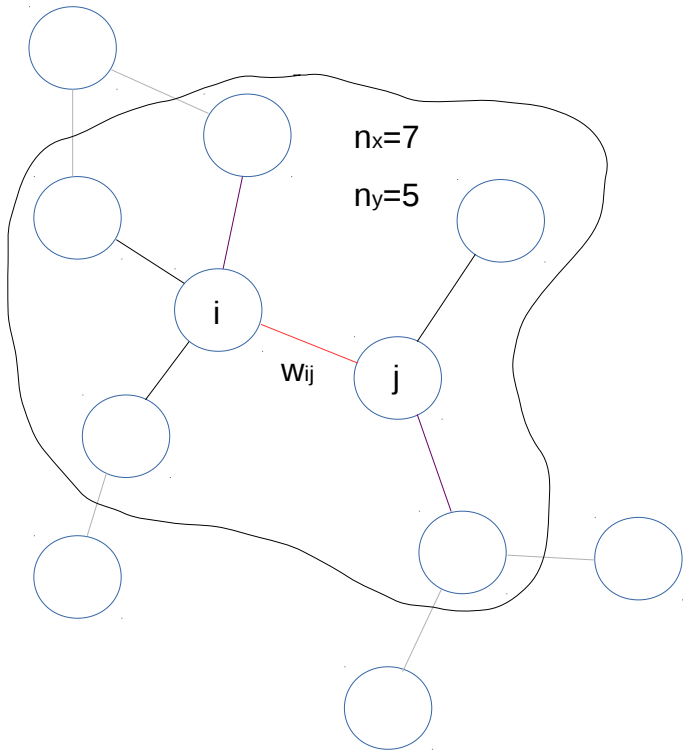


Red 30 primeros vecinos mutuos - vecindades locales

Queremos detectar zonas de alta coherencia.

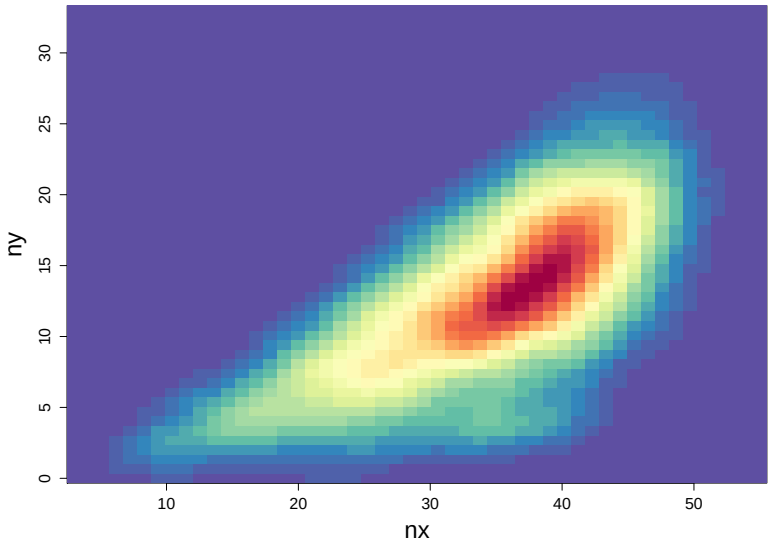
Generamos una red de 30 primeros vecinos mutuos y vamos a ver arista por arista, una localidad definida por los primeros vecinos:

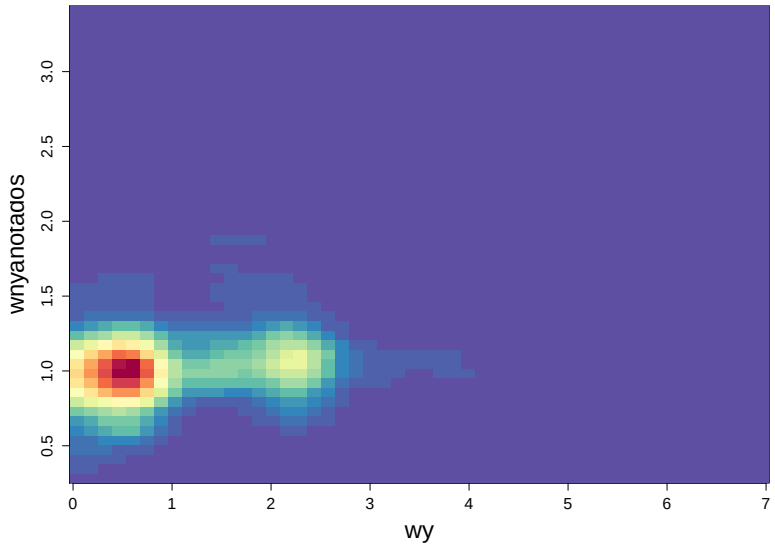
- n_x nodos.
- n_y nodos anotados.
- w_y similaridad entre nodos i y j en GO.
- w_{ny} similaridades promedio en una vecindad en GO.
- $w_{nyanotados}$ similaridades promedio en una vecindad en GO con nodos anotados.
- $ktal_{anotados}$ KTA en la vecindad de los nodos anotados.

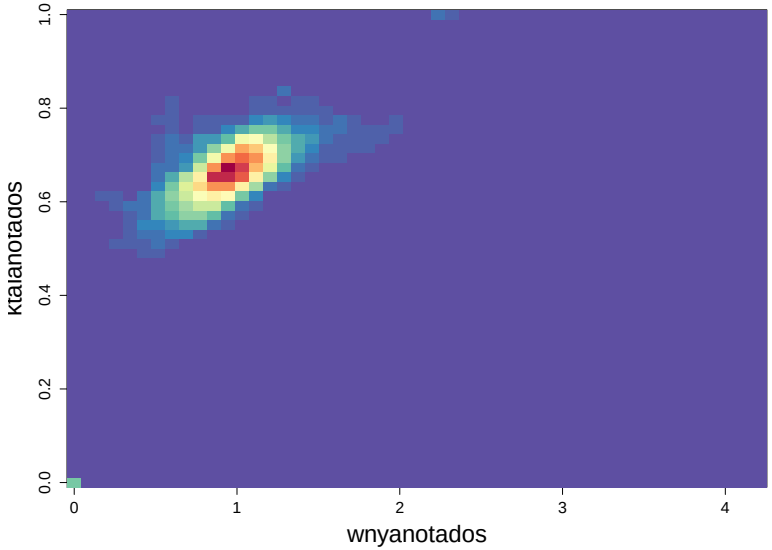


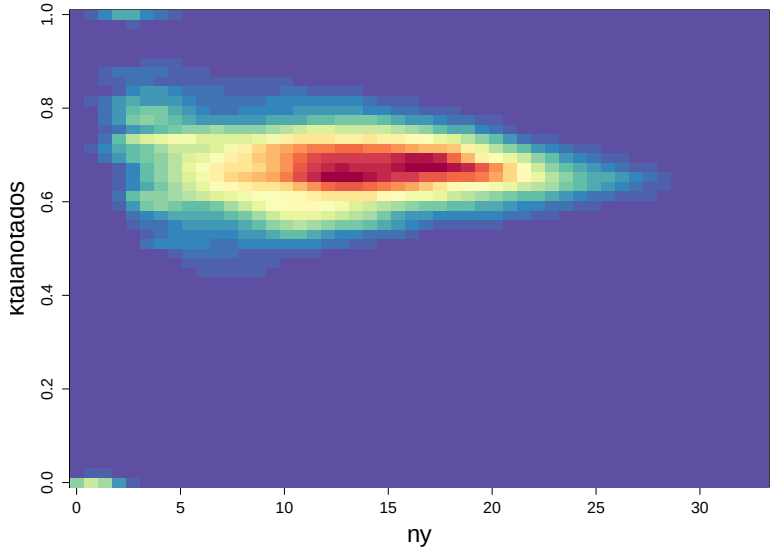
A modo de ejemplo, la red para tratamiento “Frío” consta de 1951 nodos y 18436 aristas.

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 cuantificar si 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:

$$w_{ij} = simcor_{ij}^{\beta * stress_{ij}} \tag{7}$$

Donde:

$$stress_{ij} = \frac{KTA_{fondo}}{KTAl_{ij}} \tag{8}$$

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

β es un parámetro que va a acentuar las heterogeneidades para poder detectar subgrupos.

Métrica mixta y métodos heurísticos

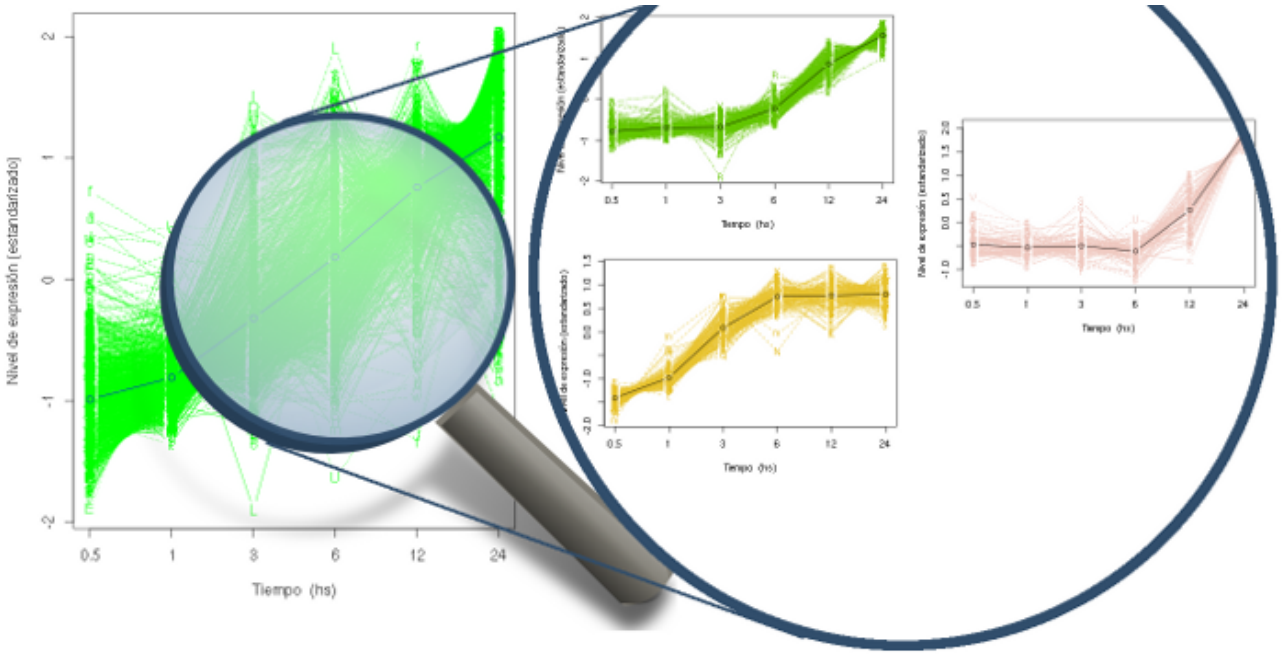
Buscamos subestructura en los grupos a partir de la métrica mixta

Heurísticas con métrica mixta:

- lkta.dtc
- lkta.cnm
- lkta.infomap

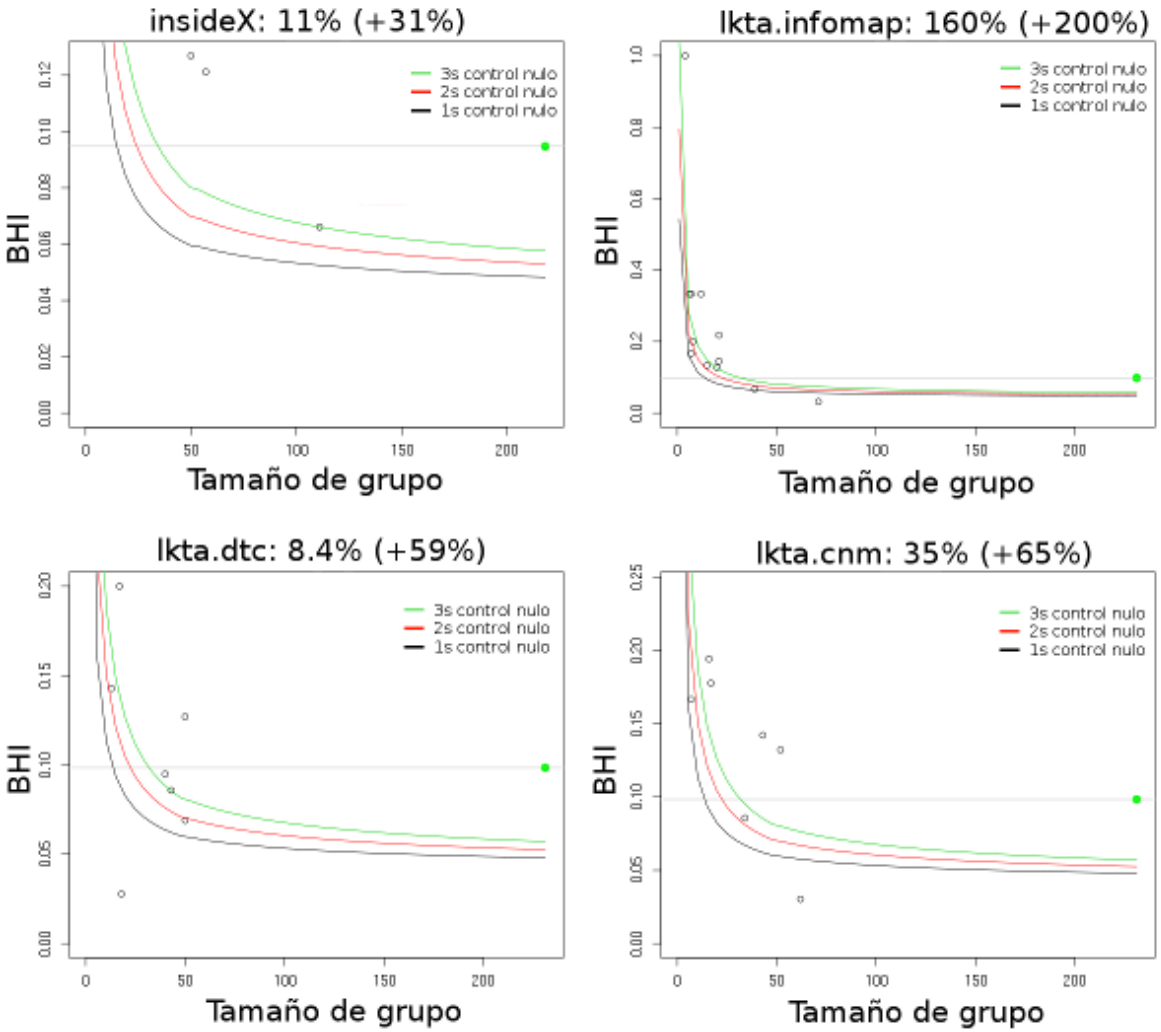
Comparación con métrica transcripcional:

- InsideX



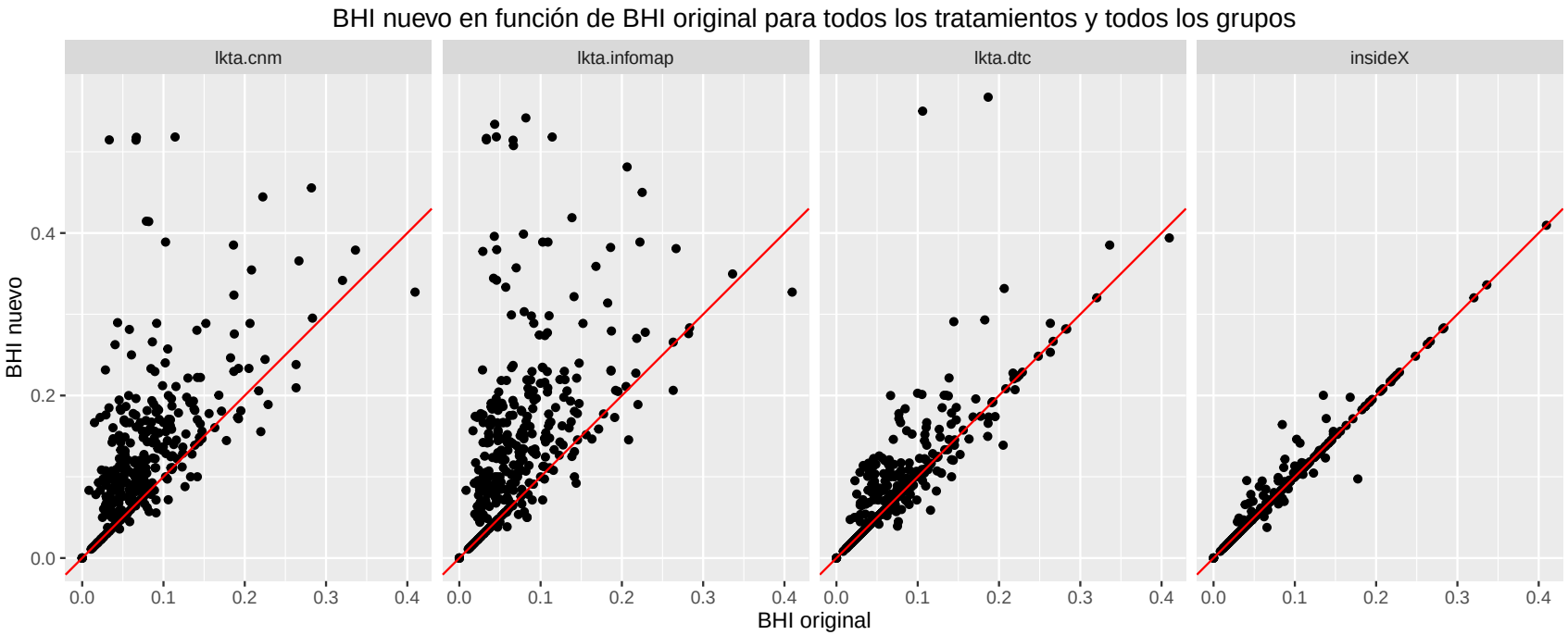
Coherencia biológica medida con BHI - Ejemplo

Subestructura en grupo 2 de tratamiento “Frío”

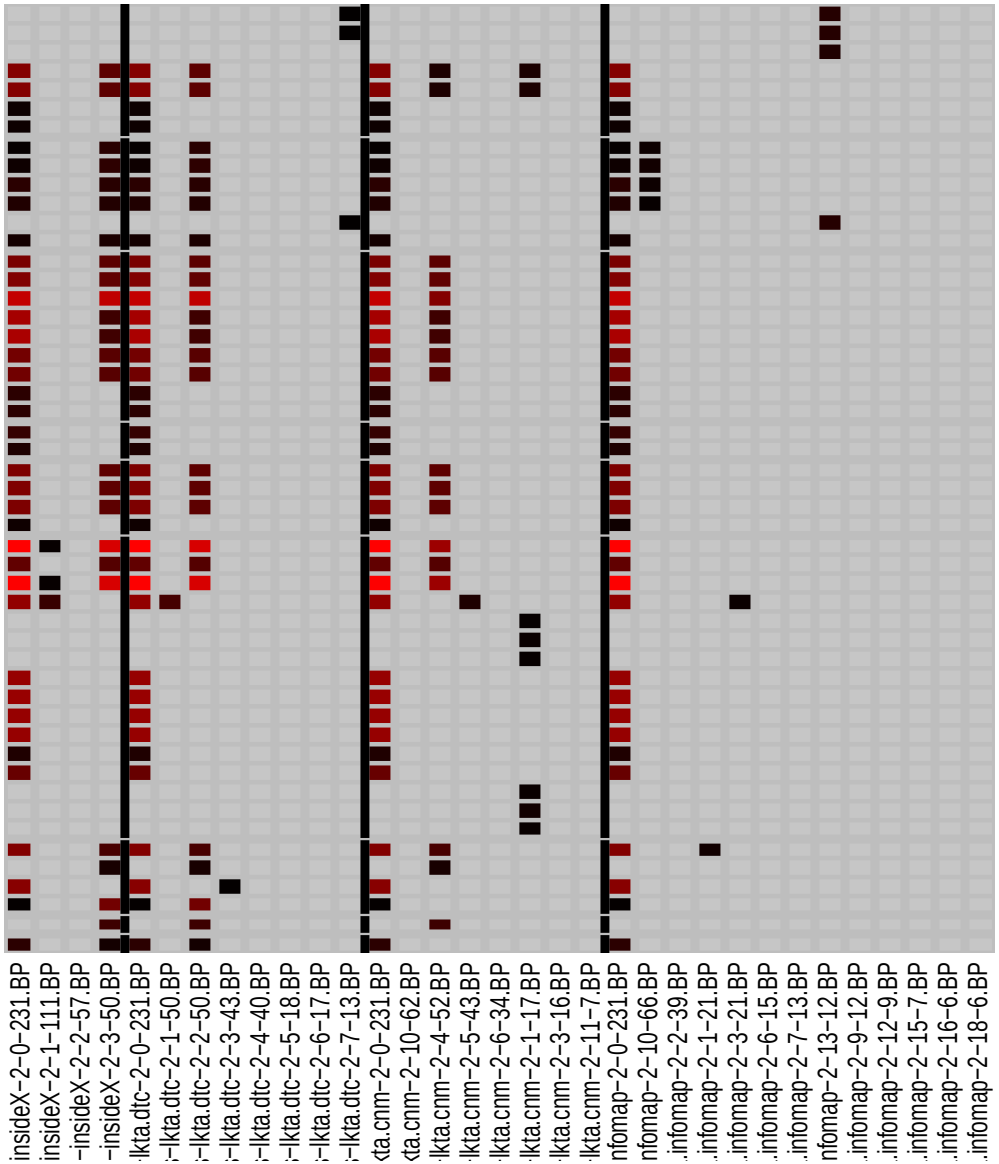


Coherencia biológica medida con BHI

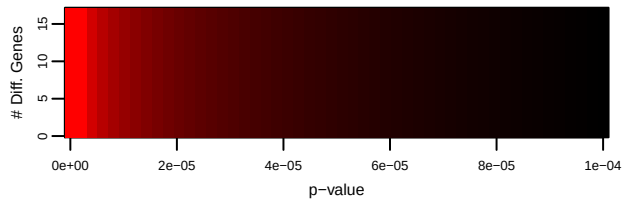
Caracterizamos los nuevos subgrupos hallados



Coherencia biológica a partir de sobrerepresentación (test de Fisher)



IC	Level	
9	5	apocarotenoid metabolic process
9	6	abscisic acid metabolic process
8.6	6	sesquiterpenoid metabolic process
11	3	respiratory burst
11	3	respiratory burst involved in defense response
7.9	4	secondary metabolite catabolic process
7.9	4	toxin catabolic process
9.1	5	phenol−containing compound biosynthetic process
10	6	salicylic acid biosynthetic process
9	5	salicylic acid metabolic process
8.2	4	phenol−containing compound metabolic process
9	5	tertiary alcohol metabolic process
8.7	4	benzene−containing compound metabolic process
8.6	3	regulation of cellular response to stress
9.8	4	regulation of plant−type hypersensitive response
8.5	3	negative regulation of defense response
9	3	negative regulation of cell death
9.3	4	negative regulation of programmed cell death
8	4	regulation of programmed cell death
7.9	4	plant−type hypersensitive response
10	3	signal transduction by protein phosphorylation
10	4	MAPK cascade
8.6	4	regulation of reactive oxygen species metabolic process
10	5	regulation of hydrogen peroxide metabolic process
9.2	4	establishment of protein localization to membrane
9.4	5	protein targeting to membrane
9.2	4	protein localization to membrane
9.2	4	defense response by callose deposition
8.3	4	salicylic acid mediated signaling pathway
9.6	5	systemic acquired resistance, salicylic acid mediated signaling pathway
8.2	5	cellular response to salicylic acid stimulus
8.3	5	response to cyclopentenone
8.7	5	cellular response to decreased oxygen levels
8.8	4	cellular response to hypoxia
8.7	4	cellular response to oxygen levels
10	5	cellular response to unfolded protein
10	4	endoplasmic reticulum unfolded protein response
10	4	cellular response to topologically incorrect protein
10	4	response to unfolded protein
9.7	3	response to topologically incorrect protein
8.9	4	response to endoplasmic reticulum stress
8.2	6	response to hexose
9.7	7	response to fructose
8.1	5	response to monosaccharide
9.2	4	cellular response to heat
11	5	cellular heat acclimation
8.4	4	heat acclimation
9	6	response to absence of light
8.1	3	response to insect
7.9	2	immune effector process



Conclusiones

- Diferentes técnicas de agrupamiento nos permitieron obtener grupos correlacionados en el espacio de expresión.
- Cada método obtiene descripciones a diferente resolución.
- Buscamos analizar estas descripciones en función de la interpretabilidad biológica.
- Presentamos los observables BHI y KTA para cuantificar la coherencia entre los espacios y corroboramos que lo detectado en el espacio transcripcional es en general coherente con el conocimiento biológico.
- Introdujimos una versión local de KTA que nos permitió definir una métrica mixta.
- Presentamos heurísticas para identificar subestructuras transcripcionales con alta interpretabilidad y coherencia biológica.

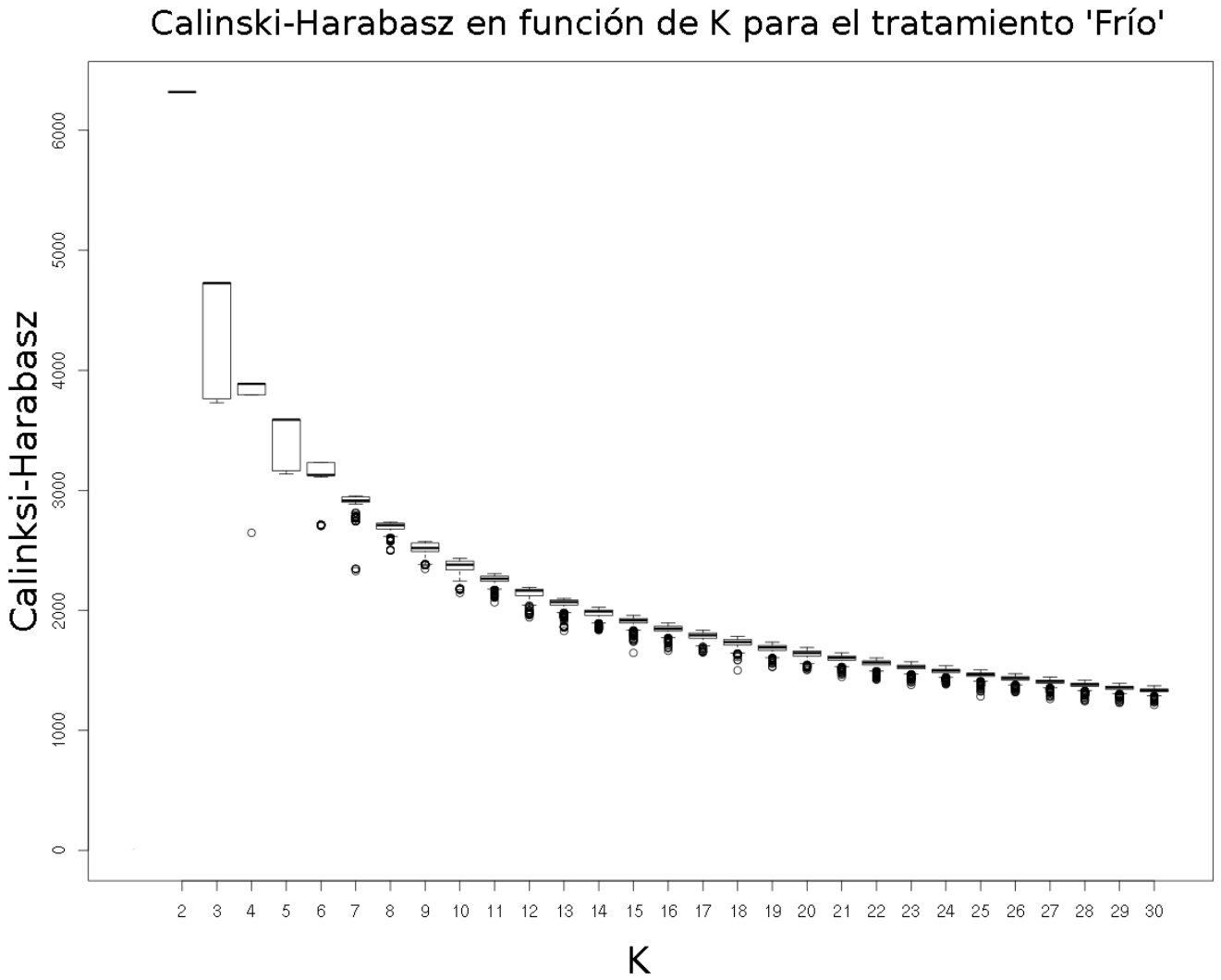
Muchas gracias

¡Muchas gracias por su atención!

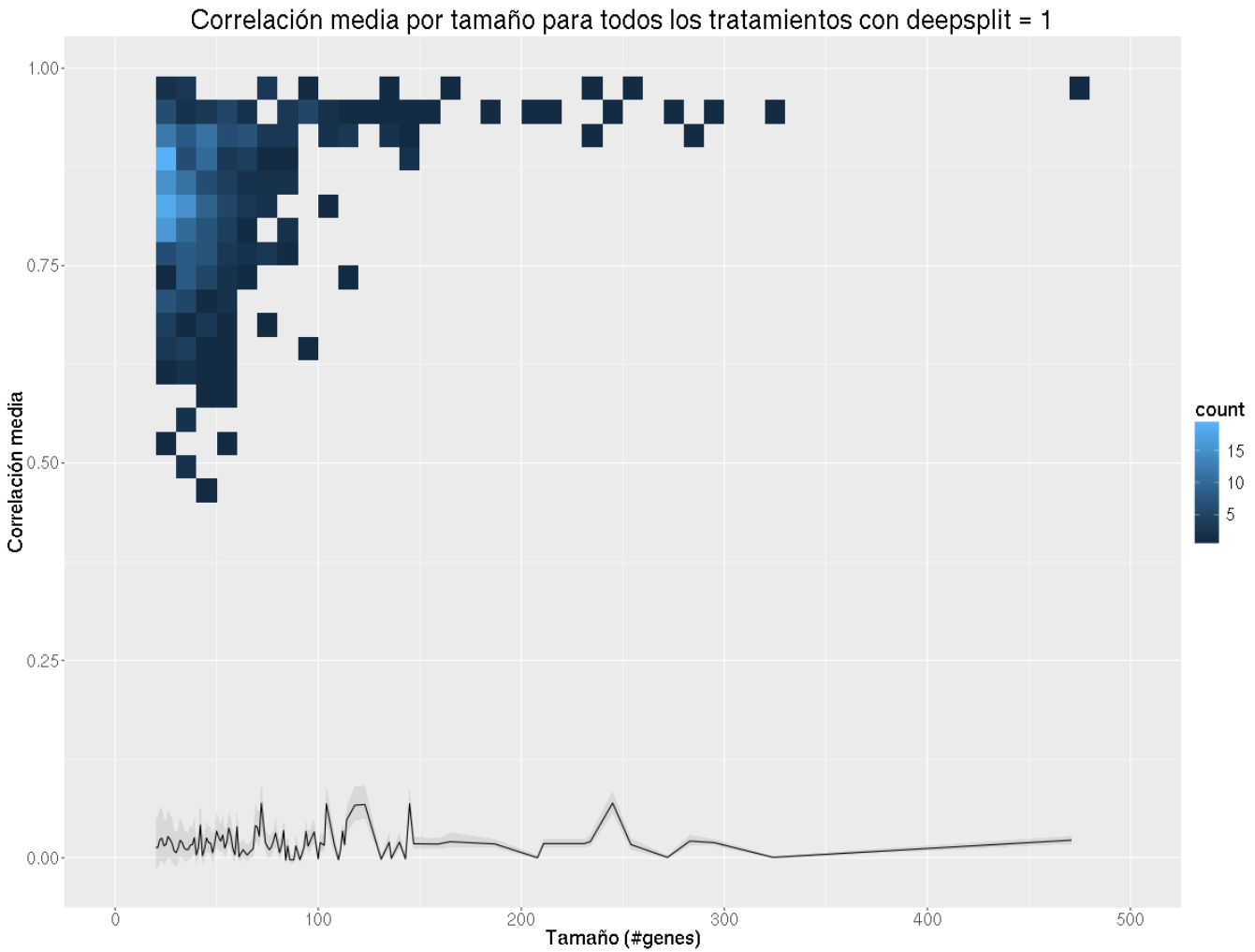
¿Preguntas o comentarios?

El grupo

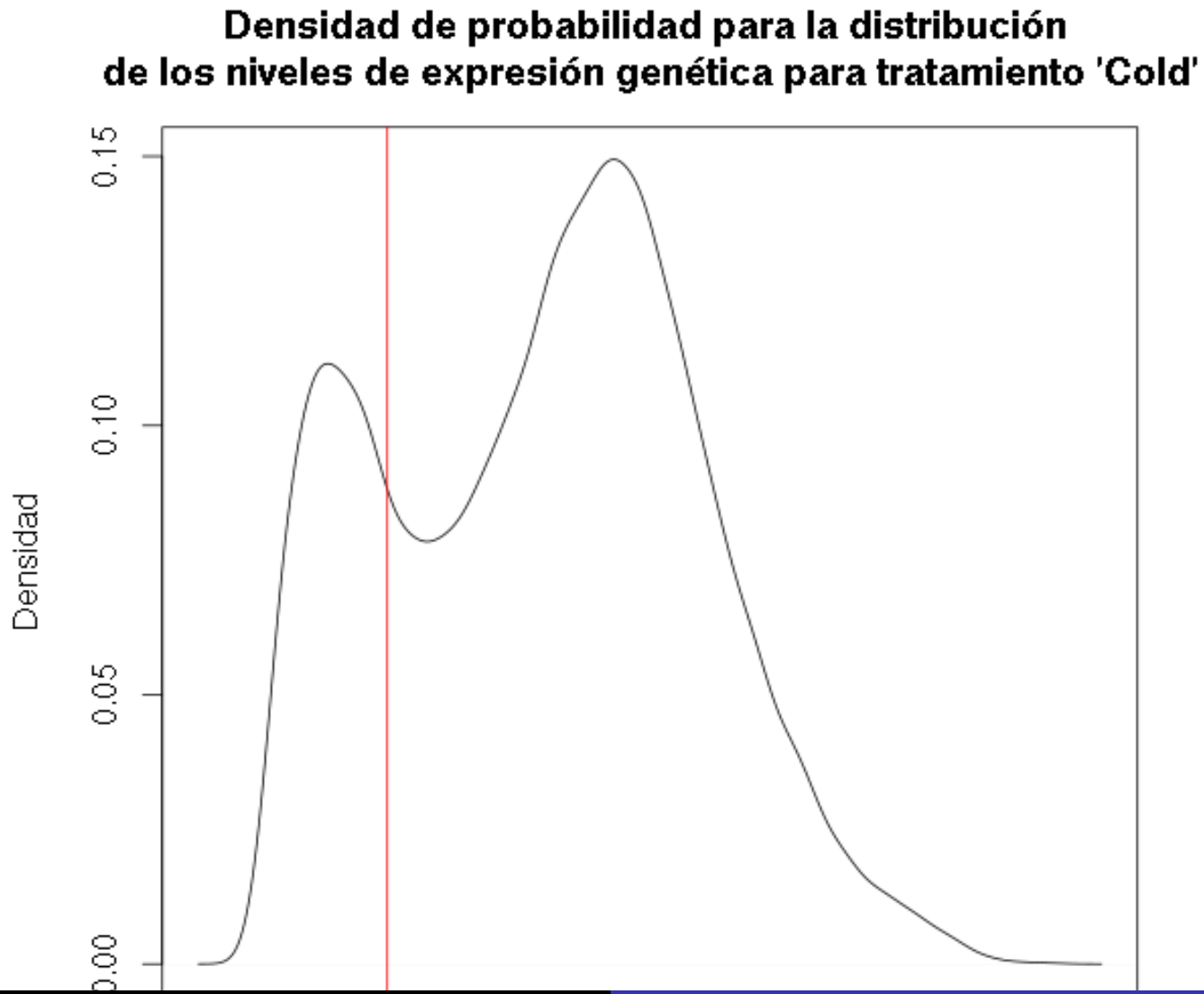
Contenido extra



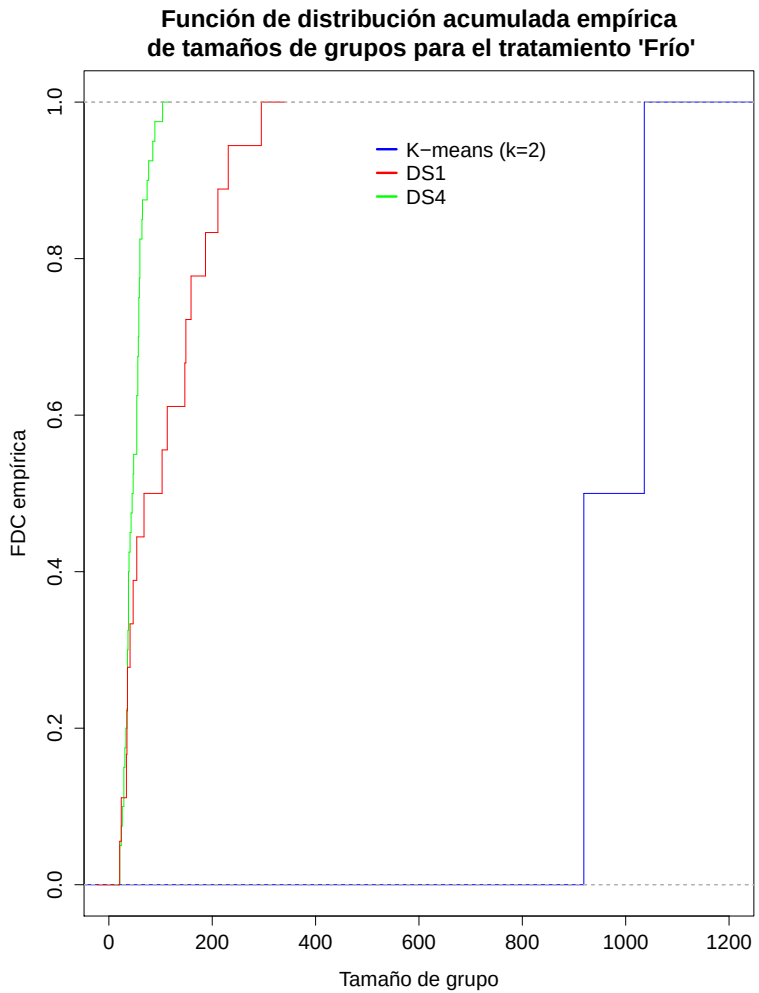
Contenido extra



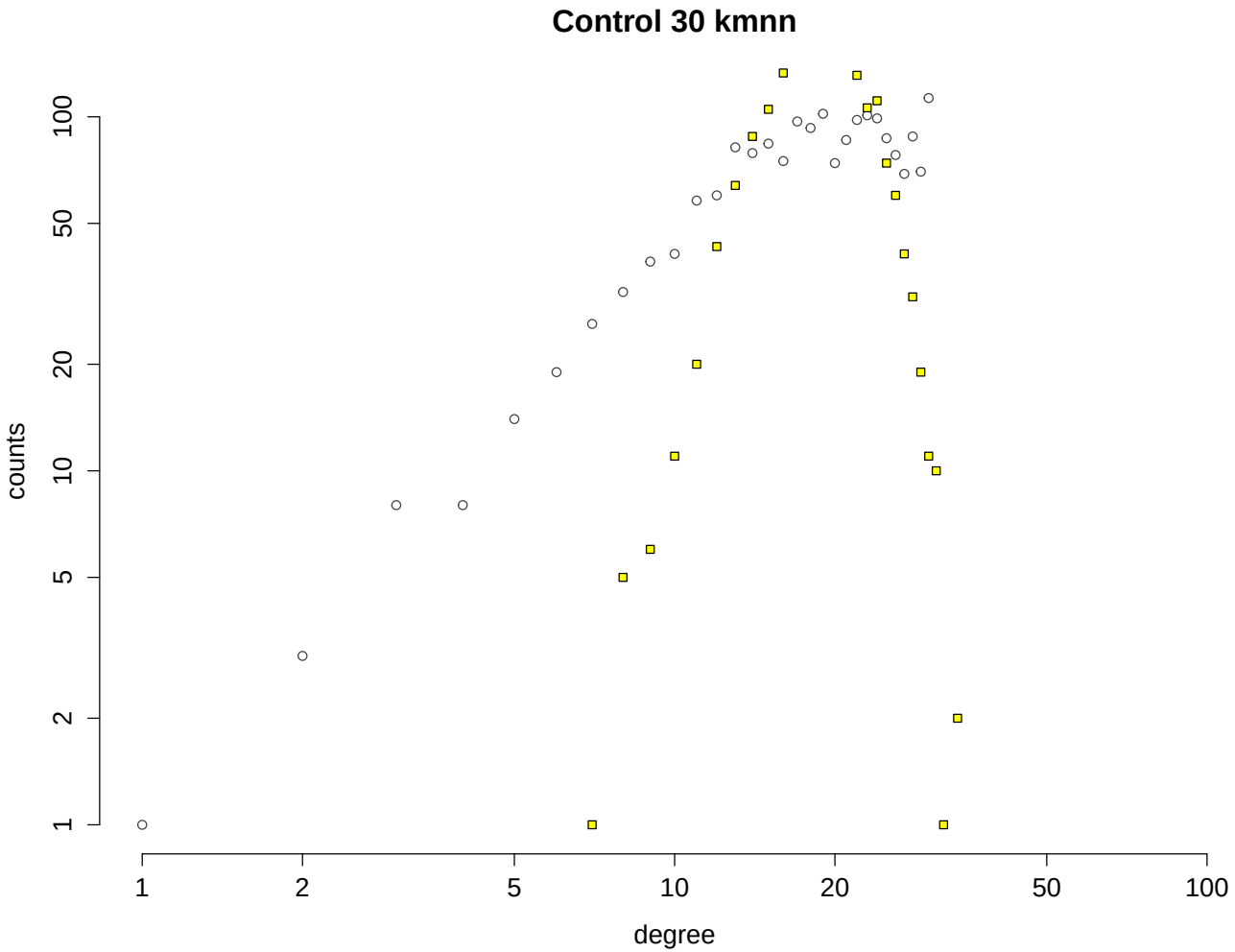
Contenido extra



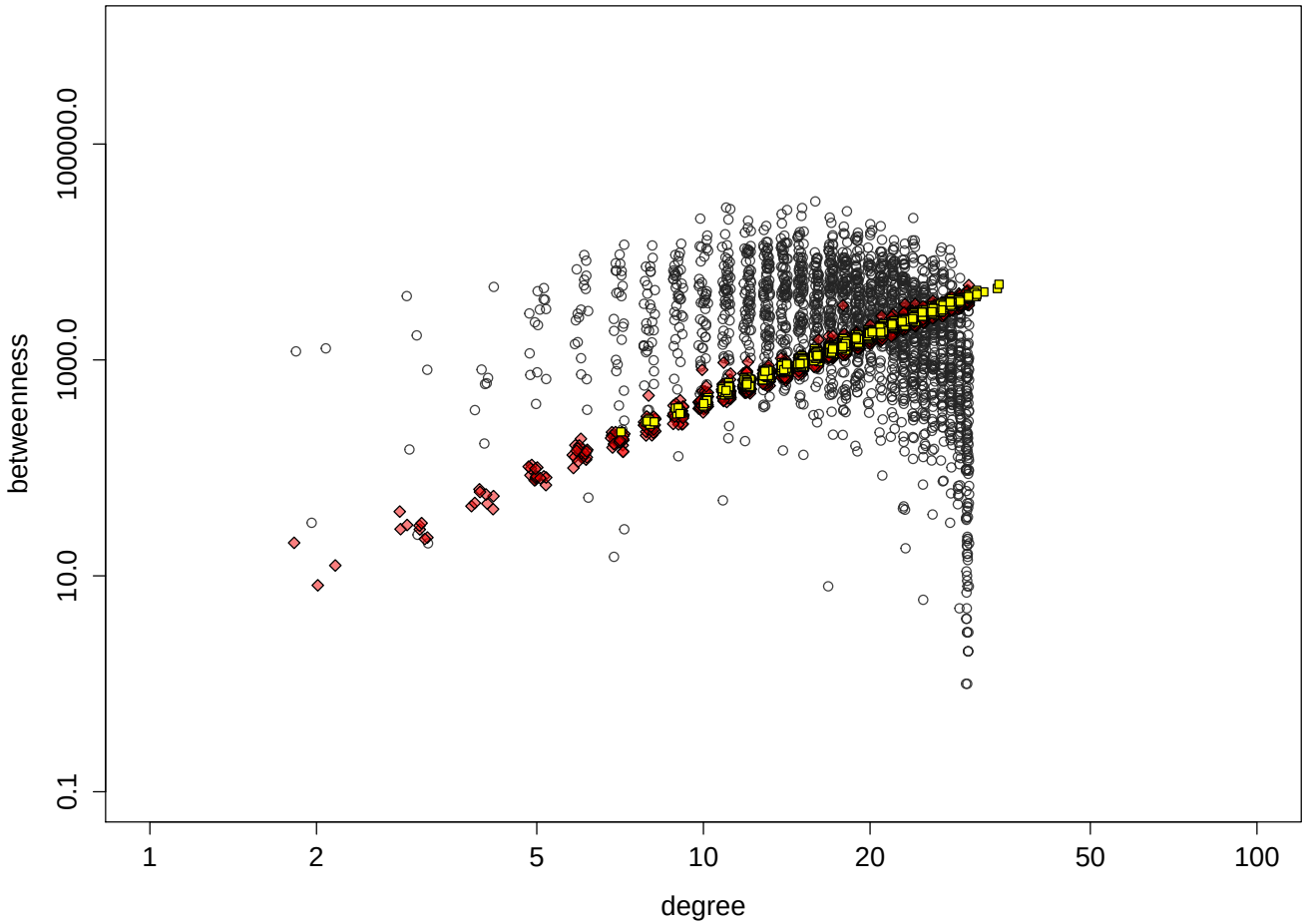
Contenido extra



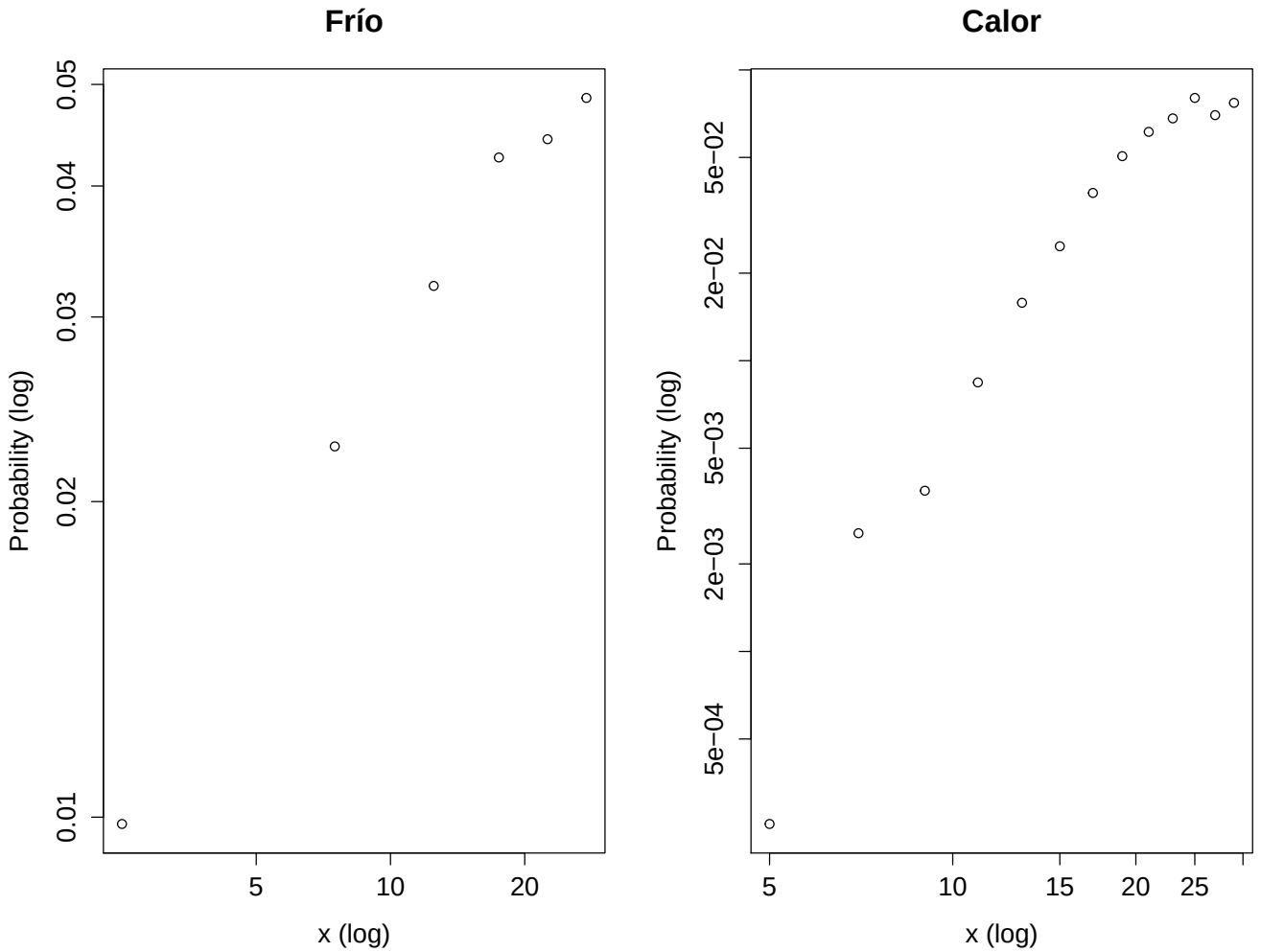
Contenido extra



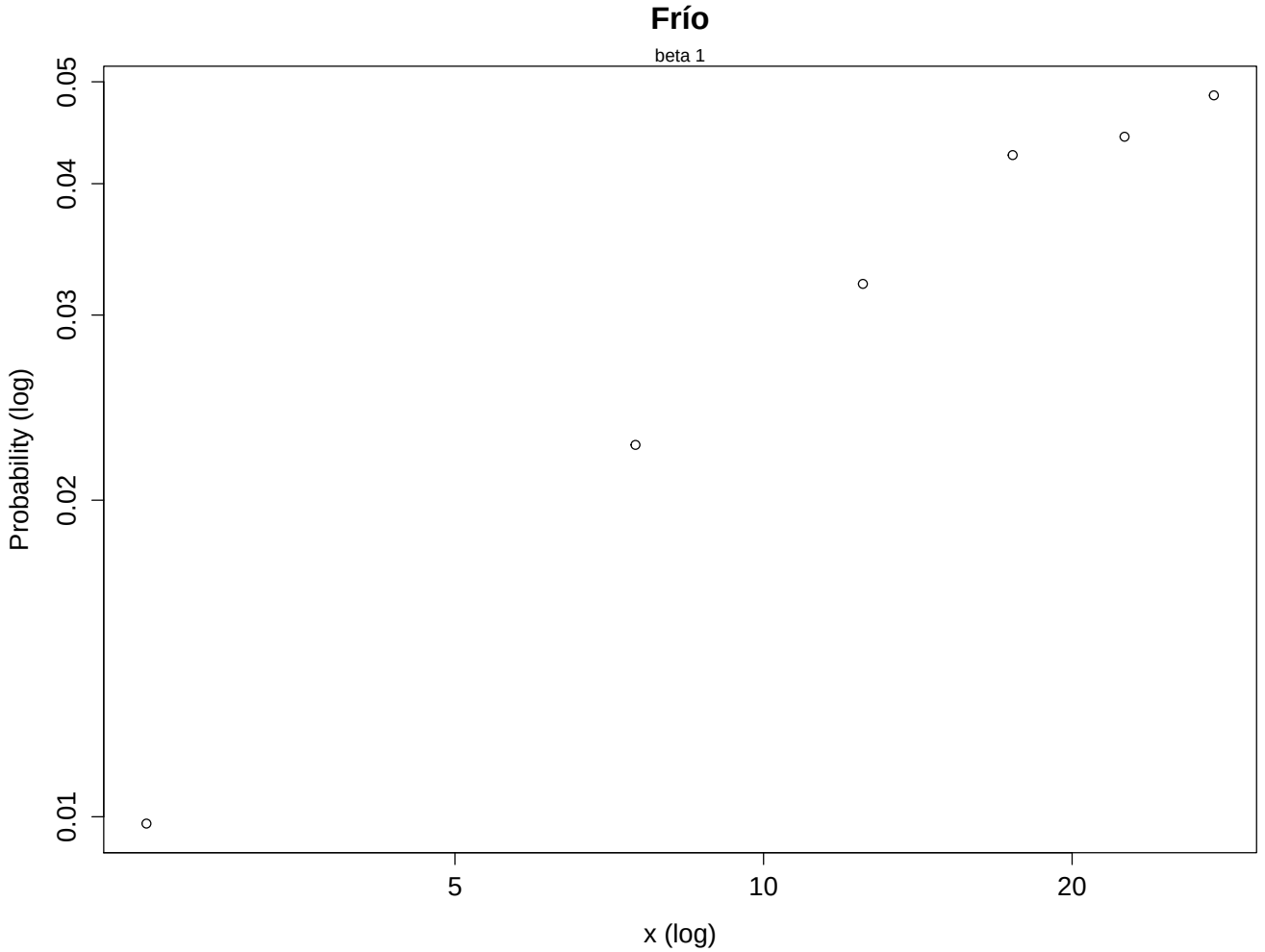
Contenido extra



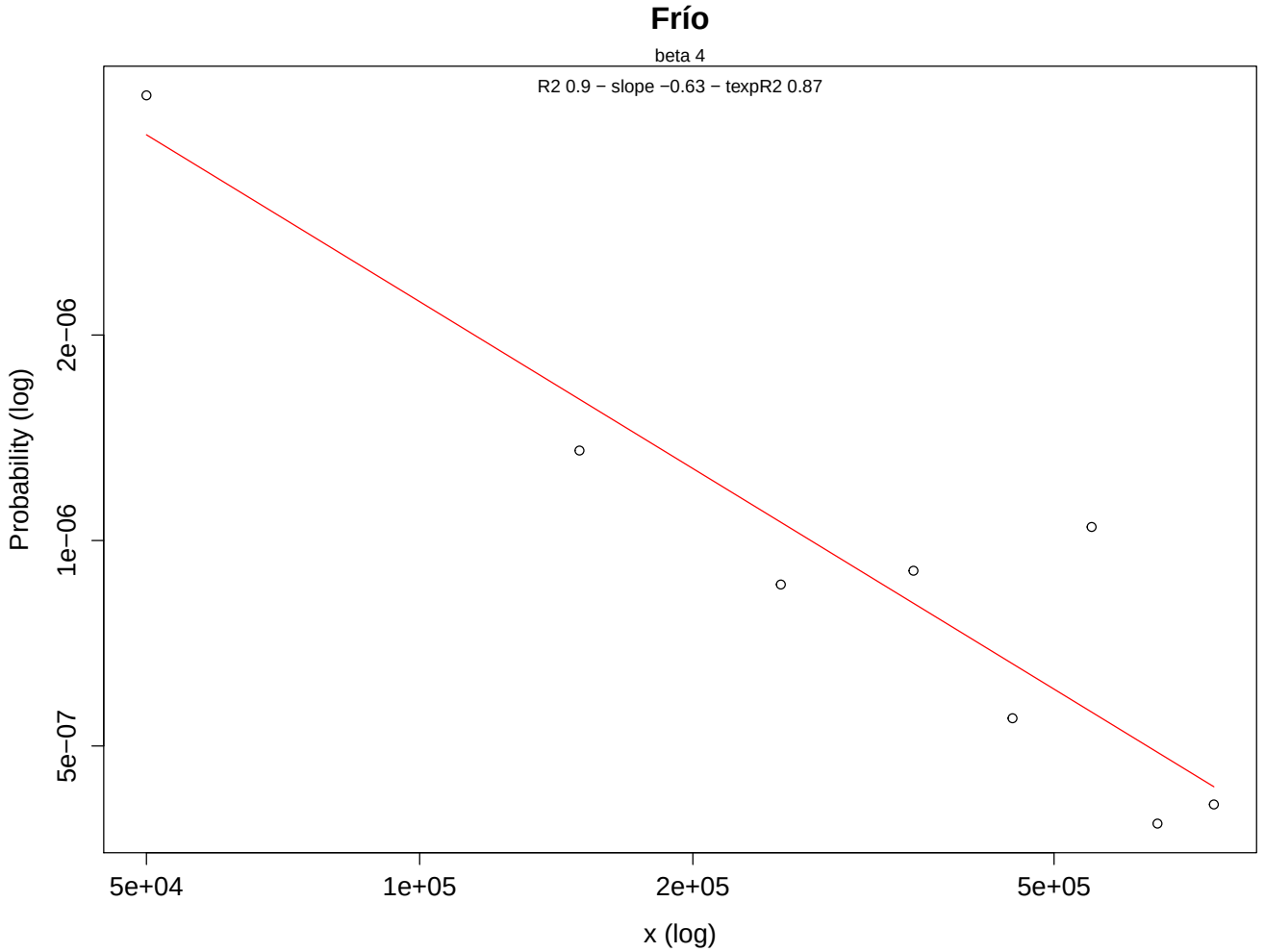
Contenido extra



Contenido extra



Contenido extra



Contenido extra