

# 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.

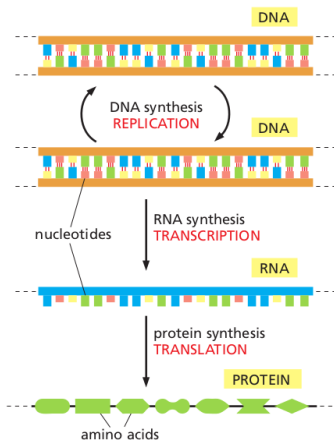
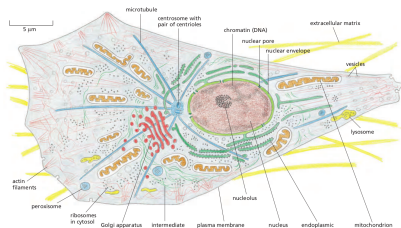


# Contenido

- 1 Introducción
  - Relevamientos transcripcionales de gran escala
  - Detección de correlaciones
- 2 Análisis de relevamientos transcripcionales
  - Medidas de similaridad y distancia
  - Métodos de agrupamiento utilizados
  - Caracterización de particiones
- 3 Congruencia biológica
  - Ontología génica (GO)
  - Cuantificando la congruencia biológica
- 4 Coherencia entre métricas
  - Métrica en GO
  - KTA global
  - Modulación de heterogeneidades transcripcionales con GO
- 5 Conclusiones y perspectivas

# Transcripción y traducción (dogma central de la biología molecular)

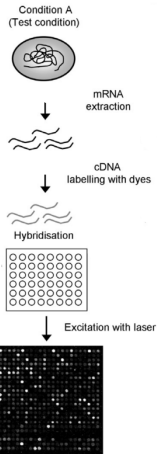
Células, ADN, ARNm, proteínas y otras yerbas...



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

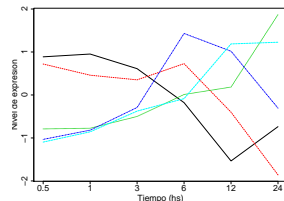


A. Thaliana



## 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.

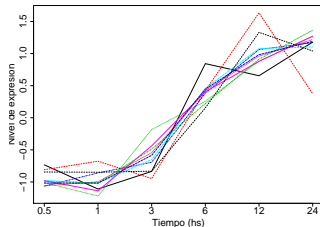
- Son métodos no supervisados.
- Consisten en agrupar elementos “similares entre si”.
- Permiten el descubrimiento de patrones en los datos.
- Posibilitan obtener conclusiones sobre los datos.

### 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  $\{2, 3\}$

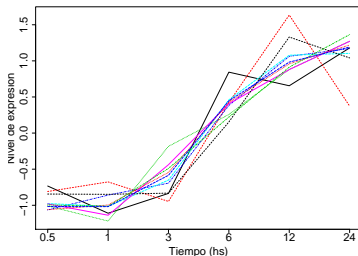


## Medidas de similitud y distancia

Distancia basada en el coeficiente de correlación de Pearson:

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

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



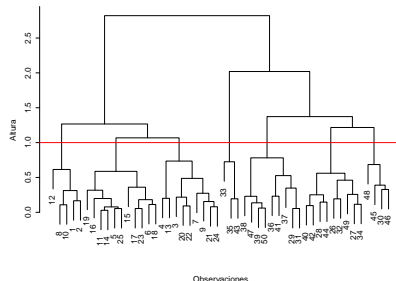
## Métodos de agrupamiento

### Método k-means

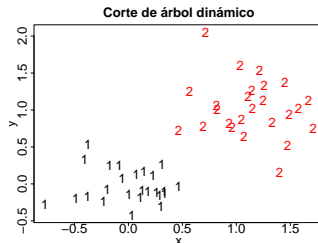
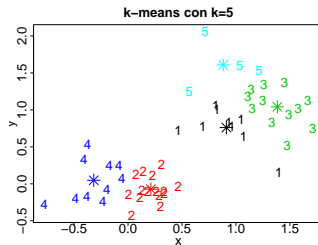
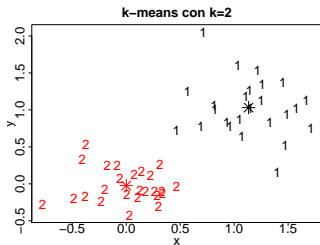
- Busca estructuras compactas.
- La cantidad k de grupos debe ser fijada a priori.
- Muy rápida ejecución.

### 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.



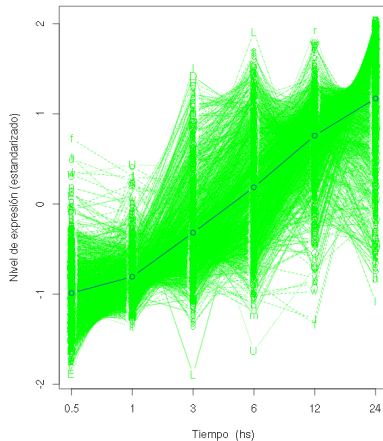
# Métodos de agrupamiento - ejemplos



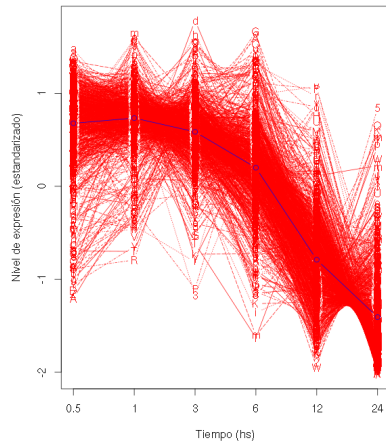


## Perfiles tratamiento “Frío” con k-means

Perfil de expresión grupo 1 tratamiento 'Frío' y k=2 ( $\rho = 0.74$ )

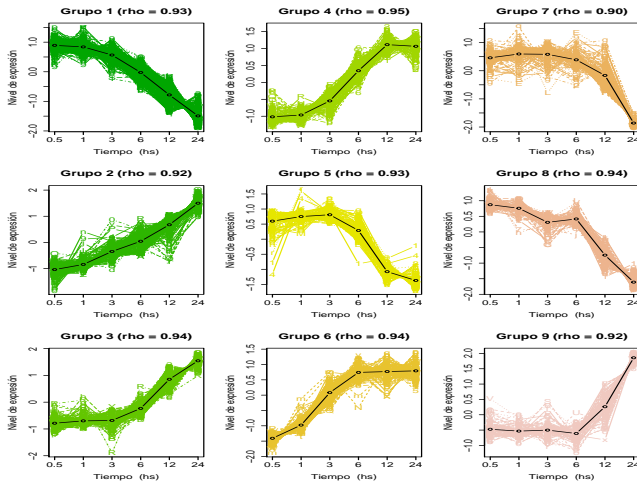


Perfil de expresión grupo 2 tratamiento 'Frío' y k=2 ( $\rho = 0.79$ )



## 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.



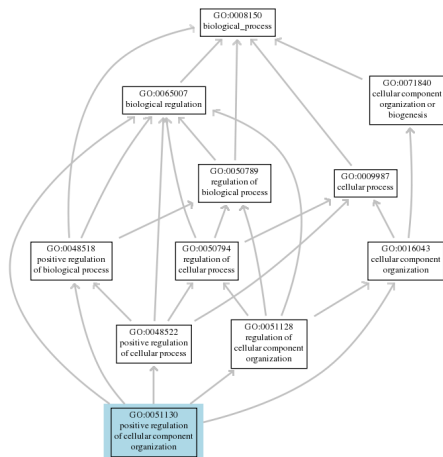
## Granularidad de las particiones

### 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 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:

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

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  $GO_j$ .

Índice de homogeneidad biológica:

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

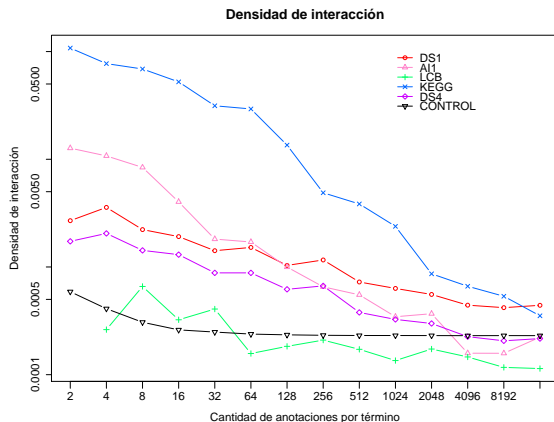
Con  $n_j$  la cantidad de genes anotados

en el grupo  $D_j$ .

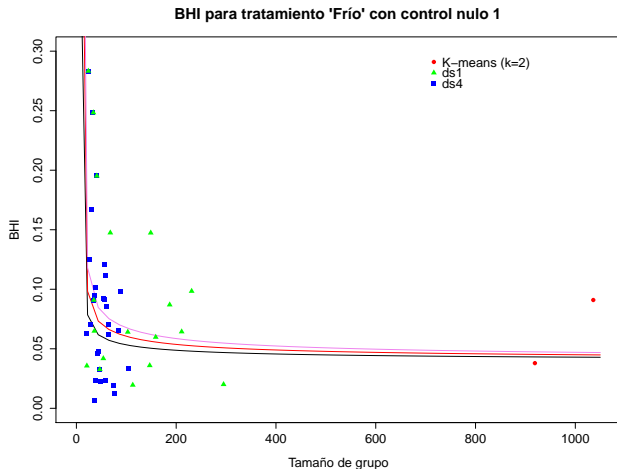
La función indicadora  $I(C(x) = C(y))$  que toma el valor 1 si hay al menos una clase en donde ambos genes estén anotados, y 0 en caso contrario.

## 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.



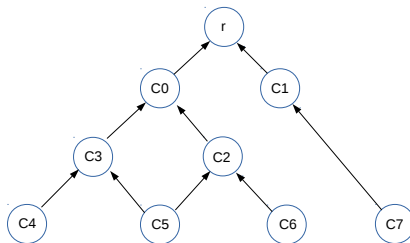
## Índice de homogeneidad biológica



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

## Similaridad entre genes en GO

Podemos definir similitudes entre genes en el espacio GO



Utilizando la similitud entre términos:

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



## KTA global

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

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

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:

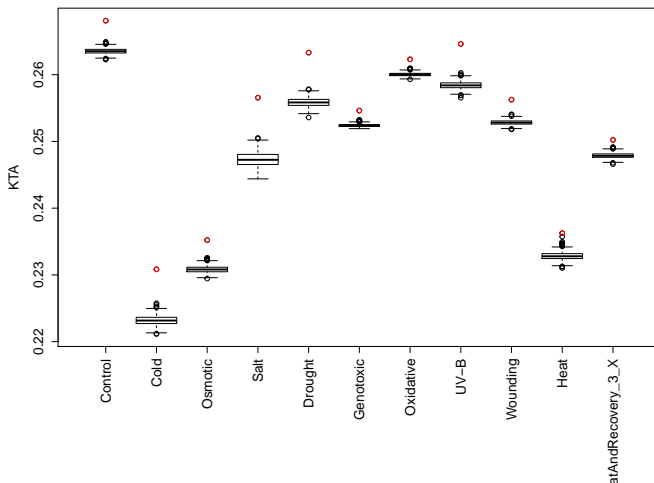
$$\hat{A}(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \quad (7)$$

con  $\langle K_1, K_1 \rangle_F = \sum_{i,j=1}^m K_1(x_i, x_j) K_1(x_i, x_j)$  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

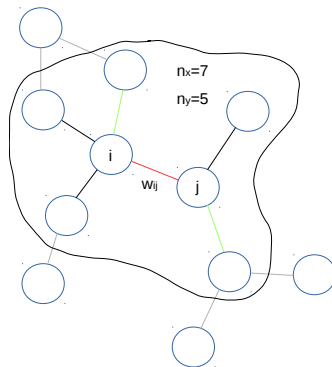


## Red 30 primeros vecinos mutuos - vecindades locales

Queremos detectar zonas de alta coherencia.

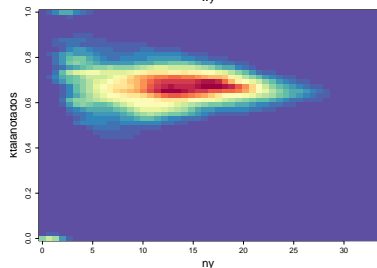
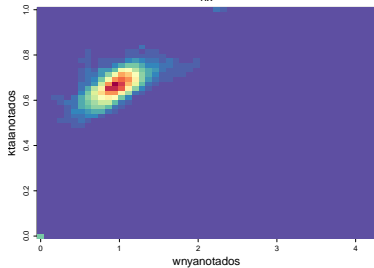
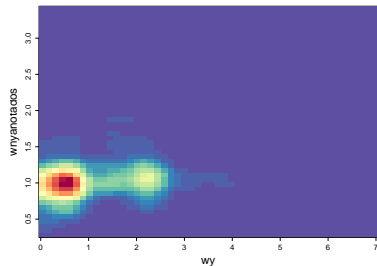
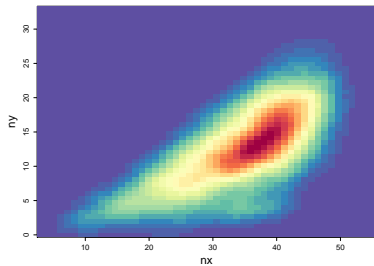
Generamos una red de 30 primeros vecinos mutuos y vamos a ver arista por arista, una localidad definida por los primeros vecinos:

- $n_x$  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:

$$w_{ij} = \text{simcor}_{ij}^{\beta * \text{stress}_{ij}} \quad (8)$$

Donde:

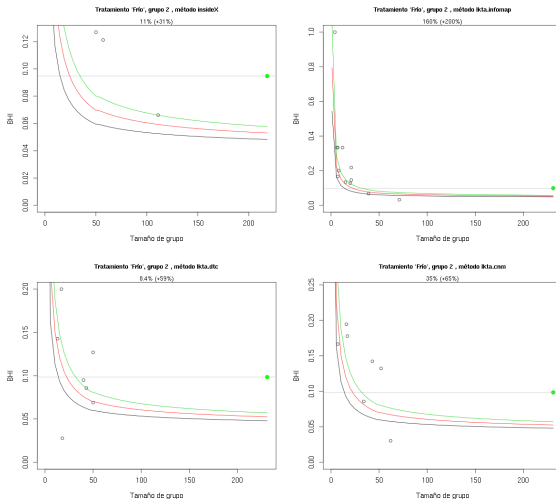
$$\text{stress}_{ij} = \frac{KTA_{fondo}}{KTA_{ij}} \quad (9)$$

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

$\beta$  es un parámetro que permite aumentar aún más la homogeneidad de la red.

## Métodos heurísticos

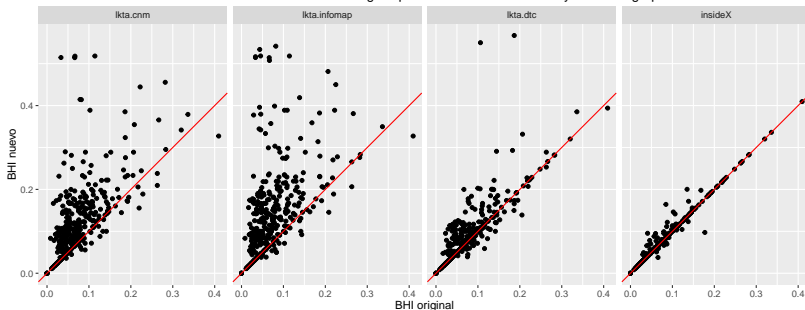
Buscamos subestructura en los grupos a partir de la métrica mixta

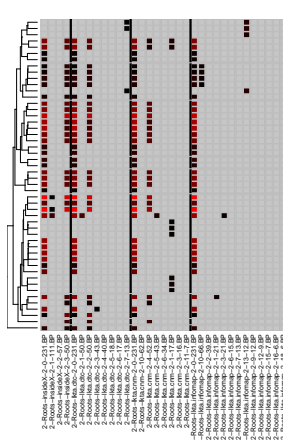


## Métodos heurísticos - caracterización de particiones

Caracterizamos los nuevos subgrupos hallados

BHI nuevo en función de BHI original para todos los tratamientos y todos los grupos





8	1	5	apoptosis	
9	2	6	apoptosis	
10	3	7	apoptosis	
11	4	8	apoptosis	
12	5	9	apoptosis	
13	6	10	apoptosis	
14	7	11	apoptosis	
15	8	12	apoptosis	
16	9	13	apoptosis	
17	10	14	apoptosis	
18	11	15	apoptosis	
19	12	16	apoptosis	
20	13	17	apoptosis	
21	14	18	apoptosis	
22	15	19	apoptosis	
23	16	20	apoptosis	
24	17	21	apoptosis	
25	18	22	apoptosis	
26	19	23	apoptosis	
27	20	24	apoptosis	
28	21	25	apoptosis	
29	22	26	apoptosis	
30	23	27	apoptosis	
31	24	28	apoptosis	
32	25	29	apoptosis	
33	26	30	apoptosis	
34	27	31	apoptosis	
35	28	32	apoptosis	
36	29	33	apoptosis	
37	30	34	apoptosis	
38	31	35	apoptosis	
39	32	36	apoptosis	
40	33	37	apoptosis	
41	34	38	apoptosis	
42	35	39	apoptosis	
43	36	40	apoptosis	
44	37	41	apoptosis	
45	38	42	apoptosis	
46	39	43	apoptosis	
47	40	44	apoptosis	
48	41	45	apoptosis	
49	42	46	apoptosis	
50	43	47	apoptosis	
51	44	48	apoptosis	
52	45	49	apoptosis	
53	46	50	apoptosis	
54	47	51	apoptosis	
55	48	52	apoptosis	
56	49	53	apoptosis	
57	50	54	apoptosis	
58	51	55	apoptosis	
59	52	56	apoptosis	
60	53	57	apoptosis	
61	54	58	apoptosis	
62	55	59	apoptosis	
63	56	60	apoptosis	
64	57	61	apoptosis	
65	58	62	apoptosis	
66	59	63	apoptosis	
67	60	64	apoptosis	
68	61	65	apoptosis	
69	62	66	apoptosis	
70	63	67	apoptosis	
71	64	68	apoptosis	
72	65	69	apoptosis	
73	66	70	apoptosis	
74	67	71	apoptosis	
75	68	72	apoptosis	
76	69	73	apoptosis	
77	70	74	apoptosis	
78	71	75	apoptosis	
79	72	76	apoptosis	
80	73	77	apoptosis	
81	74	78	apoptosis	
82	75	79	apoptosis	
83	76	80	apoptosis	
84	77	81	apoptosis	
85	78	82	apoptosis	
86	79	83	apoptosis	
87	80	84	apoptosis	
88	81	85	apoptosis	
89	82	86	apoptosis	
90	83	87	apoptosis	
91	84	88	apoptosis	
92	85	89	apoptosis	
93	86	90	apoptosis	
94	87	91	apoptosis	
95	88	92	apoptosis	
96	89	93	apoptosis	
97	90	94	apoptosis	
98	91	95	apoptosis	
99				



## 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  $\approx 50$  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