Universidad de Santiago de Chile
Laboratorio de Bioinformática (2016-2)
Análisis Funcional con DAVID

Profesor: Mario Inostroza Ponta
Ayudante: Jorge Párraga-Álava (jorge.parraga@usach.cl)

26 Noviembre, 2016

1. Introducción

Típicamente, las tareas bioinformáticas producen una basta serie de archivos o información como: listas de genes, nivel de expresión, fold change, cluster, etc. que muchas veces no son de fácil interpretación, por ello para facilitar esta actividad, se utilizan herramientas de análisis funcional. El análisis funcional consiste en definir la función de un grupo de genes (cluster, diferenciales, etc.) y gráficamente comparar los resultados correspondientes.

2. Objetivo

Este laboratorio tiene como objetivo entregar al estudiante conceptos básicos sobre la utilización de DAVID como herramienta para el análisis funcional de datos de expresión génica.

3. DAVID

Uno de los sistemas más consultados para el análisis funcional de genómica/proteómica es DAVID (*Database for Annotation, Visualization and Integrated Discovery*). DAVID es un recurso bioinformático online que proporciona herramientas para la interpretación funcional de grandes listas de genes / proteínas.

Para el analisis funcional de genes, DAVID, dispone de tres herramientas **Functional Annotation Clustering**, **Functional Annotation Chart** y **Functional Annotation Table** las que obtienen informacion biológica en base a diferentes anotaciones. Cada anotación aporta con datos en diferentes contextos, por ejemplo: **KEGG PATHWAY** es una colección de mapas de rutas biológicas que representan conocimiento sobre las redes de interacción y reacción moleculares para procesos celulares, sistemas del organismo, enfermedades humanas, etc. **OMIM_DISEASE** es un catálogo de genes humanos que contiene información sobre rasgos y trastornos genéticos.

A pesar que DAVID ofrece una interfaz web, la manipulación de las anotaciones se vuelve compleja cuando se requiere realizar análisis detallados o visualizaciones personalizadas de los resultados obtenidos en ella. Por ello, en esta ayudantía tiene como objetivo ofrecer a los alumnos una guía para la obtención de anotaciones funcionales de datos génicos usando la librería **RDAVIDWebService**.

4. Uso de librería RDAVIDWebService

RDAVIDWebService es una librería de R para la obtención de datos desde DAVID usando servicios web, con ella se pueden obtener, en base a una lista de genes, informes sobre grupos de genes, anotaciones por categorías y resumen de las mismas.

A continuación usaremos esta libreria para obtener anotaciones funcionales de una lista de 1203 genes con identificador "ENTREZ_GENE_ID", los cuales han sido proporcionados por el ayudante o puede descargarlos **aquí**. Al contar con el archivo *.text* de los genes, procedemos de la siguiente forma:

4.1 Conexión al servicio web

- Cree una carpeta, guarde el archivo de genes descargado.
- Cargue las librerías **RDAVIDWebService** y **ggplot2**.

```
source("http://bioconductor.org/biocLite.R")
```

```
## Bioconductor version 3.2 (BiocInstaller 1.20.3), ?biocLite for help

## A new version of Bioconductor is available after installing the most
##   recent version of R; see http://bioconductor.org/install
```

```
library("RDAVIDWebService")
```

```
## Loading required package: graph

## Loading required package: GOstats

## Loading required package: Biobase

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
##     grep, grepl, intersect, is.unsorted, lapply, lengths, Map,
##     mapply, match, mget, order, paste, pmax, pmax.int, pmin,
##     pmin.int, Position, rank, rbind, Reduce, rownames, sapply,
##     setdiff, sort, table, tapply, union, unique, unlist, unsplit
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: Category

## Loading required package: stats4

## Loading required package: Matrix

## Loading required package: AnnotationDbi

## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:Matrix':
##
##     expand

## Loading required package: GO.db

## Loading required package: DBI

##

##
## Attaching package: 'GOstats'

## The following object is masked from 'package:AnnotationDbi':
##
##     makeGOGraph

## Loading required package: ggplot2

##
## Attaching package: 'RDAVIDWebService'

## The following object is masked from 'package:AnnotationDbi':
##
##     species

## The following object is masked from 'package:IRanges':
##
##     members

## The following objects are masked from 'package:BiocGenerics':
##
##     counts, species
```

```
library("ggplot2")
library("Rgraphviz")
```

```
## Loading required package: grid
```

- Cargue la lista de genes.

```
ListaGenes<-read.csv("C:/Users/jbele/Desktop/genesDAVID.txt",stringsAsFactors=FALSE)
```

- La librería **RDAVIDWebService** requiere en primer lugar establecer una conexión al servicio web de DAVID, para ello se requiere contar con un email registrado en tal servicio. Puede usar el usuario jorge.parraga@usach.cl o acceder a http://david.abcc.ncifcrf.gov/webservice/register.htm y registrar su email.

```
david<-DAVIDWebService$new(email="jorge.parraga@usach.cl",
url="https://david.ncifcrf.gov/webservice/services/DAVIDWebService.DAVIDWebServiceHttpSoap12Endpoint/")
```

- Averigue las categorías y tipos de ID de genes que estan disponibles.

```
getIdTypes(david) #Obtiene la lista de tipos de ID de genes que pueden usarse.
getAllAnnotationCategoryNames(david) #Lista de categorías funcionales que pueden usarse.
getDefaultCategoryNames(david) #Obtiene las categorías activas por default.
```

- Agregue el archivo descargado como lista de genes a consultar.

```
  result<-addList(david,ListaGenes,
                  idType="ENTREZ_GENE_ID",
                  listName=colnames(ListaGenes),
                  listType="Gene")
```

- Ahora configuramos solo dos categorías de anotaciones funcionales a consultar.

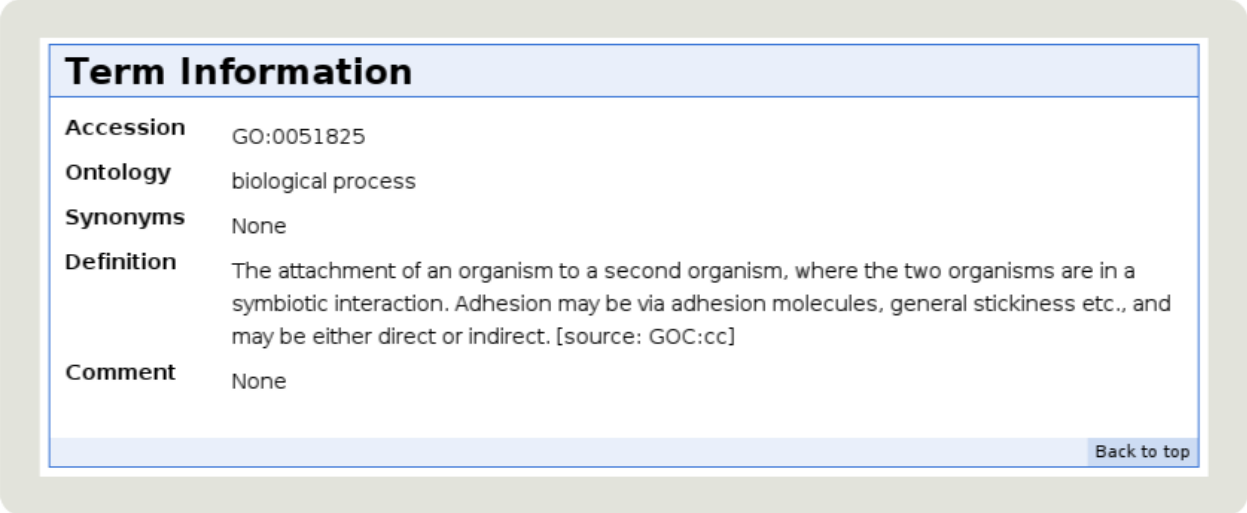
```
setAnnotationCategories(david, c("GOTERM_BP_ALL","KEGG_PATHWAY"))
```

4.2 Obtención de anotaciones funcionales y visualización

- Una vez que se han configurado todos los parámetros, es posible usar las diversas herramientas de DAVID para obtener el análisis funcional de los genes.

```
resul_funcional_anotacion_cuadro<-getFunctionalAnnotationChart(david)
resul_funcional_anotacion_tabla<-getFunctionalAnnotationTable(david)
resul_cluster<-getClusterReport(david)
resul_resumen_anotacion<-getAnnotationSummary(david)
```

- Habitualmente estas herramientas ofrecen información en base a ontologías biológicas, siendo el principal proyecto Gene Ontology (GO) el cual está compuesto de tres ontologías: procesos biológicos, componentes celulares y funciones moleculares. Cada gen, puede tener varios términos ('term') biológicos de cada ontología, y cada categoría de anotación está asociada a varias ontologías, de este modo un ejemplo de términos es mostrado en la figura 2.



Term Information

Accession	GO:0051825
Ontology	biological process
Synonyms	None
Definition	The attachment of an organism to a second organism, where the two organisms are in a symbiotic interaction. Adhesion may be via adhesion molecules, general stickiness etc., and may be either direct or indirect. [source: GOC:cc]
Comment	None

Back to top

Figura 1.- Ejemplo de Términos encontrados en GO.

4.2.1 Functional Annotation Chart

Esta herramienta ofrece anotaciones en forma de tabla con varios índices estadísticos, y donde las filas corresponden a sublistas de genes que están presentes en diferentes términos biológicos de las categorias consultadas.

Functional Annotation Chart

Help and Manual

Current Gene List: genesDAVID
Current Background: Homo sapiens
1089 DAVID IDs
418 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	LT	PH	PT	%	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR	Fisher Exact
	GOTERM_BP_ALL	phosphorylation	RT		78	809	800	14116	7,2	4,1E-6	1,7	1,3E-2	1,3E-2	7,5E-3	2,3E-6
	GOTERM_BP_ALL	regulation of developmental process	RT		68	809	674	14116	6,2	6,3E-6	1,8	1,9E-2	9,8E-3	1,1E-2	3,4E-6
	GOTERM_BP_ALL	protein kinase cascade	RT		44	809	370	14116	4,0	7,5E-6	2,1	2,3E-2	7,7E-3	1,4E-2	3,3E-6
	GOTERM_BP_ALL	cellular process	RT		655	809	10541	14116	60,1	8,8E-6	1,1	2,7E-2	6,9E-3	1,6E-2	8,1E-6
	GOTERM_BP_ALL	cellular component organization	RT		190	809	2498	14116	17,4	1,4E-5	1,3	4,3E-2	8,8E-3	2,6E-2	1,0E-5
	GOTERM_BP_ALL	regulation of cell morphogenesis	RT		22	809	131	14116	2,0	1,6E-5	2,9	5,0E-2	8,5E-3	3,0E-2	4,9E-6
	GOTERM_BP_ALL	regulation of anatomical structure morphogenesis	RT		30	809	219	14116	2,8	2,1E-5	2,4	6,3E-2	9,3E-3	3,8E-2	7,9E-6
	GOTERM_BP_ALL	protein amino acid phosphorylation	RT		65	809	667	14116	6,0	3,0E-5	1,7	9,0E-2	1,2E-2	5,5E-2	1,7E-5
	GOTERM_BP_ALL	phosphorus metabolic process	RT		85	809	973	14116	7,8	8,6E-5	1,5	2,3E-1	2,9E-2	1,5E-1	5,3E-5
	GOTERM_BP_ALL	phosphate metabolic process	RT		85	809	973	14116	7,8	8,6E-5	1,5	2,3E-1	2,9E-2	1,5E-1	5,3E-5
	GOTERM_BP_ALL	post-translational protein modification	RT		99	809	1182	14116	9,1	1,0E-4	1,5	2,7E-1	3,1E-2	1,8E-1	6,7E-5
	GOTERM_BP_ALL	enzyme linked receptor protein signaling pathway	RT		38	809	342	14116	3,5	1,5E-4	1,9	3,6E-1	4,0E-2	2,6E-1	6,9E-5

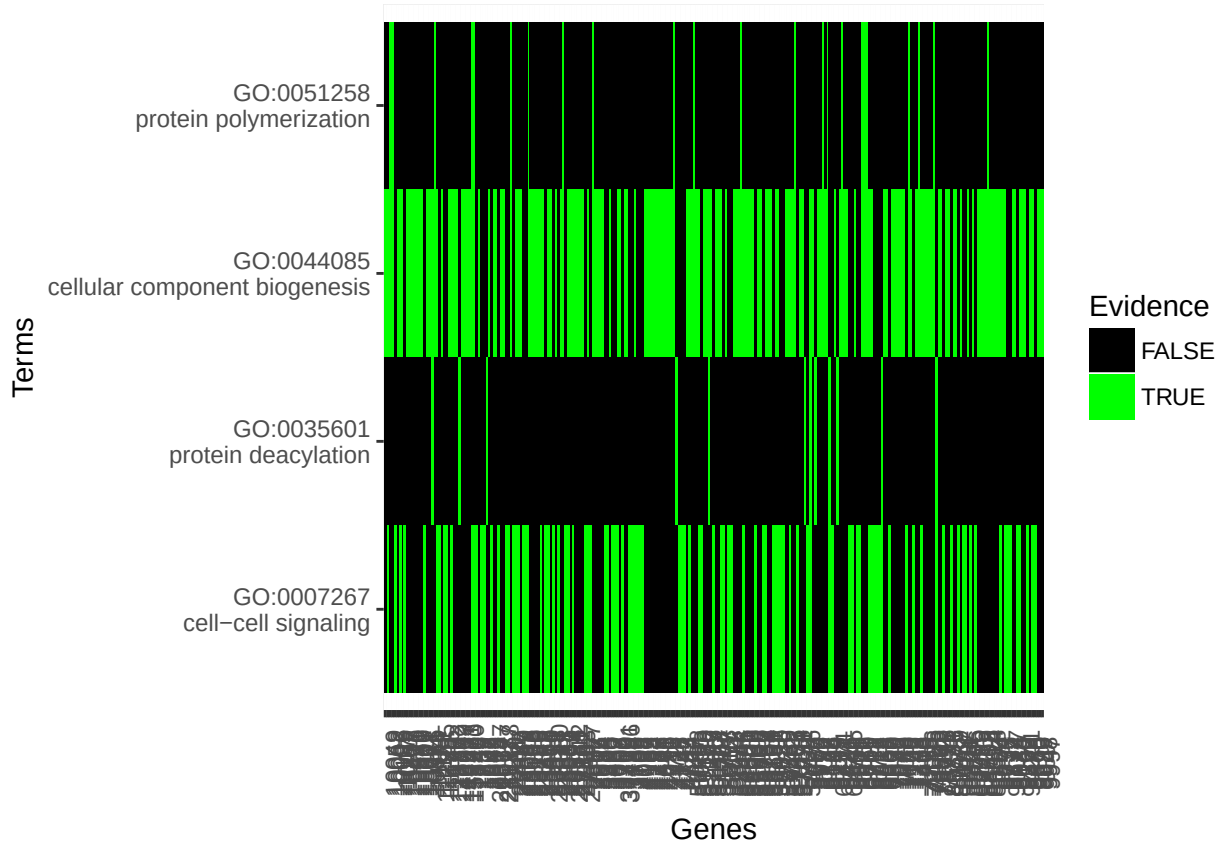
Figure 2.- Ejemplo de Functional Annotation Chart.

- Para visualizar los genes que aparecen en cada 'term' use el comando

```
ids(resul_funcional_anotacion_cuadro)
```

- Ahora usaremos el comando de 'plot2D" de la librería **ggplot** para visualizar las relaciones biológicas entre términos de las categorías de anotación y los genes de la lista. El color verde corresponde a la presencia de cierto gen en cierto término biológico, y el color negro a la ausencia.

```
plot2D(DAVIDFunctionalAnnotationChart(resul_funcional_anotacion_cuadro[388:391, ]),
        color=c("FALSE"="black", "TRUE"="green"))
```



4.2.2 Functional Annotation Table

Genera una tabla con funciones biológicas de las categoría de anotación que involucran a los genes de la lista.

Functional Annotation Table

Help and Manual

Current Gene List: genesDAVID
Current Background: Homo sapiens
1089 DAVID IDs
819 record(s)

Download File

4939	2'-5'-oligoadenylate synthetase 2, 69/71kDa	Related Genes	Homo sapiens
GOTERM_BP_ALL	immune system process, nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, RNA catabolic process, nitrogen compound metabolic process, immune response, metabolic process, catabolic process, macromolecule catabolic process, cellular process, RNA metabolic process, cellular nitrogen compound metabolic process, macromolecule metabolic process, cellular metabolic process, primary metabolic process, cellular catabolic process, cellular macromolecule metabolic process, cellular macromolecule catabolic process, response to stimulus,		
2531	**3-ketodihydrosphingosine reductase**	**Related Genes**	**Homo sapiens**
GOTERM_BP_ALL	metabolic process, oxidation reduction,		
KEGG_PATHWAY	Sphingolipid metabolism,		
64343	**5-azacytidine induced 2**	**Related Genes**	**Homo sapiens**
GOTERM_BP_ALL	signal transduction, intracellular signaling cascade, protein kinase cascade, I-kappaB kinase/NF-kappaB cascade, regulation of biological process, regulation of cellular process, biological regulation,		
KEGG_PATHWAY	RIG-I-like receptor signaling pathway,		
8745	**ADAM metallopeptidase domain 23**	**Related Genes**	**Homo sapiens**
GOTERM_BP_ALL	proteolysis, cell adhesion, cell surface receptor linked signal transduction, integrin-mediated signaling pathway, multicellular organismal development, nervous system development, central nervous system development, metabolic process, cellular process, protein metabolic process, biological adhesion, multicellular organismal process, developmental process, macromolecule metabolic process, primary metabolic process, system development, anatomical structure development,		
377	**ADP-ribosylation factor 3**	**Related Genes**	**Homo sapiens**
GOTERM_BP_ALL	transport, signal transduction, intracellular signaling cascade, small GTPase mediated signal transduction, protein localization, cellular process, protein transport, vesicle-mediated transport, macromolecule localization, establishment of protein localization, regulation of biological process, regulation of cellular process, localization, establishment of localization, biological regulation,		
4299	**AF4/FMR2 family, member 1**	**Related Genes**	**Homo sapiens**
GOTERM_BP_ALL	regulation of transcription, DNA-dependent, regulation of biosynthetic process, positive regulation of biosynthetic process, positive regulation of metabolic process, regulation of gene expression, regulation of macromolecule biosynthetic process, positive regulation of macromolecule biosynthetic process, positive regulation of macromolecule metabolic process, positive regulation of gene expression, regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, regulation of metabolic process, regulation of cellular metabolic process, positive regulation of cellular metabolic process, regulation of cellular biosynthetic process, positive regulation of cellular biosynthetic process, regulation of transcription, positive regulation of transcription, DNA-dependent, positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, positive regulation of transcription, positive regulation of biological process, positive regulation of cellular process, regulation of biological process, regulation of cellular process, regulation of nitrogen compound metabolic process, positive regulation of nitrogen compound metabolic process, regulation of RNA metabolic process, positive regulation of RNA metabolic process, regulation of macromolecule metabolic process, biological regulation, regulation of primary metabolic process,		

Figura 3.- Ejemplo de Functional Annotation Table.

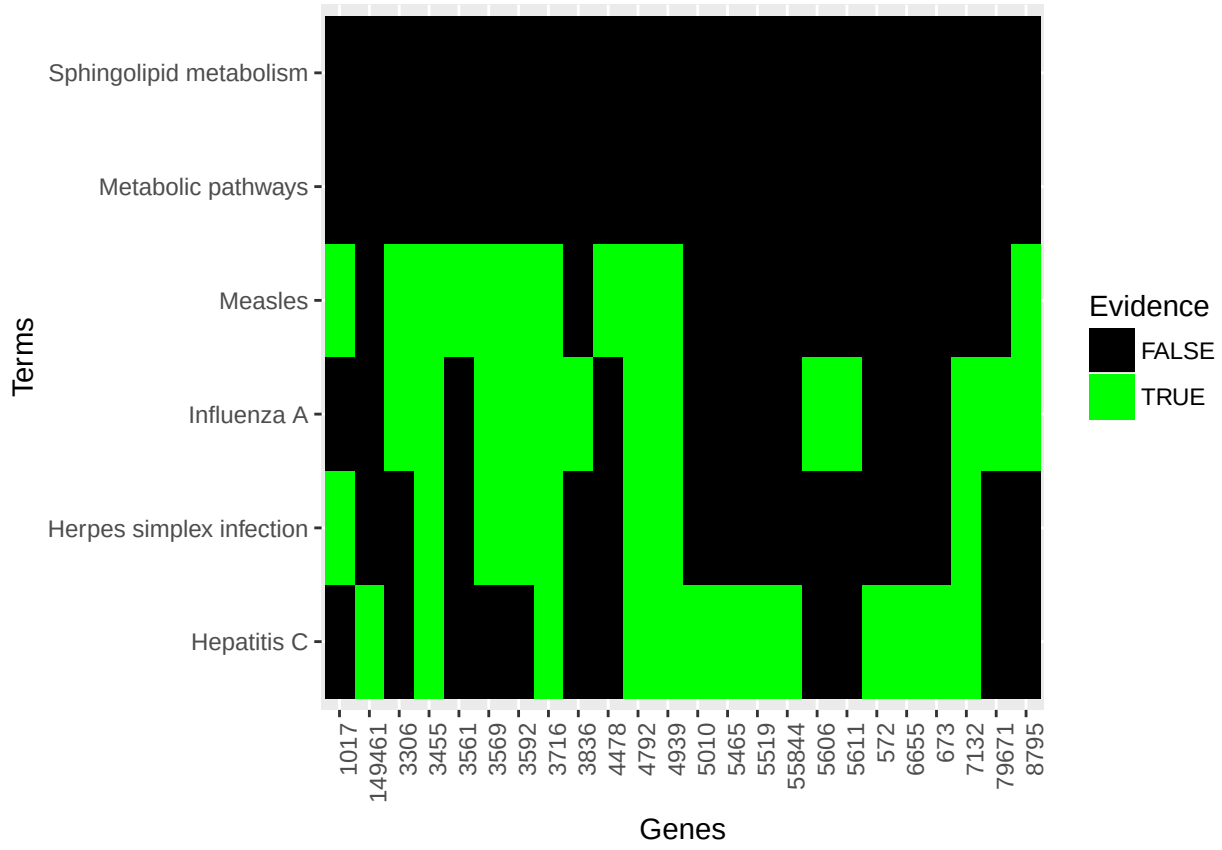
Sobre los resultados obtenidos por esta herramienta, aplicaremos procesos similares al anterior pero filtrando la categoría de anotación a sólo *KEGG_PATHWAY* y generando un 'plot2D" que muestre los genes y los términos pero sin código de la categoría de anotación.

```
#A cada término en categoria KEGG_PATHWAY extraemos el código asociado.
#El 2 indica que es cat. KEGG_PATHWAY
categorySelection<-list(head(
  dictionary(resul_funcional_anotacion_tabla,
             categories(resul_funcional_anotacion_tabla)[2])$ID)
                        )
#y lo agregamos como name de la categoria KEGG_PATHWAY
names(categorySelection)<-categories(resul_funcional_anotacion_tabla)[2]

#Con membership se obtiene una matriz binaria que indica que término de
#KEGG_PATHWAY se encuentra cada gen.
#[,1:3] indica sólo escogo los tres primeros grupos de funciones biológicas.
id<-membership(resul_funcional_anotacion_tabla,
               categories(resul_funcional_anotacion_tabla)[2])[,1:3]

#Se escogo los ID de aquellos genes que se encuentren en los términos de KEGG_PATHWAY
id<-ids(genes(resul_funcional_anotacion_tabla))[rowSums(id)>0]

plot2D(resul_funcional_anotacion_tabla, category=categorySelection, id=id,
       names.category=TRUE)
```



4.2.3 Functional Annotation Clustering

Genera grupo de anotaciones según las coincidencias de términos biólogicos y genes involucrados en ellos.

Functional Annotation Clustering

Help and Manua

Current Gene List: genesDAVID
Current Background: Homo sapiens
1089 DAVID IDs
168 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 3.43			Count	P_Value	Benjami
	GOTERM_BP_ALL	phosphorylation	RT		78	4.1E-6	1.3E-2
	GOTERM_BP_ALL	protein amino acid phosphorylation	RT		65	3.0E-5	1.2E-2
	GOTERM_BP_ALL	phosphate metabolic process	RT		85	8.6E-5	2.9E-2
	GOTERM_BP_ALL	phosphorus metabolic process	RT		85	8.6E-5	2.9E-2
	GOTERM_BP_ALL	post-translational protein modification	RT		99	1.0E-4	3.1E-2
	GOTERM_BP_ALL	protein modification process	RT		113	5.6E-4	9.2E-2
	GOTERM_BP_ALL	biopolymer modification	RT		114	2.2E-3	1.5E-1
	GOTERM_BP_ALL	cellular protein metabolic process	RT		158	1.9E-2	3.5E-1
	GOTERM_BP_ALL	protein metabolic process	RT		179	6.7E-2	5.6E-1
	Annotation Cluster 2	**Enrichment Score: 2.45**			**Count**	**P_Value**	**Benjami**
	GOTERM_BP_ALL	regulation of developmental process	RT		68	6.3E-6	9.8E-3
	GOTERM_BP_ALL	regulation of cell differentiation	RT		46	1.3E-3	1.4E-1
	GOTERM_BP_ALL	negative regulation of developmental process	RT		28	1.9E-3	1.4E-1
	GOTERM_BP_ALL	negative regulation of cell differentiation	RT		22	1.2E-2	3.0E-1
	GOTERM_BP_ALL	positive regulation of developmental process	RT		26	1.7E-2	3.4E-1

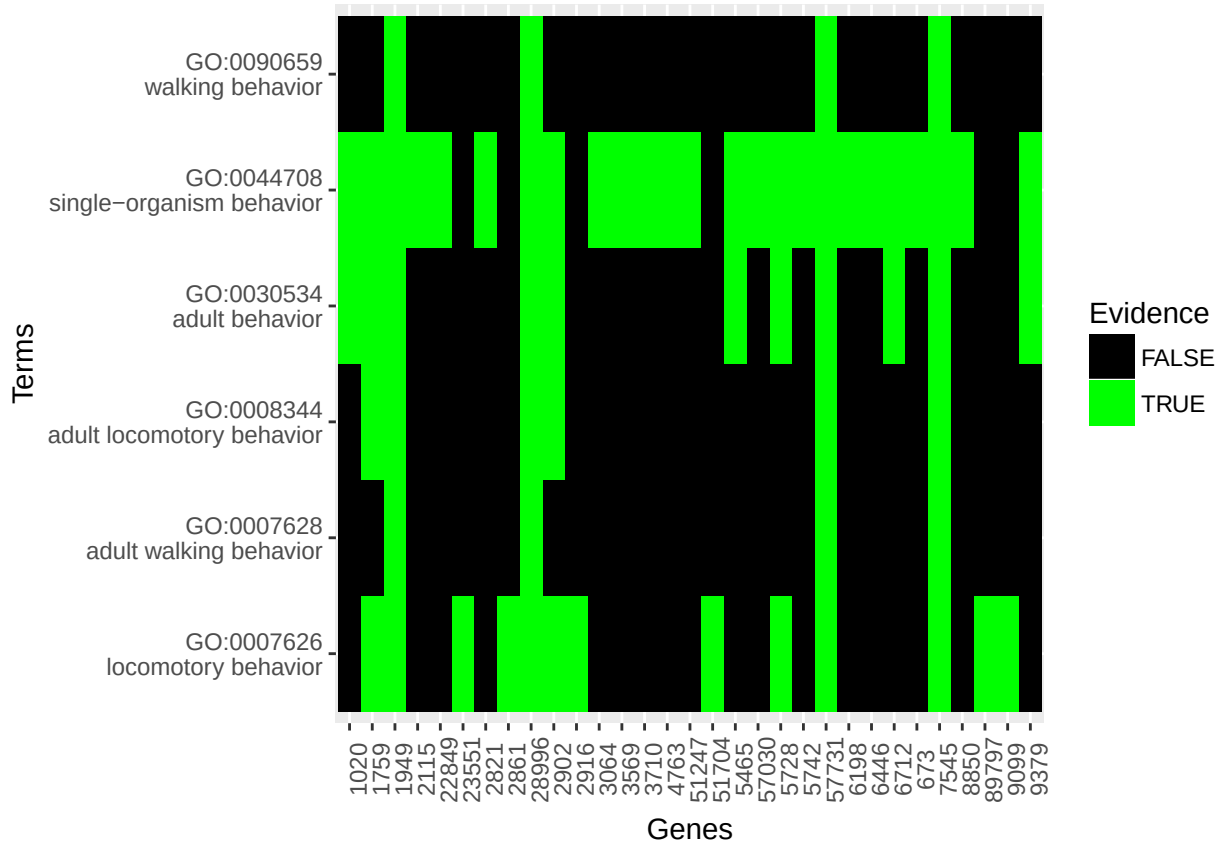
Figure 4.- Ejemplo de Functional Annotation Clustering.

- Obtenga información sobre los grupos.

```
cluster(resul_cluster)
summary(resul_cluster)
```

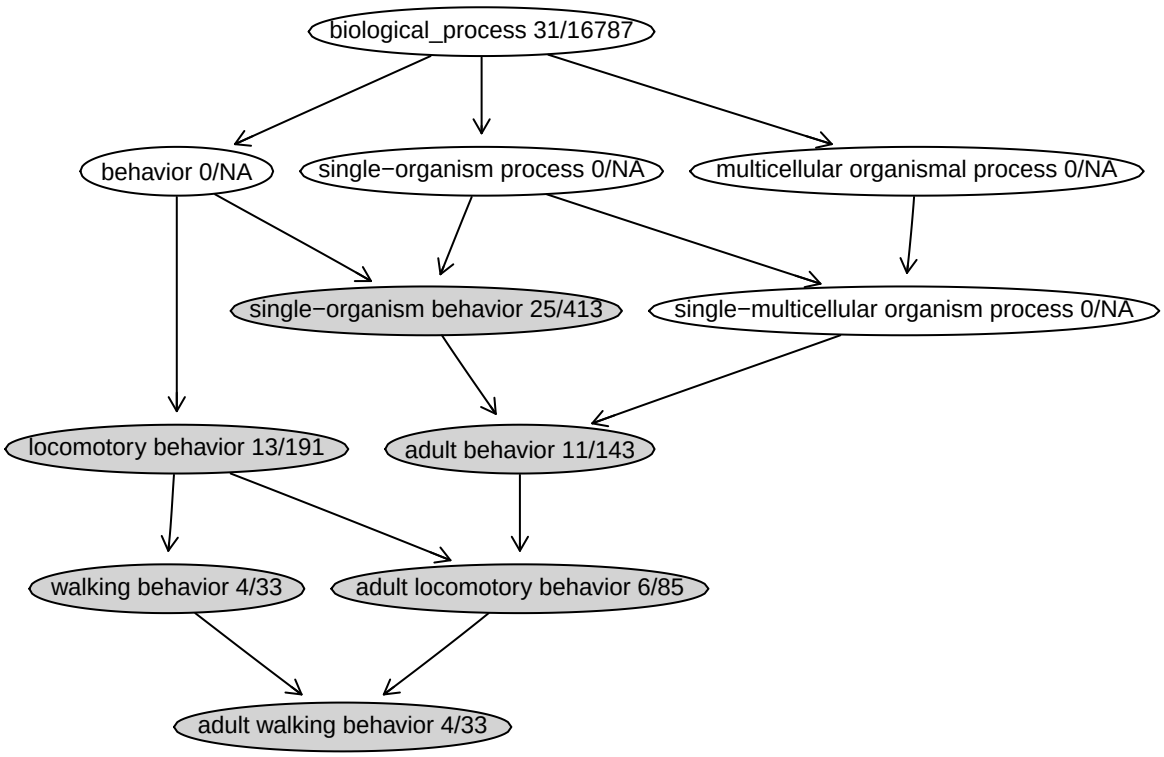
- Genere un 'plot2d"" del grupo 97.

```
clustNumber=97
plot2D(resul_cluster, clustNumber)
```



- Ahora generamos un grafo que representa la relación entre los términos del grupo 97. Los números a la derecha de cada nodo la relación entre los genes en la lista vs. los de referencia de la categoría de anotación. Si no hay información asociada entonces NAs. Nodos grises son aquellos cuyo *EASE Score* es $<$ *pvalueCutoff*.

```
davidGODag<-DAVIDGODag(members(resul_cluster)[[clustNumber]], pvalueCutoff=0.90)
 plotGOTermGraph(g=goDag(davidGODag), r=davidGODag, max.nchar=40, node.shape="ellipse")
```



5 Actividad

Con los genes diferencialmente expresados encontrados en la ayudantía anterior realice un análisis funcional. Para ello considere sólo uno de los métodos de selección de genes.