TALLER 8: USO DE R PARA LA EJECUCIÓN DE ANÁLISIS DE ENRIQUECIMIENTO FUNCIONAL

Noviembre 25 de 2021

Chrystian C. Sosa Y Mauricio A. Quimbaya

Doctorado en ingeniería y ciencias aplicadas

El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA

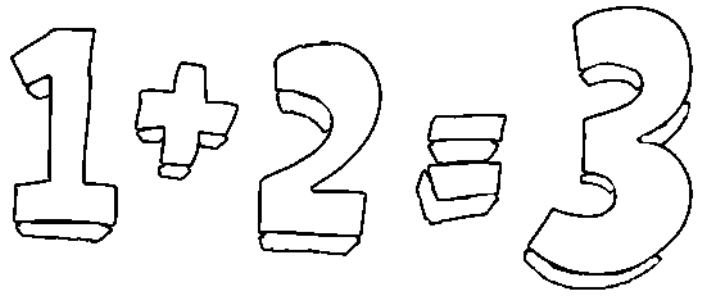
Contenido

- ¿Qué es un análisis de enriquecimiento funcional y qué es un grafo?
- Realización de un análisis de enriquecimiento funcional usando una lista de genes.
- Análisis de enriquecimiento funcional en R.
- Análisis de enriquecimiento funcional por categorías y entre dos especies usando R.

¿Qué es un análisis de enriquecimiento funcional y qué es un grafo?

El contexto de los análisis bioinformáticos

¿Qué quisiéramos tener?



¿Qué tenemos en biología?

¿Cómo se puede saber que funciones tiene una proteína?

- Una gran proporción de proteínas estan pobremente caracterizadas experimentalmente
- Homología
- Similitud estructural

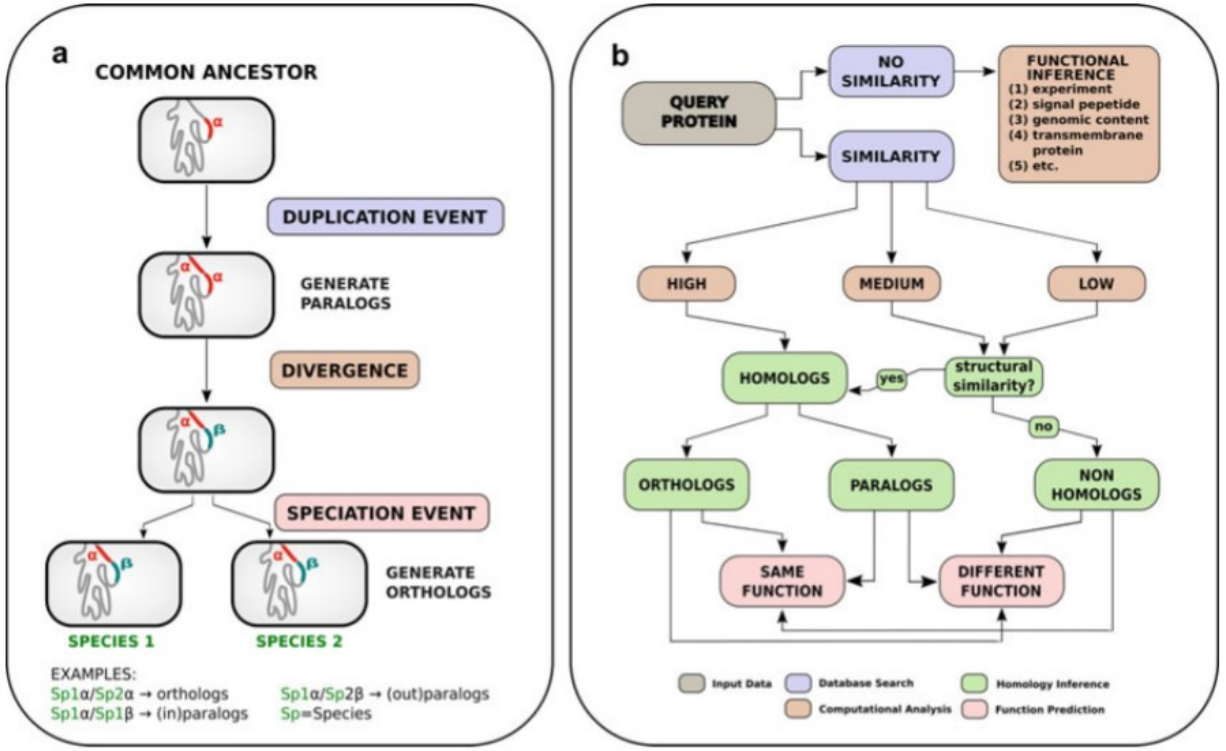


Fig. 1 Sequence similarity and homology in protein function prediction. Flowcharts summarizing (**a**) basic concepts on homology and sequence divergence and (**b**) possible strategies in protein annotation using sequence similarity

Cruz et al., 2017

¿Qué necesitamos para poder comprender la complejidad de funciones biológicas de una proteína?

¿Ideas?

¿Qué es una función biológica?

- Mecanismo: qué partes contribuyen a un sistema
  - Entidades y actividades
  - ¿Qué es una entidad?: Proteína
  - ¿Qué es una actividad?: Traducción

- Bajo la visión de un biólogo molecular:
  - Actividades coordinadas que tienen la apariencia de haber sido diseñadas por un ***propósito***

El proyecto *gene ontology*

ontología

Del lat. mod. *ontologia*, de *onto-* 'onto-' y *-logia* '-logía'.

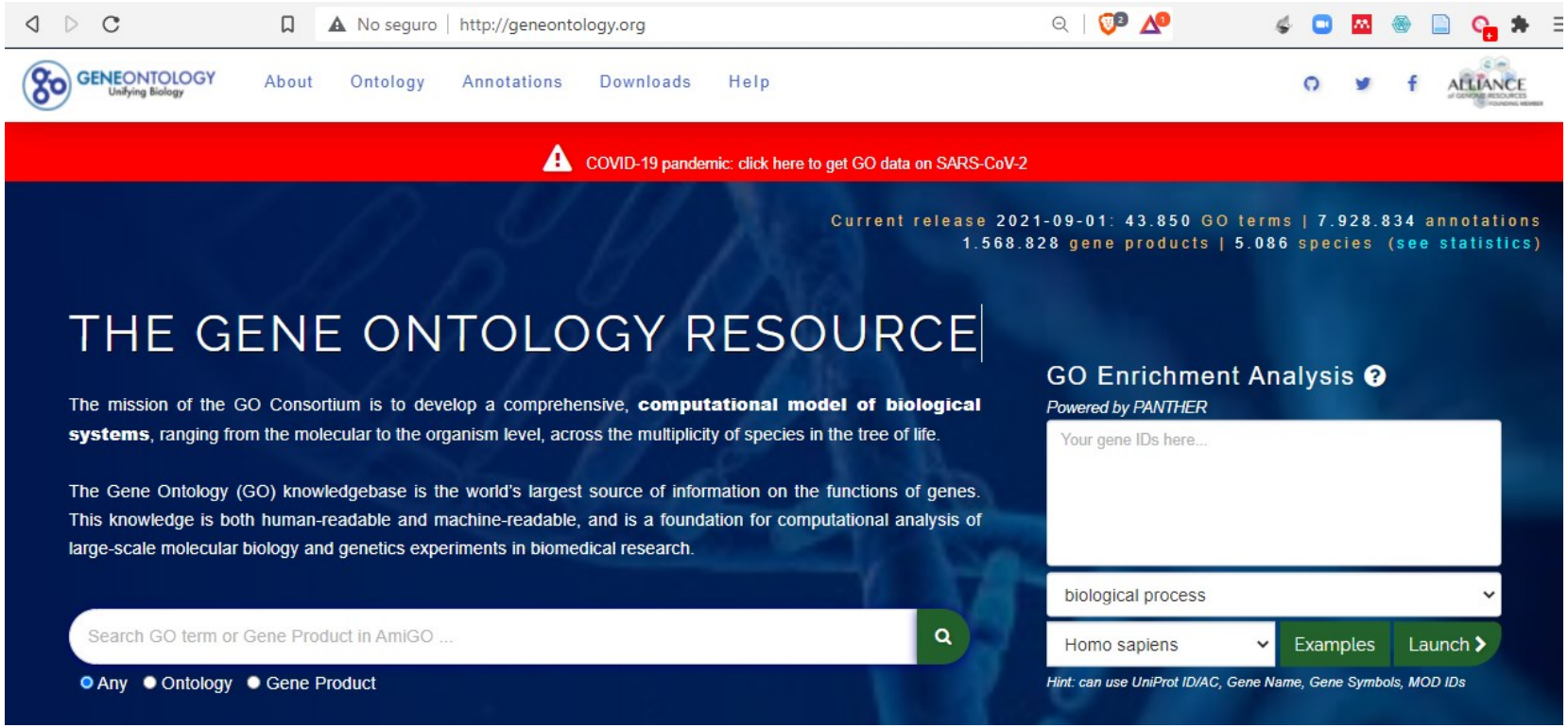
1. f. *Fil.* Parte de la metafísica que trata del ser en general y de sus propiedades trascendentales.

2. f. En ciencias de la comunicación y en inteligencia artificial, red o sistema de datos que define las relaciones existentes entre los conceptos de un dominio o área del conocimiento.

Gene Ontology: tool for the unification of biology

The Gene Ontology Consortium*

- Lenguaje estructurado
- Base de datos estructurada
- http://geneontology.org/



Ashburner et al.,2000

¿Qué sabemos de los términos de ontología (GO)?

- Cambian con el tiempo
- Existen métodos automáticos y curación manual

Statistics for release 2021-09

Ontology

Property	Value
Valid terms	43850 (Δ = -28)
Obsoleted terms	3378 (Δ = 27)
Merged terms	2335 (Δ = 9)
Biological process terms	28503
Molecular function terms	11168
Cellular component terms	4179

Annotations

Property	Value
Number of annotations	7,928,834
Annotations for biological process	2,970,944
Annotations for molecular function	2,538,694
Annotations for cellular component	2,419,196
Annotations for evidence PHYLO	3,886,746
Annotations for evidence IEA	2,020,599
Annotations for evidence OTHER	829,676
Annotations for evidence EXP	875,954
Annotations for evidence ND	260,621
Annotations for evidence HTP	55,238
Number of annotated scientific publications	165,158

Gene products and species

Property	Value
Annotated gene products	1,568,828
Annotated species	5,086
Annotated species with over 1,000 annotations	202

Statistics for release 2021-10

Ontology

Property	Value
Valid terms	43832 (Δ = -18)
Obsoleted terms	3394 (Δ = 16)
Merged terms	2348 (Δ = 25)
Biological process terms	28484
Molecular function terms	11166
Cellular component terms	4182

Annotations

Property	Value
Number of annotations	7,827,476
Annotations for biological process	2,950,031
Annotations for molecular function	2,525,429
Annotations for cellular component	2,352,016
Annotations for evidence PHYLO	3,782,649
Annotations for evidence IEA	2,019,391
Annotations for evidence OTHER	830,737
Annotations for evidence EXP	880,087
Annotations for evidence ND	259,029
Annotations for evidence HTP	55,583
Number of annotated scientific publications	165,816

Gene products and species

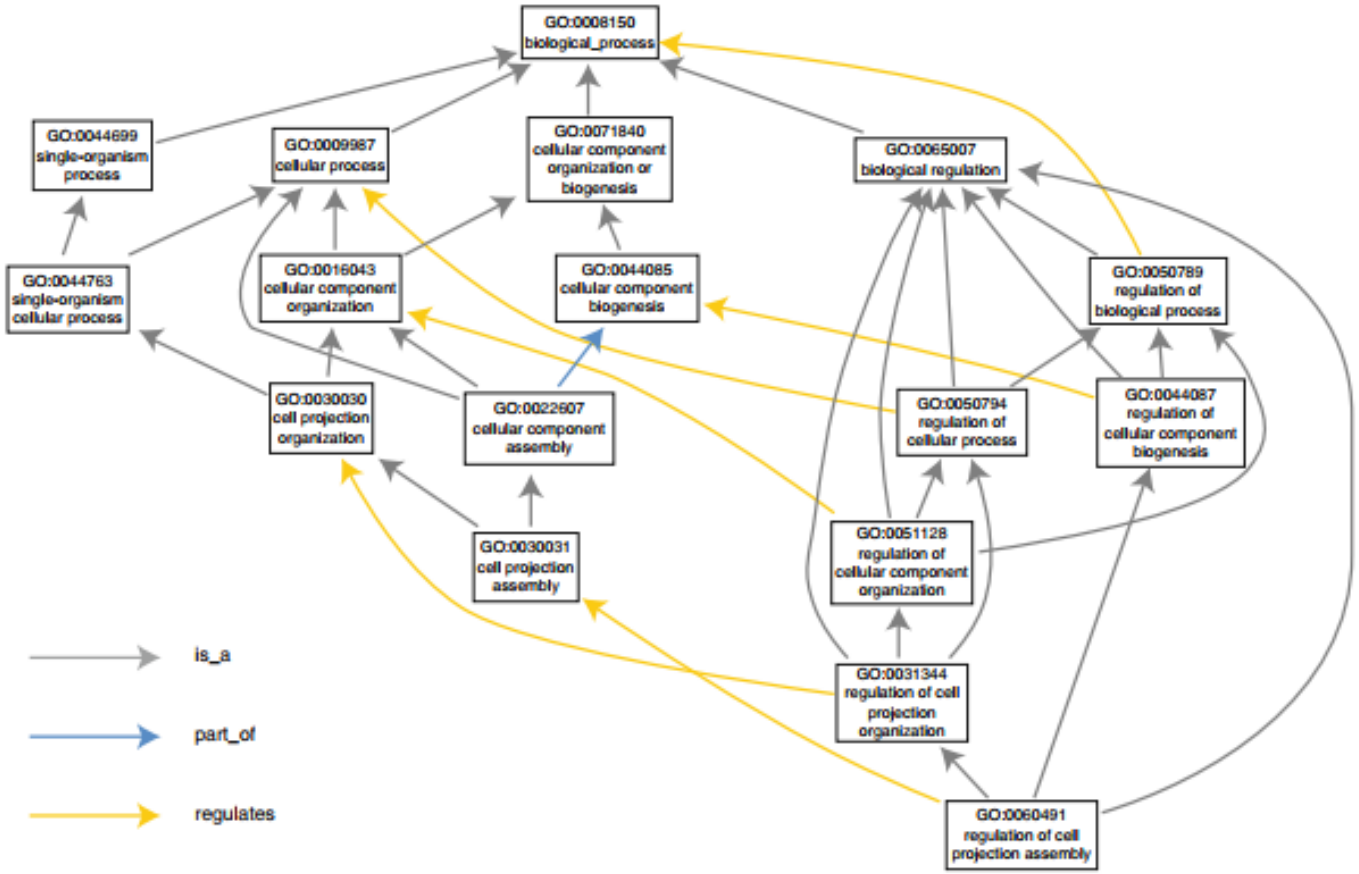
Property	Value
Annotated gene products	1,542,582
Annotated species	5,086
Annotated species with over 1,000 annotations	201

http://geneontology.org/stats.html

Gene ontology

¡La representación de una ontología de genes es un grafo!

- Tres aspectos son representados:
  - Función molecular
  - Proceso biológico
  - Componente celular
- No son redundantes
- Estructura de vocabulario en forn jerárquica
  - Los nodos son términos
  - Las aristas son las relaciones



Gaudet et al.,2017

Función de un GO

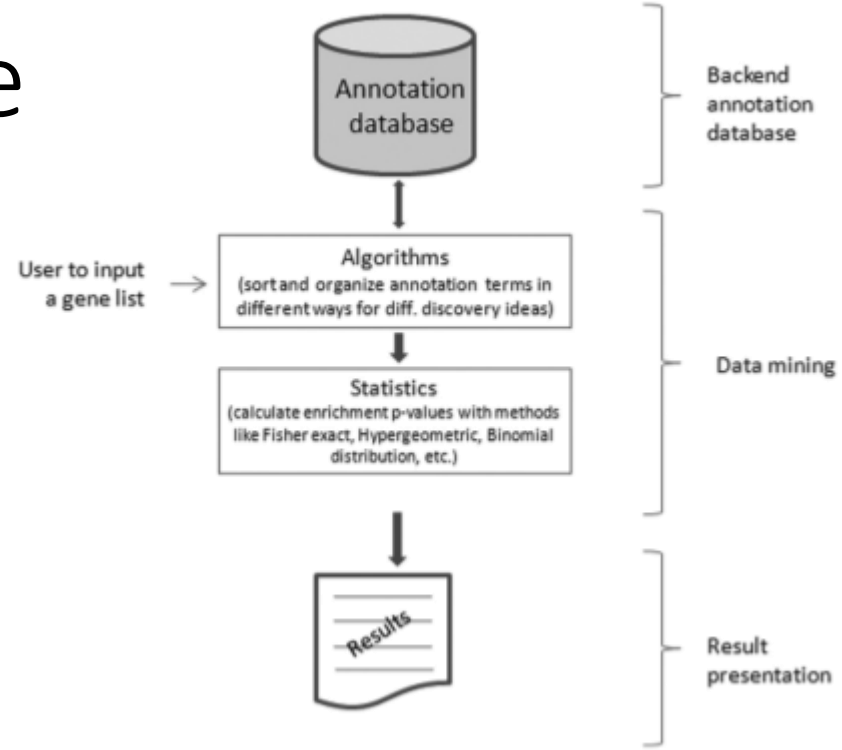
- **Gen**: Es una región contigua de ADN que codifica instrucciones para como la célula genera macromoléculas
- **Macromolécula**: Producto de un gen (Proteína o ARN no codificante)
- Un producto de un gen puede actuar como una máquina molecular para desarrollar una acción química (**actividad**)
- Los productos de genes de diferentes genes se pueden combinar en máquinas moleculares mas grandes (**complejos**)

GO define el posible universo de funciones de un gen particular!

- **Función molecular**: actividad o proceso a nivel molecular
- **Componente celular**: localización relativa en la célula
- **Proceso biológico**: procesos moleculares vinculados a un objetivo biológico

¿Cómo funciona un algoritmo de enriquecimiento funcional?

- Hasta el año 2009 habían más de 68 herramientas disponibles
- Capas:
  - Datos de soporte
  - Algoritmo de minería de datos incluido estadística
  - Presentación de datos



Published online 25 November 2008 *Nucleic Acids Research, 2009, Vol. 37, No. 1* **1–13**
doi:10.1093/nar/gkn923

SURVEY AND SUMMARY

Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists

Da Wei Huang, Brad T. Sherman and Richard A. Lempicki*

Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD 21702, USA

Huang et al., 2009

Table 2. Categorization of enrichment analysis tools

Tool category	Description	Indication and limitation	Sub-type of algorithms	Methods	Example tool
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.	Capable of analyzing any gene list, which could be selected from any high-throughput biological studies/technologies (e.g. Microarray, ChIP-on-CHIP, ChIP-on-sequence, SNP array, EXON array, large scale sequence, etc.). However, the deeper inter-relationships among the terms may not be fully captured in linear format report.	Global reference background	Fisher's exact hypergeometric chi-square binomial	GoStat, GoMiner, GOTM, BinGO, GOtoolBox, GFinder, etc.
			Local reference background	Fisher's Exact hypergeometric chi-square binomial	DAVID, Onto-Express, GARBAN, FatiGO, etc.
			Neural network	Bayesian	BayGO
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.	Suitable for pair-wide biological studies (e.g. disease versus control). Currently, may be difficult to be applied to the diverse data structures derived by a complex experimental design and some of the new technologies (e.g. SNP, EXON, Promoter arrays).	Based on ranked gene list	Kolmogorov–Smirnov-like	GSEA, CapMap, etc.
			Based on continuous gene values	*t*-Test permutation *Z*-score	FatiScan, ADGO, ermineJ, PAGE, iGA, GO-Mapper, GOdist, FINA, T-profiler, MetaGP, etc.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.	Capable of analyzing any gene lists, which could be selected from any high-throughput biological studies/technologies, like Class I. Emphasis on network relationships during analysis. 'Orphan' gene/term (with little relationships to other genes/terms), that sometimes could be very interesting, too, may be left out from the analysis.	Composite annotations	Measure enrichment on joint terms	ADGO, GeneCodis, ProfCom, etc.
			DAG Structure	Measure enrichment by considering parents-child relationships	topGO, Ontologizer, POSOC, etc.
			Global annotation relationship	Measure term–term global similarity with Kappa Statistics Czekanowski-Dice Pearson's correlation	DAVID, GoToolBox, etc.

¿Qué hace el análisis de enriquecimiento funcional? (I)

- Distribución hipergeométrica

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}.$$

- N es el número de genes en la distribución *background* (genes anotados en el genoma)
- M es el número de genes dentro de la distribución que están anotados
- n es el tamaño de la lista de genes de interés
- k es el número de genes de la lista que están anotados

De cada 20 piezas fabricadas por una máquina, hay 2 que son defectuosas. Para realizar un control de calidad, se observan 15 elementos y se rechaza el lote si hay alguna que sea defectuosa. Vamos a calcular la probabilidad de que el lote sea rechazado

$$N = 20$$

$$n = 15$$

$$X = \text{número de piezas defectuosa s de las } 15 \text{ escogidas}$$

$$P(X \geq 1) = 1 - p(X < 1) = 1 - p(X = 0)$$

$$1 - \frac{\binom{2}{0} \cdot \binom{20-2}{15}}{\binom{20}{15}} = 1 - \frac{816}{15504} = \frac{18}{19} = 0{,}947$$

https://bit.ly/3F1hCn7

¿Qué hace el análisis de enriquecimiento funcional? (II) (Test exacto de Fisher)

- Basado en distribución hipergeométrica

	A	Not *A*	Total
In sample	a	b	r_1
Not in sample	c	d	r_2
	c_1	c_2	N

- Posibles muestras $\binom{N}{r_1}$
- Formas en que A esta en una muestra de tamaño c1 $\binom{c_1}{a}$
- Formas en que no esta A esta en una muestra de tamaño N- c1=c2 $\binom{c_2}{b}$
- Formas en que elegir a de As y b que no esta en As $\binom{c_1}{a}\binom{c_2}{b}$

$$A_s = \frac{\binom{c_1}{a}\binom{c_2}{b}}{\binom{N}{r_1}} = \frac{\frac{c_1!}{a!c!} \times \frac{c_2!}{b!d!}}{\frac{N!}{r_1!r_2!}} = \frac{c_1!c_2!r_1!r_2!}{N!a!b!c!d!}$$

(Hoffman., 2019)

Table 13.1 Basis of Fisher's test

	Women	Men	Total
In sample	3	2	5
Not in sample	17	8	25
	20	10	30

Example 13.1

A medical clinic has 30 patients, 20 women and 10 men. A random sample of 5 patients is drawn. What is the probability that there will 2 men?

A sample of 5 patients out of 30 can be chosen in $\binom{30}{5}$ ways $= 142{,}506$ ways.

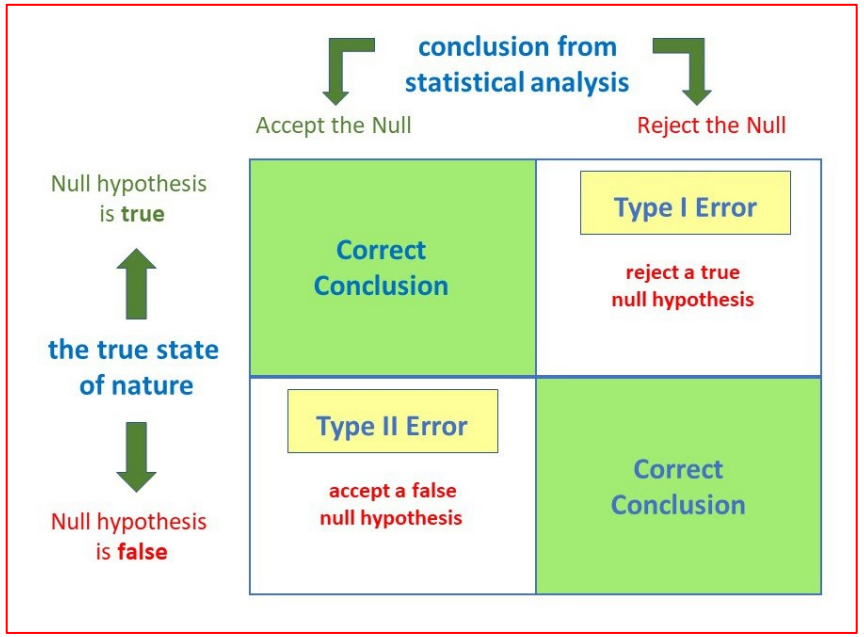
A sample of 2 men and 3 women can be drawn in $\binom{10}{2} \times \binom{20}{3}$ ways $= 51{,}300$ ways.

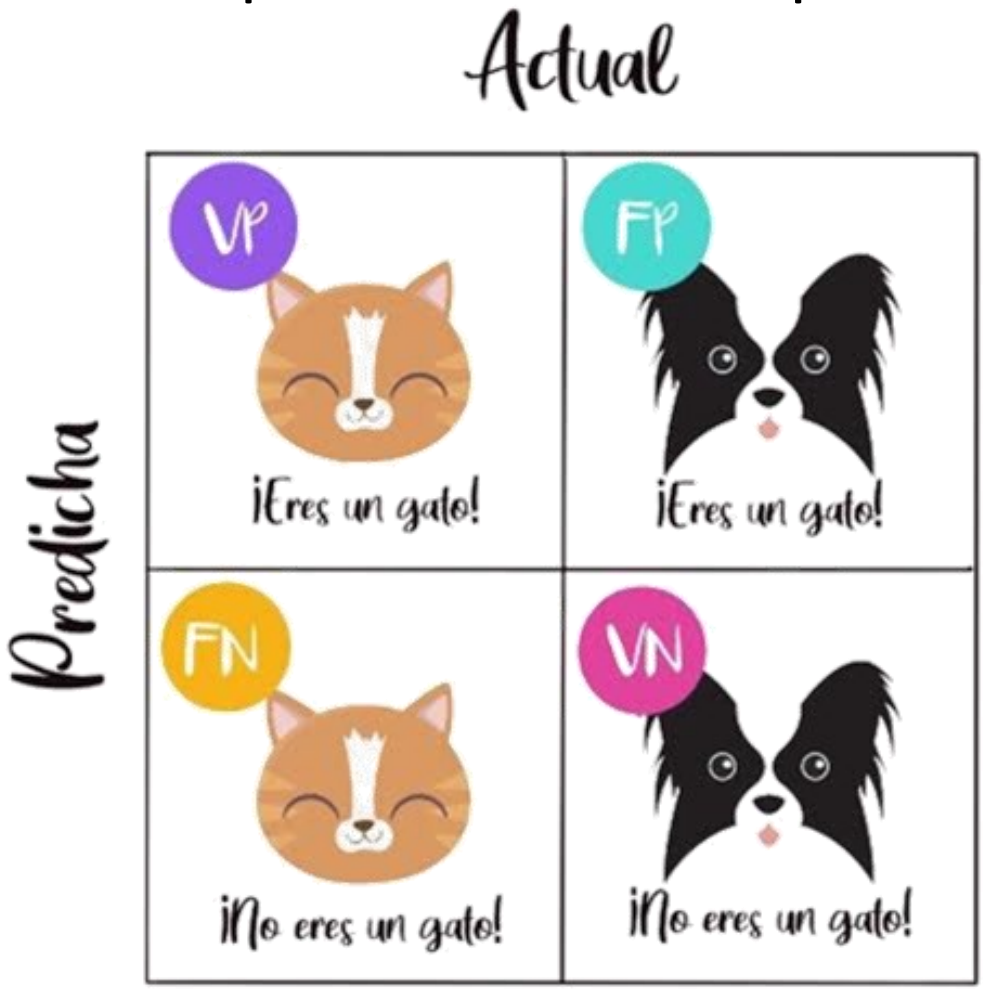
Therefore $P(2 \text{ men, } 3 \text{ women}) = \dfrac{\binom{10}{2} \times \binom{20}{3}}{\binom{30}{5}} = 51{,}300/142{,}506 = 0.359985.$

(Hoffman., 2019)

¿Qué hace el análisis de enriquecimiento funcional? (III) Corrección de hipótesis múltiples

- Hay que reducir el riesgo de errores de omisión y comisión
- Problema de la propagación de genes (Una función compartida entre genes)





https://bit.ly/3mYCgOI;https://bit.ly/3o7AfyU

Bonferroni y *False Discovery Rate*

1. Bonferroni=α/k (punto critico)
2. FDR:
   - Ordenar de la forma:
   - $p_1 \leq p_2 \leq \cdots \leq p_k$
   - Comenzando con el $p_i$mas alto, buscar $p_i \leq \frac{i}{k} * \alpha$ (punto critico)

- Bonferroni es muy conservador

Conservation Genetics (2006) 7:783–787
DOI 10.1007/s10592-005-9056-y

Beyond Bonferroni: Less conservative analyses for conservation genetics

Shawn R. Narum
Columbia River Inter-Tribal Fish Commission, 3059-F National Fish Hatchery Road, Hagerman, ID, 83332, USA (Corresponding author: Phone: +1-208-837-4072; Fax: +1-208-837-6047; E-mail: nars@critfc.org)

Received 4 March 2005; accepted 4 September 2005

Key words: Bonferroni, conservation genetics, false discovery rate, multiple comparison tests

2/15

10/15

	pi	BONFERRONI	FDR			
alpha		P VALUE< VALOR CRITICO	i	pi <=i/k*alpha	FDR	P VALUE< VALOR CRITICO
0.05	0.0001	VERDADERO	1	VERDADERO	0.00333333	VERDADERO
alpha/k	0.001	VERDADERO	2	VERDADERO	0.00666667	VERDADERO
0.00333333	0.0062	FALSO	3	VERDADERO	0.01	VERDADERO
k	0.0101	FALSO	4	VERDADERO	0.01333333	VERDADERO
15	0.0214	FALSO	5	FALSO	0.01666667	VERDADERO
	0.0227	FALSO	6	FALSO	0.02	VERDADERO
	0.0273	FALSO	7	FALSO	0.02333333	VERDADERO
	0.0292	FALSO	8	FALSO	0.02666667	VERDADERO
	0.0311	FALSO	9	FALSO	0.03	VERDADERO
	0.0323	FALSO	10	VERDADERO	0.03333333	VERDADERO
	0.0441	FALSO	11	FALSO	0.03666667	FALSO
	0.049	FALSO	12	FALSO	0.04	FALSO
	0.0573	FALSO	13	FALSO	0.04333333	FALSO
	0.1262	FALSO	14	FALSO	0.04666667	FALSO
	0.5794	FALSO	15	FALSO	0.05	FALSO

Narum.,2006

Trampas en el análisis de enriquecimiento funcional

¡Los GO evolucionan!

Chen *et al. BMC Bioinformatics* (2021) 22:178
https://doi.org/10.1186/s12859-021-04105-8

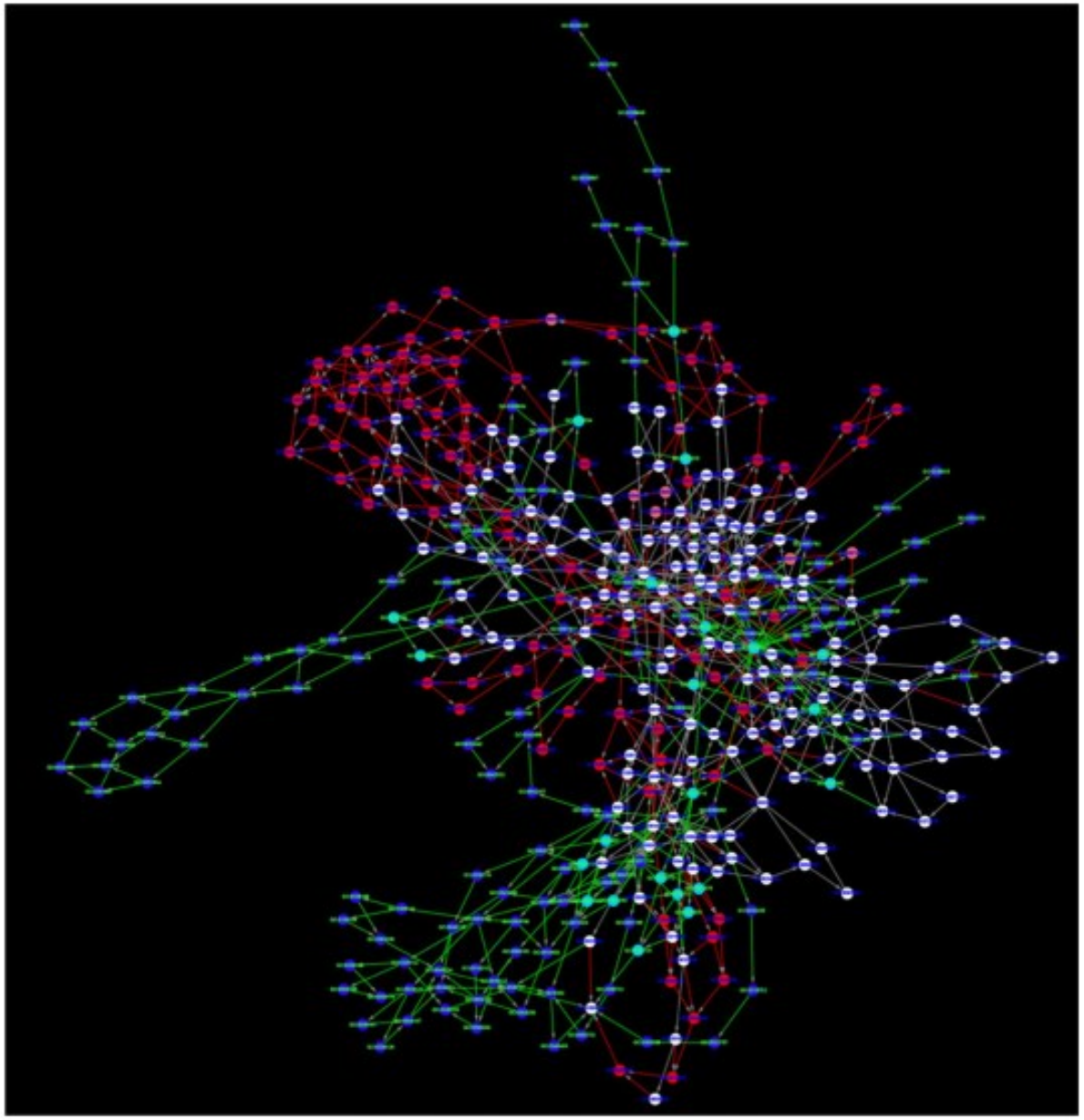
BMC Bioinformatics

RESEARCH ARTICLE Open Access

Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations

Yi Chen*, Fons. J. Verbeek and Katherine Wolstencroft

- ¡ Los *GO terms* cambian en el tiempo!
- Nodos rojos oscuros: existen en 2012 y 2016 pero solo seleccionados por GO2
- Nodos rojos claros: obsoletos en 2016
- Nodos azules oscuros: existen en 2012 y 2016 pero solos seleccionados por GO1
- Nodos azules claros: Nuevos nodos creados



Cheng et al.,2021

Las trampas:

- Paradoja de Simpson: Análisis de datos agregados puede llevar a diferentes conclusiones
- Los datos de GO pueden estar incompletos
- Resultados contradictorios
- Diferencias entre especie (Ejemplo, *zebrafish* tiene más datos relacionados a desarrollo)
- Sesgo del anotador
- Desbalance de anotaciones positivas y negativas
- Conversión entre identificadores de genes o proteínas

Table 1
Main pitfalls or biases discussed in the chapter and their remedies

Pitfall or bias	Remedy
Wrongly assume that absence of annotation implies absence of function.	Account for the fact that both ontology and annotations are necessary incomplete, for instance by assessing the impact of incompleteness on one's analyses and findings.
Not all directed edges in the ontology structure have the same meaning: depending on their type, the relationship they represent may or may not be transitive.	The transitivity of each type of relations must be taken into account when reasoning over the GO. "Is a" and "part of" are transitive, but "regulates" is not.
To yield meaningful results, GO enrichment analyses require accurate specification of the background distribution, which can vary substantially across releases, species, etc.	Specify the actual background distribution used in the analysis of interest. Short of this, ensure that the enrichment analysis is performed on consistent database release and subsets of species, terms, etc. To test the robustness of results, consider repeating the analysis using several releases of GO ontology/annotation databases. Avoid tools that are not regularly updated.
Inter-ontology links and annotation extensions can result in large variations in the number of annotations. Furthermore, annotation extensions may not be consistently implemented, if at all, across analyses tools or workflows.	Keep track of database releases in analyses. If they are relevant, make sure that annotation extensions are implemented consistently.
Qualifiers such as "NOT" or "co-localizes with" are important parts of a gene annotation in that they fundamentally change the meaning of annotations. Because only a small minority of all annotations have qualifiers, such errors can easily go unnoticed.	Remember to take into account qualifiers. When using tools or software libraries, make sure that these take qualifiers into account as well.
Annotations are supported by different types of evidence (categorized by evidence codes). The annotations associated with each code vary in their scope, specificity, and number. These differences can confound some analyses.	Take evidence code into account. In statistical analyses, consider the distribution of annotations in terms of evidence codes, and, if needed, control for this potential confounder.
Different species tend to have very different types of annotations. For instance, model species have many more experiment-based annotations.	When performing statistical analyses or using information-theoretic similarity measures, use species-specific frequencies of GO term.
Experiment-based annotations derived from the same research article tend to be more similar than annotations derived from different articles. Similar trends hold for annotations derived from same versus different authors, and same versus different annotators.	Control for authorship bias in analyses that may have varying proportion of annotations stemming from the same article, lab, or annotation team.
Because annotations are preferentially propagated among closely related sequences, electronic annotations can confound analyses seeking to characterize relationships between evolution and function.	Restrict such analyses to experiment-based annotations. Avoid circularity.
There are many more positive annotations than negative annotations. As a result, standard accuracy measures used by machine learning methods may be misleading ("class imbalance problem").	Consider false-positive and false-negative rates separately. Focus on subset of data for which the class imbalance problem is less pronounced.

¿Qué es un grafo?

- Una red es un catálogo de:
  - Componentes de un sistema (nodos)
  - Interacciones directas (aristas)
- Número de nodos (N): número de componentes (tamaño) i=1,2,…,N
- Número de vínculos (L): número total de interacciones entre nodos

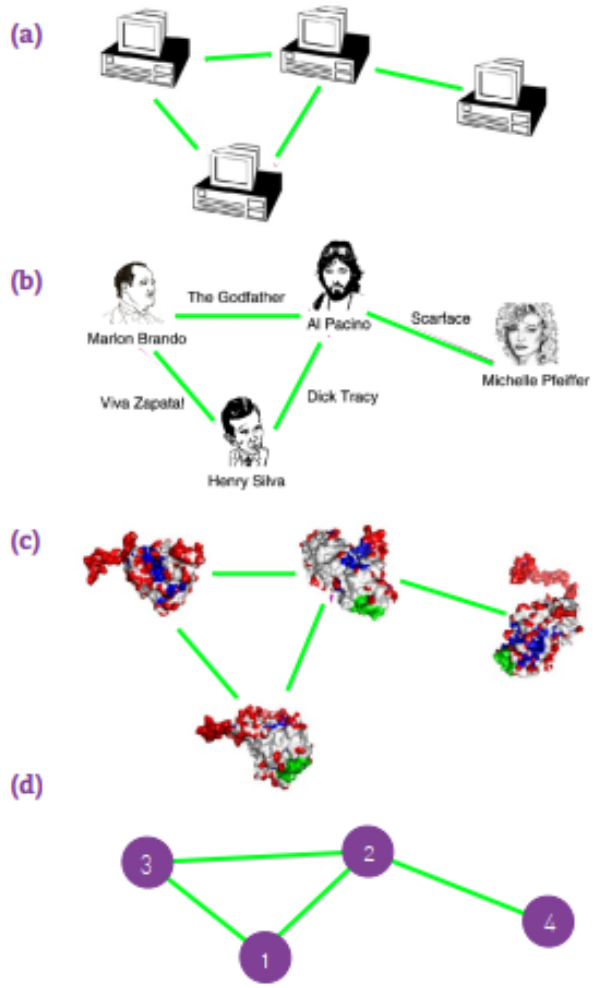
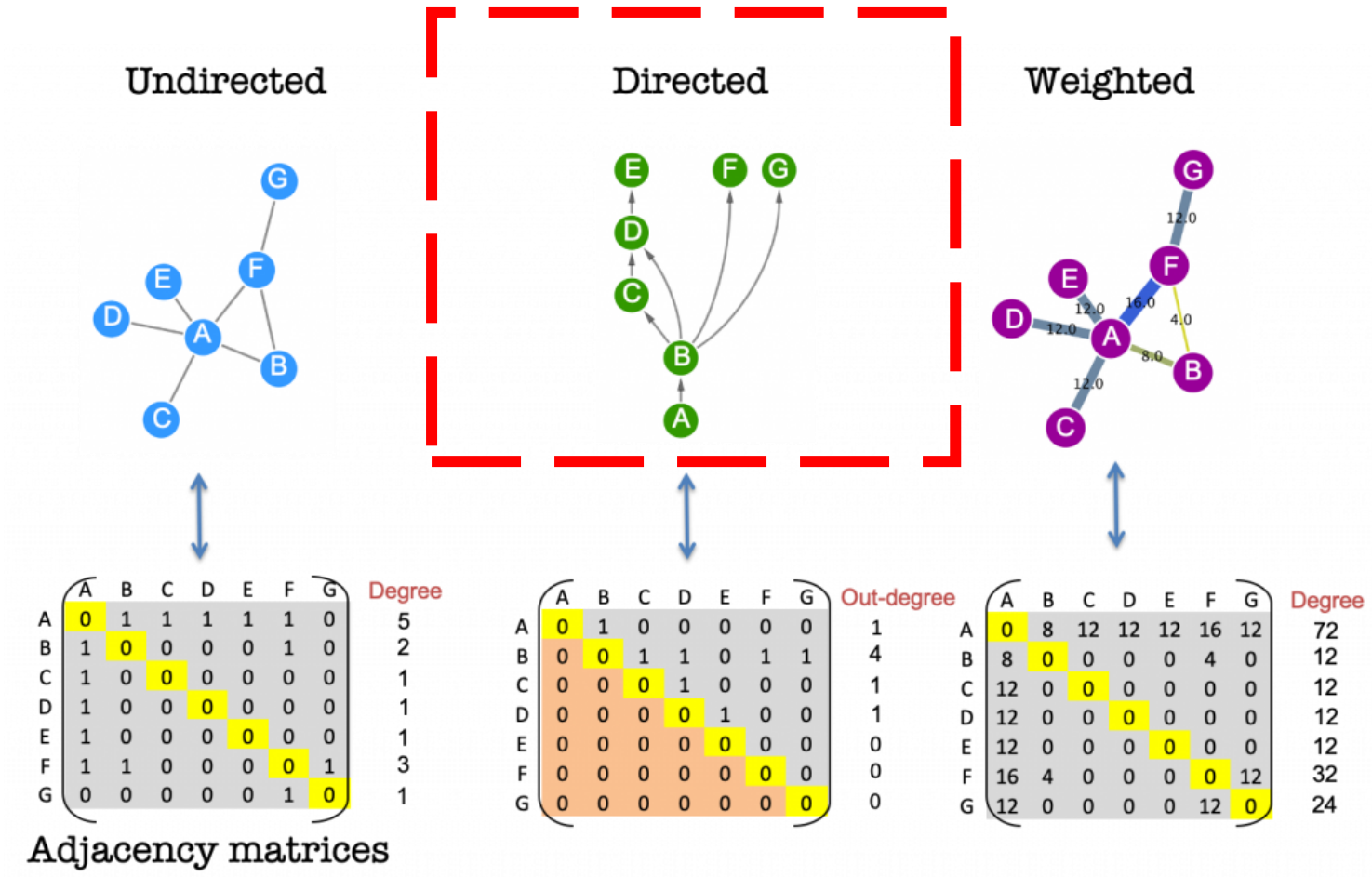
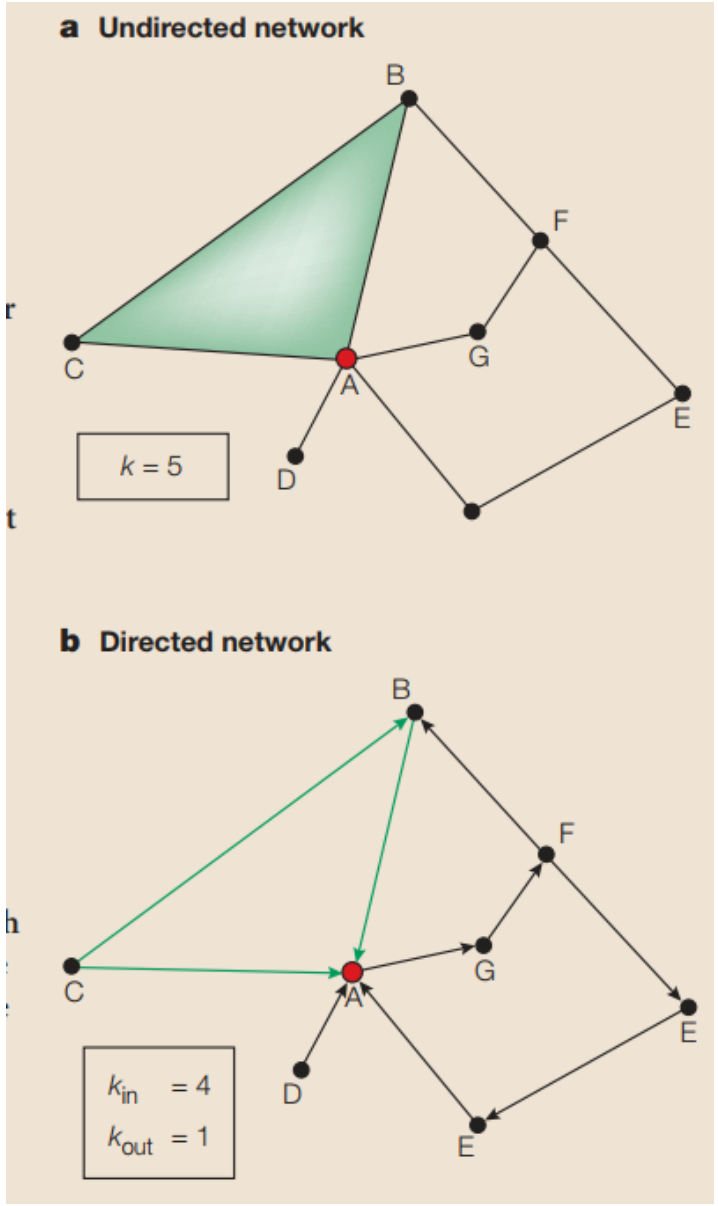


Figure 2.2
Different Networks, Same Graph

Barabasi, 2016

Tipos de grafos





Barabasi et al., 2004; https://bit.ly/3Hbcf6z

Algunos ejemplos de herramientas para enriquecimiento funcional

Herramientas más usadas

- Programas:
  - **BinGO**
  - DAVID
  - **g:Profiler**
  - **AmiGO**
- Paquetes de R:
  - **topGO**
  - **clusterprofiler**
  - **gprofiler2**

TOOLBOX

PROGRAMMING TOOLS: ADVENTURES WITH R

A guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis.

https://www.nature.com/articles/517109a.pdf

Recursos para explorar

- https://yulab-smu.top/biomedical-knowledge-mining-book/index.html
- https://gohandbook.org/doku.php
- https://apps.cytoscape.org/apps/bingo
- http://bioinformatics.sdstate.edu/go/
- http://amigo.geneontology.org/amigo
- http://geneontology.org/
- https://biit.cs.ut.ee/gprofiler/
- http://revigo.irb.hr/

BiNGO

BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3448–3449
doi:10.1093/bioinformatics/bti551

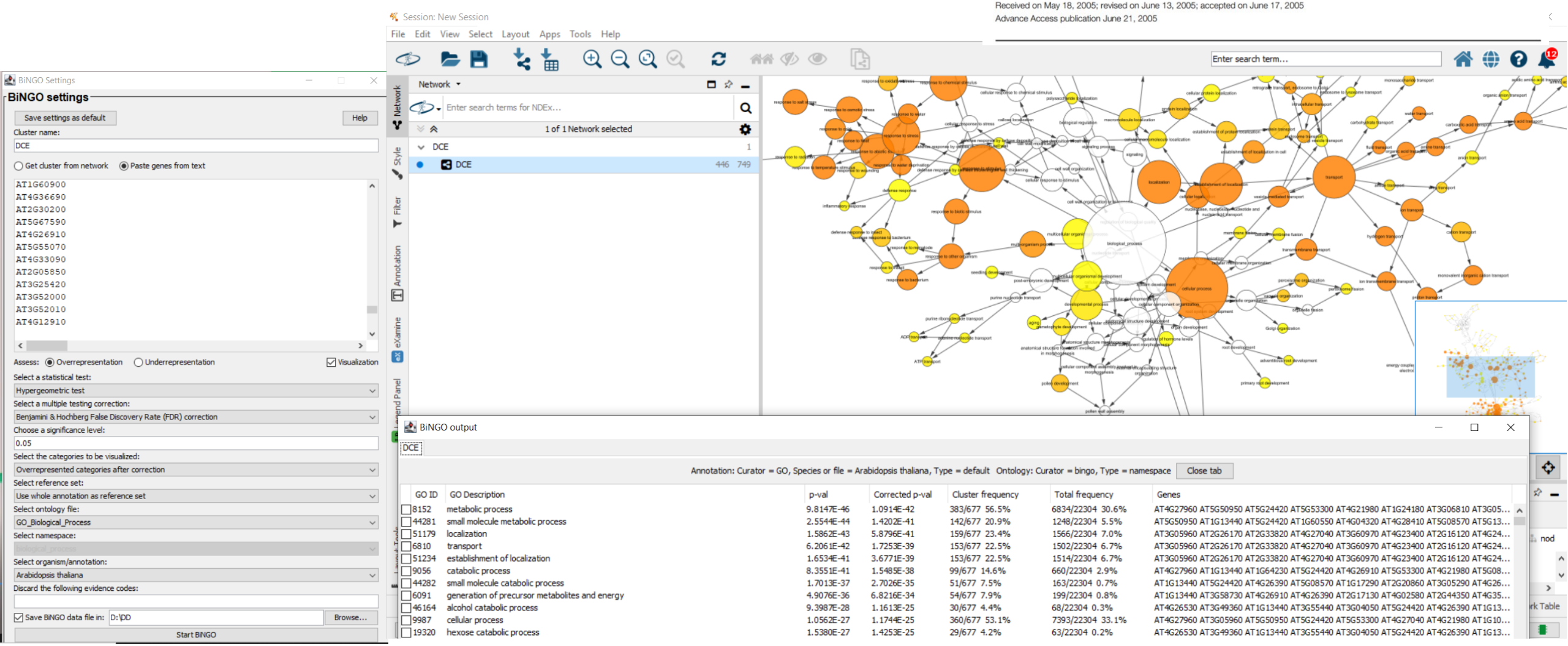
Systems biology

BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks

Steven Maere, Karel Heymans and Martin Kuiper*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052, Ghent, Belgium

Received on May 18, 2005; revised on June 13, 2005; accepted on June 17, 2005
Advance Access publication June 21, 2005



AmiGO

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 2 2009, pages 288–289
doi:10.1093/bioinformatics/btn615

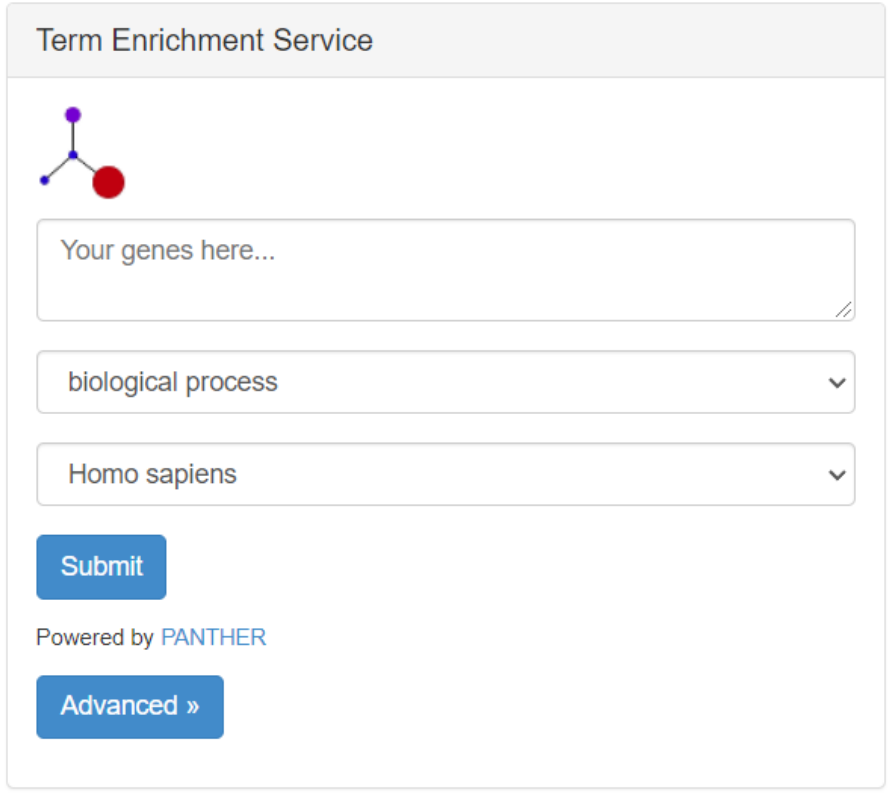
Databases and ontologies

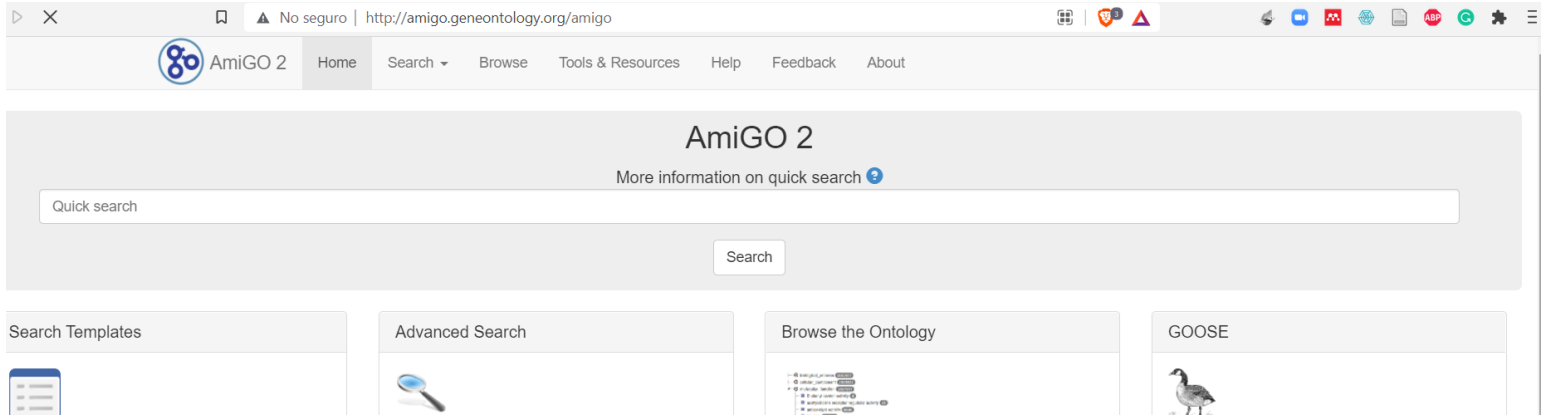
AmiGO: online access to ontology and annotation data

Seth Carbon[1,*,†], Amelia Ireland[2,†], Christopher J. Mungall[1], ShengQiang Shu[3], Brad Marshall[1], Suzanna Lewis[1], the AmiGO Hub[‡] and the Web Presence Working Group[‡]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]GO Editorial Office, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK and [3]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

Received on October 2, 2008; revised on November 20, 2008; accepted on November 21, 2008
Advance Access publication November 25, 2008
Associate Editor: Martin Bishop





http://amigo.geneontology.org/amigo

g:Profiler

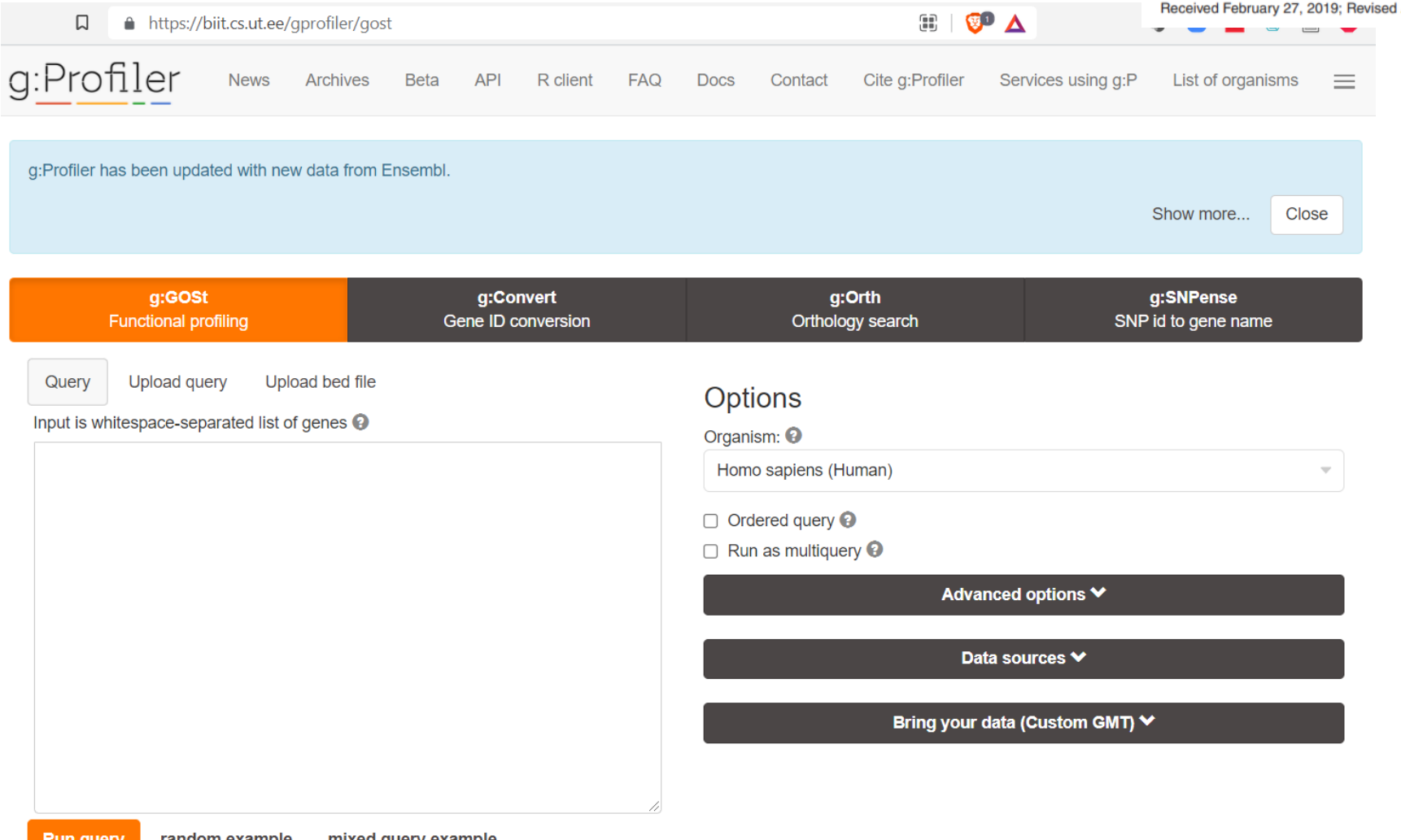
Published online 8 May 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W191–W198***
doi: 10.1093/nar/gkz369

g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)

Uku Raudvere[1,†], Liis Kolberg[1,†], Ivan Kuzmin[1], Tambet Arak[1], Priit Adler[1,2], Hedi Peterson[1,2,*] and Jaak Vilo[1,2,3,*]

[1]Institute of Computer Science, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia, [2]Quretec Ltd, Ülikooli 6a, 51003, Tartu, Estonia and [3]Software Technology and Applications Competence Centre, Ülikooli 2, 51003 Tartu, Estonia

Received February 27, 2019; Revised April 07, 2019; Editorial Decision April 16, 2019; Accepted April 29, 2019



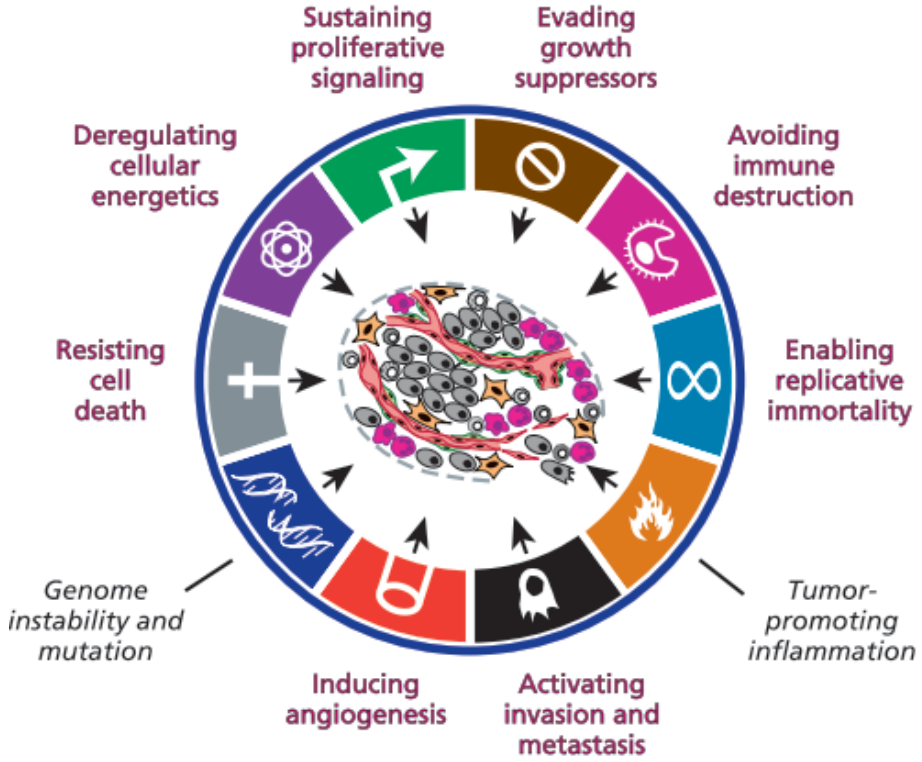
https://biit.cs.ut.ee/gprofiler/gost
https://biit.cs.ut.ee/gprofiler/page/r

Realización de un análisis de enriquecimiento funcional usando una lista de genes

El dataset que vamos a usar

- 5494 genes de *H. sapiens* en diez categorías:
- Inestabilidad genómica (167 genes)
- Dos ejercicios: BinGO y ShinnyGO

	A	B	C	D	E	F	G	H	I	J
1	AvoidingImmu	ActivatingInva	Deregulationg	EnablingReplic	EvadingGrowt	GenomeInstal	InducingAngio	ResistingCellD	SustainingPro	TumorPromot
2	CYBB	TSTA3	SUCLG1	PRIM2	VPS53	CDK5RAP3	SLC12A6	CDK5RAP3	CDK5RAP3	CYBB
3	PSMC1	DUSP26	CHMP4C	PCNA	BARD1	CNOT7	ACTA2	CSNK2A1	CHMP4C	RPS19
4	POLR3C	BZW1	VPS36	CCT4	RBL1	RNASEH2A	AQP1	CAT	CSNK2A1	CAMK1D
5	PSMD12	GSK3B	CYP2C9	NSMCE2	STK11	PCNA	HMGB1	GSK3B	PPP6C	CHIA
6	POLR3F	BZW2	ADIPOR1	TFIP11	SIN3A	POLR2D	PRKX	PDK4	CNOT7	HMGB1
7	CYFIP1	DENND6A	SORD	POLD1	CDC73	NSMCE2	AAMP	SLC25A4	PRIM2	TLR7
8	SLC11A1	RPL23A	GSK3B	GAR1	MYBBP1A	POLR2G	SAT1	ACSL5	PSMC1	IRAK4
9	SERINC3	OLA1	VPS35	UPF1	NDRG1	CDK1	NCL	PDK2	RPL23A	MAPK14
10	RPS19	DOCK7	YIPF1	MAPK14	MSH2	COPS7A	CTH	HMGN5	STAG3	LBP
11	PSME4	CYFIP1	PDK4	DKC1	FHIT	CUL4A	PDCL3	PRDX2	PCNA	NOX4
12	GNL1	RPS19	TAZ	RFC3	RAD51C	BARD1	NOX1	ATAD3A	PA2G4	CALR
13	PIK3R4	CDK1	MLYCD	FEN1	PLCD1	GTF2H1	ANXA3	NMT1	HMGN5	ANXA1
14	POLR3A	RPL34	CHAF1B	CALR	HINT1	POLR2I	EDF1	ZFAND6	CDKL1	PPIA
15	POLR3B	MMP7	ACSL5	RAD50	RBM5	POLD1	CHI3L1	PRDX3	CHMP4A	PHB
16	CHIA	ATP6V1D	PDK2	TERT	ASS1	CCNB1	ACVR1	SMNDC1	UFL1	GBF1
17	CTSH	GTPBP4	AKR1A1	MIF	CUL3	HMGB1	SPHK1	CDK1	DOCK7	MIF
18	PSMD1	CAMK1D	NDUFAF1	XRCC5	PPP2R1B	CUL4B	PDCD6	GLO1	PSMD12	RAC1
19	MAPK10	ENTPD1	CHMP4A	MAPKAPK5	WWOX	UPF1	SETSIP	CUL4A	SLC25A33	MVK
20	DDX3X	BYSL	IPPK	MAPK1	HPGD	MSH6	SERPINE1	CTNNBL1	NSMCE2	NCKAP1L
21	SERINC5	SERPINB3	SCD	RAD51D	NDRG2	EXO5	XDH	BARD1	PRMT5	CHID1
22	HMGB1	CTSH	CYP2C8	SMC6	CHMP1A	PPP4C	NOX5	CSE1L	PRDX3	PPIL2




```
("AID","AIM","DCE","ERI","EGS","GIM","IA","RCD","SPS","TPI")
```

Knijnenburg *et al. Chin J Cancer* (2015) 34:48
DOI 10.1186/s40880-015-0050-6

Chinese Journal of Cancer

ORIGINAL ARTICLE Open Access

A multilevel pan-cancer map links gene mutations to cancer hallmarks

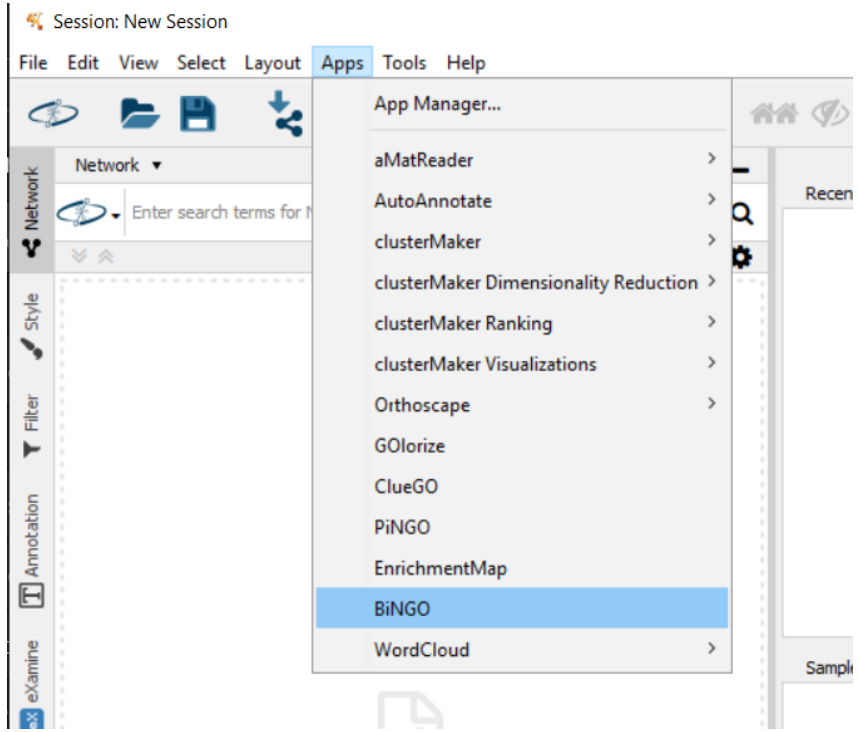
Theo A. Knijnenburg[1*], Tycho Bismeijer[2], Lodewyk F. A. Wessels[2] and Ilya Shmulevich[1]

Hanahan & Weinberg, 2011

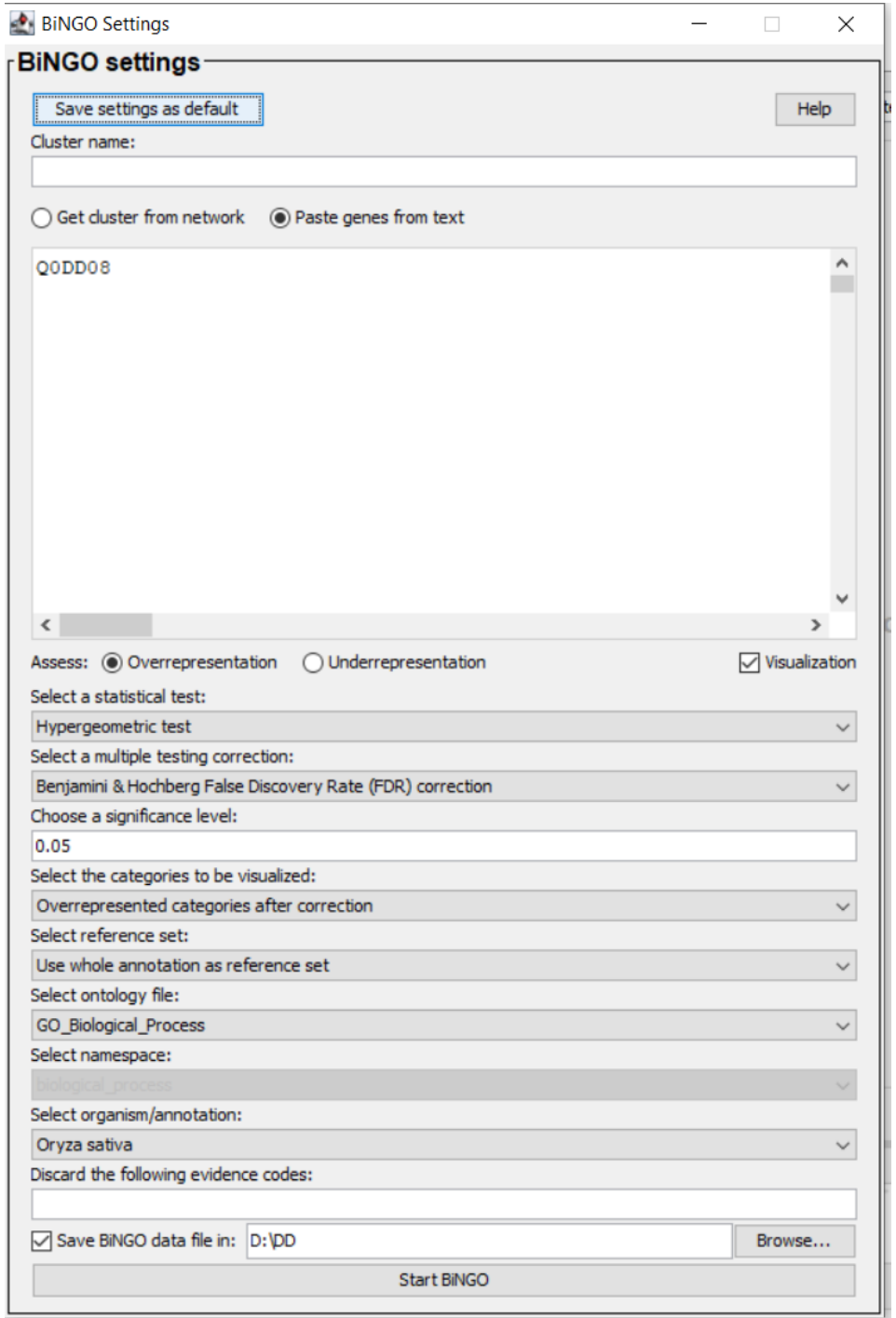
BinGO

- 1. Descargar el archivo disponible en https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv

- 2. Abra Cytoscape y el busque el plugin BinGO en el menú *Apps*



- **Nombre del análisis**
- **Espacio para incluir los ids de los genes**
- **Selección de evaluación a ser realizada**
- **Prueba estadística a ser usada**
- **Selección del método de corrección de hipótesis múltiples**
- **Elección del alfa**
- Selección de los GO a ser visualizados en un grafo
- Selección de la referencia a usar
- **Información a ser obtenida**
- Selección del organismo a ser analizado
- Espacio para guardar el archivo localmente
- **Botón para correr el análisis**



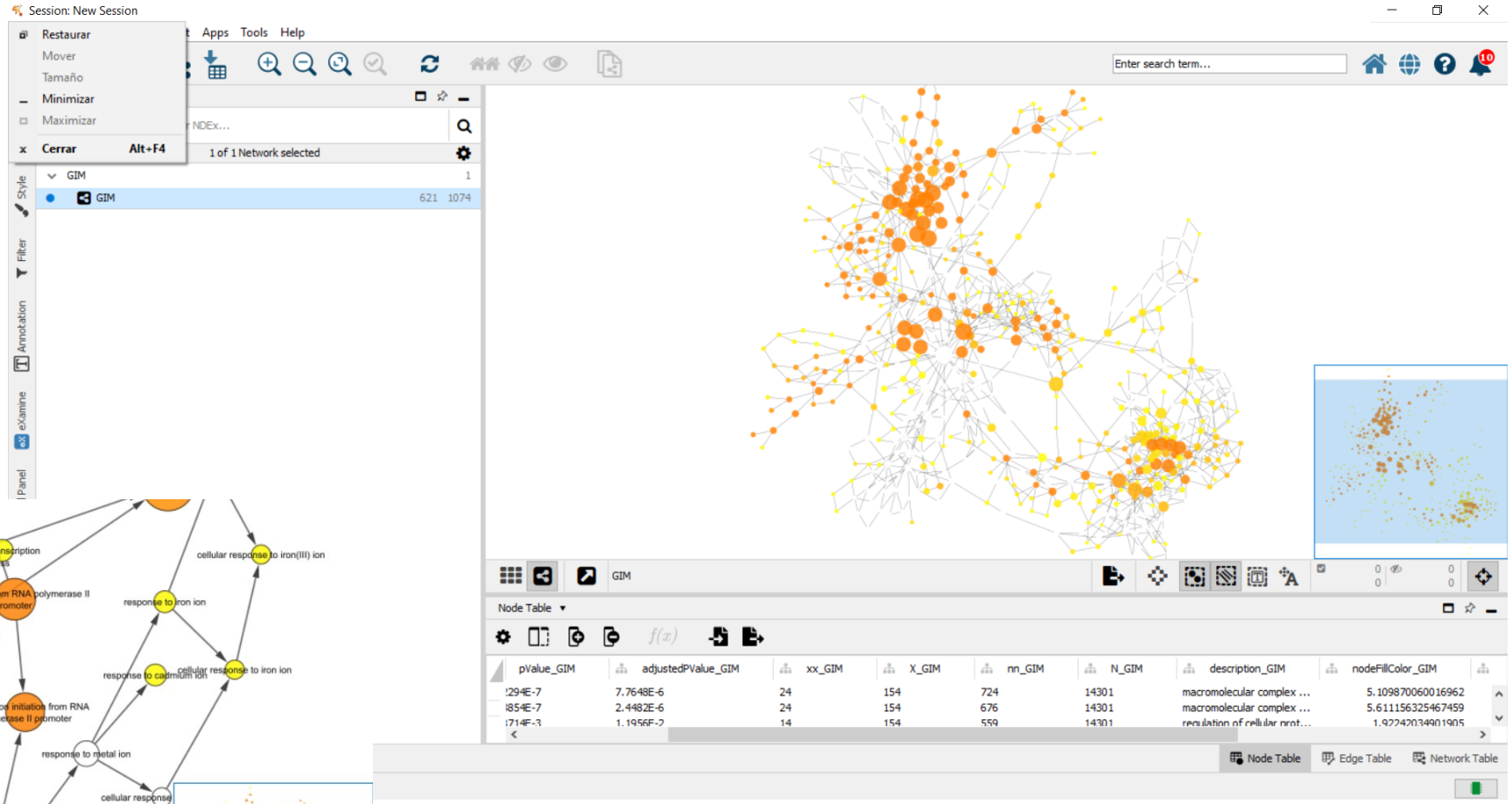
¿Qué es cada cosa en un análisis de enriquecimiento funcional?

Id, descripción, p valores, FDR

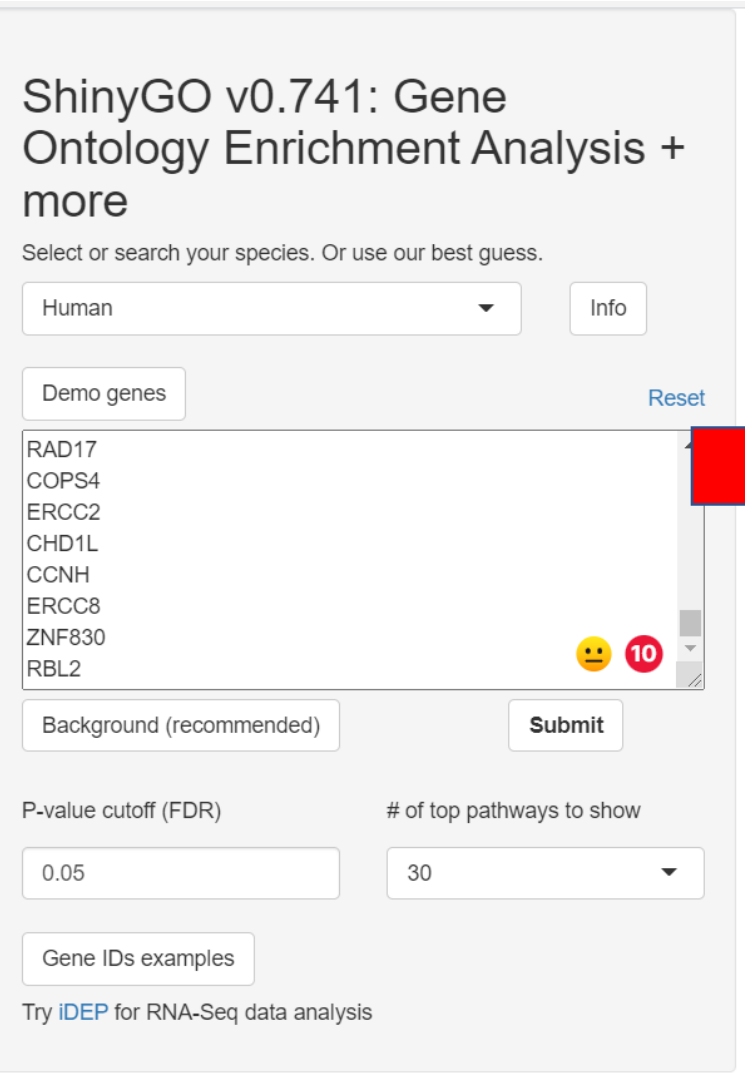
Annotation: Curator = GO, Species or file = Homo sapiens, Type = default Ontology: Curator = bingo, Type = namespace Close tab

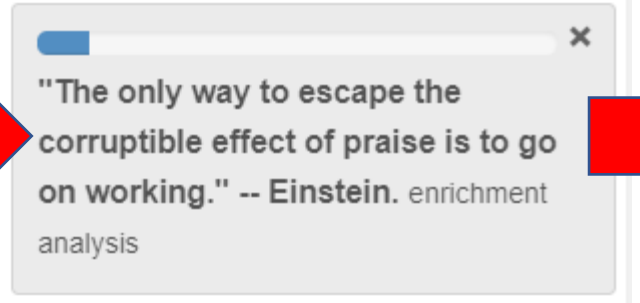
GO ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes
6974	response to DNA damage stimulus	8.0209E-123	0.0000E-100	101/154 65.5%	394/14301 2.7%	MPG CCNH SMC5 CCND1 RUVBL2 ALKBH1 CHEK1 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L
6281	DNA repair	1.2554E-122	0.0000E-100	94/154 61.0%	298/14301 2.0%	FEN1 MPG CCNH OGG1 SMC5 HMGB1 PRPF19 POLB RUVBL2 ALKBH1 CHEK1 POLE PRMT6 RFC5 UPF1
6259	DNA metabolic process	6.1519E-108	0.0000E-100	100/154 64.9%	515/14301 3.6%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L RBX1 M
33554	cellular response to stress	7.0582E-103	2.0000E-100	102/154 66.2%	617/14301 4.3%	MPG CCNH SMC5 CCND1 RUVBL2 ALKBH1 CHEK1 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L
51716	cellular response to stimulus	9.4405E-83	2.2091E-80	103/154 66.8%	985/14301 6.8%	MPG CCNH SMC5 CCND1 RUVBL2 ALKBH1 CHEK1 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L
90304	nucleic acid metabolic process	3.7304E-80	7.2742E-78	114/154 74.0%	1456/14301 10.1%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L RE
6139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3.1703E-70	5.2990E-68	114/154 74.0%	1782/14301 12.4%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L RE
34641	cellular nitrogen compound metabolic process	6.8807E-63	1.0063E-60	114/154 74.0%	2074/14301 14.5%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L RE
6807	nitrogen compound metabolic process	3.0593E-60	3.9771E-58	114/154 74.0%	2192/14301 15.3%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L RE
44260	cellular macromolecule metabolic process	2.3271E-58	2.7228E-56	132/154 85.7%	3504/14301 24.5%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR USP7 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CH
6950	response to stress	9.3223E-57	9.9155E-55	103/154 66.8%	1771/14301 12.3%	MPG CCNH SMC5 CCND1 RUVBL2 ALKBH1 CHEK1 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CHD1L
6289	nucleotide-excision repair	2.6415E-53	2.5755E-51	34/154 22.0%	58/14301 0.4%	PCNA CCNH OGG1 XPC POLD3 POLD1 POLD2 POLE RFC5 RFC3 RFC4 LIG1 RFC2 RPA1 LIG4 RPA2 RA
43170	macromolecule metabolic process	6.6767E-51	6.0090E-49	132/154 85.7%	4013/14301 28.0%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR USP7 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CH
44237	cellular metabolic process	3.2887E-39	2.7484E-37	132/154 85.7%	4986/14301 34.8%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR USP7 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CH
6310	DNA recombination	1.7287E-38	1.3483E-36	32/154 20.7%	105/14301 0.7%	IGHMBP2 SMC5 HMGB1 UNG RAD54B RECQL4 RUVBL2 CHEK1 RAD54L ZSWIM7 XRCC6 RBM14 LIG1 X
44238	primary metabolic process	3.5899E-37	2.6251E-35	133/154 86.3%	5284/14301 36.9%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR USP7 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CH
6260	DNA replication	1.1688E-32	8.0440E-31	34/154 22.0%	185/14301 1.2%	FEN1 RNASEH2A PCNA IGHMBP2 POLB POLD3 CDC45 POLD1 CHEK1 POLD2 NAE1 POLE RFC5 UPF1
8152	metabolic process	7.0191E-31	4.5624E-29	133/154 86.3%	5955/14301 41.6%	MPG CCNH SMC5 RUVBL2 ALKBH1 CHEK1 AQR USP7 LIG1 NSMCE2 RECQL LIG4 CSNK1D CSNK1E CH

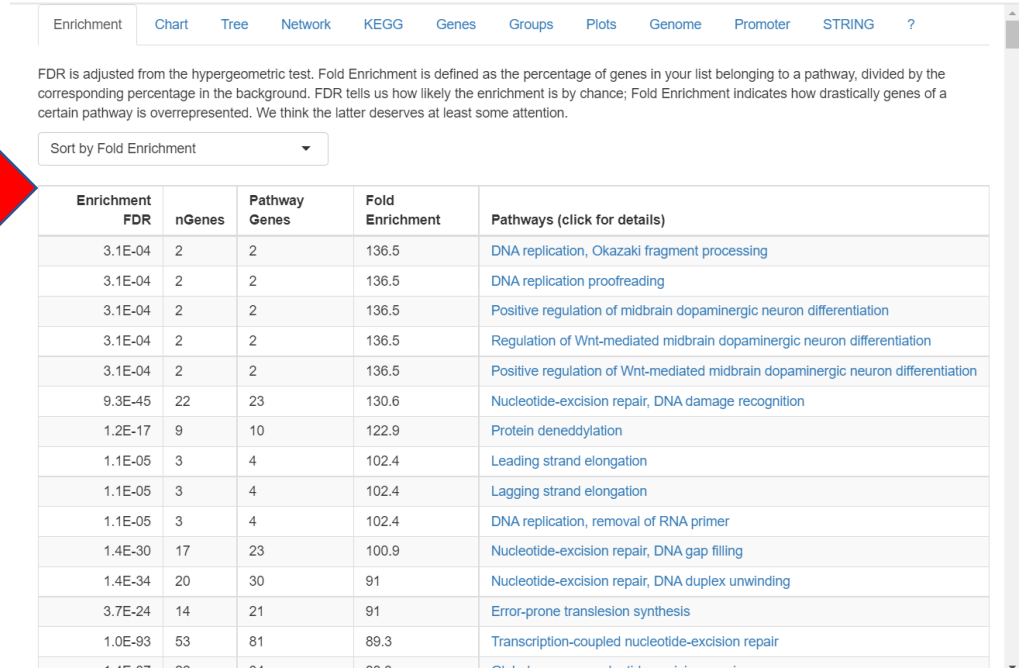
Resultado visual



ShinnyGO http://bioinformatics.sdstate.edu/go/



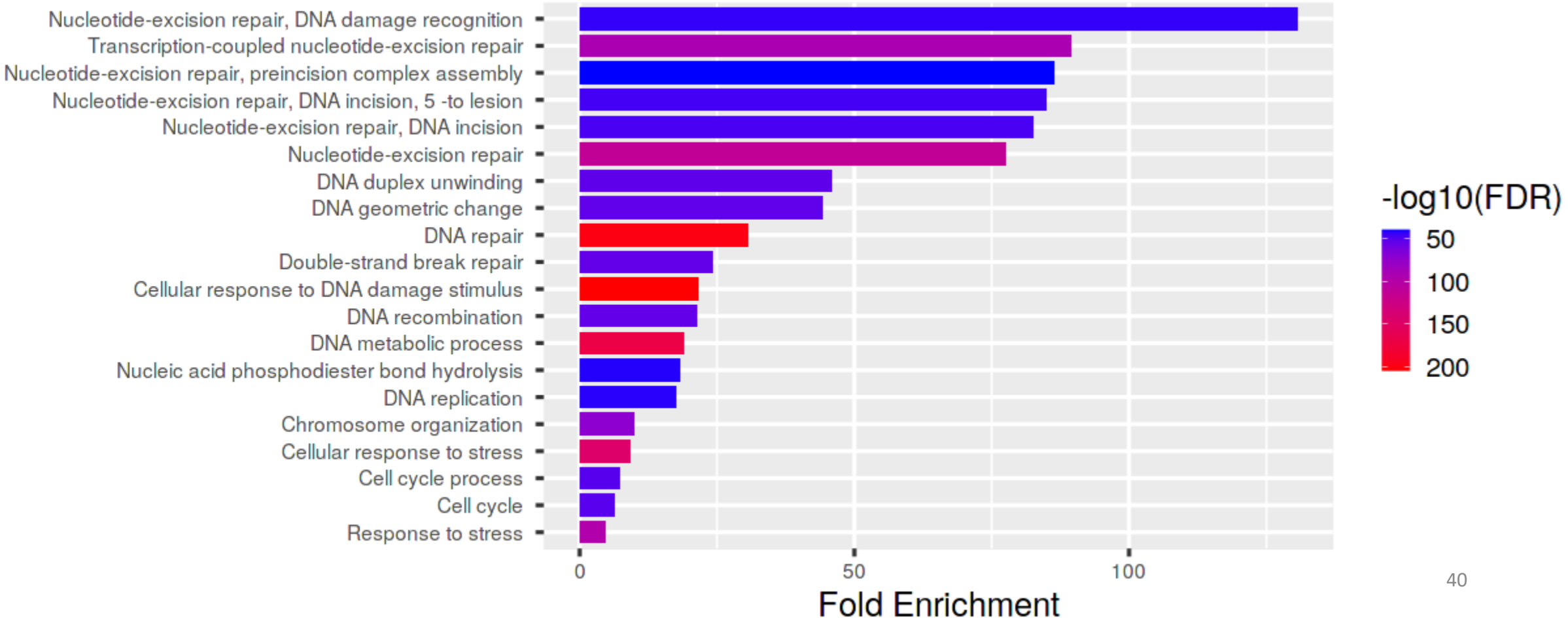




Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
3.1E-04	2	2	136.5	DNA replication, Okazaki fragment processing
3.1E-04	2	2	136.5	DNA replication proofreading
3.1E-04	2	2	136.5	Positive regulation of midbrain dopaminergic neuron differentiation
3.1E-04	2	2	136.5	Regulation of Wnt-mediated midbrain dopaminergic neuron differentiation
3.1E-04	2	2	136.5	Positive regulation of Wnt-mediated midbrain dopaminergic neuron differentiation
9.3E-45	22	23	130.6	Nucleotide-excision repair, DNA damage recognition
1.2E-17	9	10	122.9	Protein deneddylation
1.1E-05	3	4	102.4	Leading strand elongation
1.1E-05	3	4	102.4	Lagging strand elongation
1.1E-05	3	4	102.4	DNA replication, removal of RNA primer
1.4E-30	17	23	100.9	Nucleotide-excision repair, DNA gap filling
1.4E-34	20	30	91	Nucleotide-excision repair, DNA duplex unwinding
3.7E-24	14	21	91	Error-prone translesion synthesis
1.0E-93	53	81	89.3	Transcription-coupled nucleotide-excision repair

Fold enrichment: $\frac{\%\ genes\ \ de\ lista\ en\ una\ ruta}{\%\ genes\ en\ background}$

(como los genes son sobrerepresentados)



Análisis de enriquecimiento funcional en R

Recomendaciones

- Instale las librerías antes de realizar los ejercicios
- Corra el código línea a línea
- Intente no cambiar el código, hágalo solo si es necesario
- Algunos ejercicios pueden tomar algo de tiempo, por favor sea paciente
- Para futuros trabajos no guarde los resultados en su escritorio

Paquete gProfiler2

Tareas

- Cargar librerías
- Cargar un archivo CSV
- Correr un análisis de enriquecimiento similar al de BiNGO, solo con procesos biológicos

Código

```
#cargar librerias
  require(gprofiler2);library(biomaRt);library(topGO);
require(clusterProfiler);require(GOsummaries)

#cargar ejemplo desde CSV
  url_file = "https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv"
  x <- read.csv(url_file,header = T)
```

```
#cargar ejemplo desde CSV
  url_file = "https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv"
  x <- read.csv(url_file,header = T)
```

```
#Functional enrichment analysis
  x_s <-  gprofiler2::gost(query = x[,1],
                           organism = "hsapiens", ordered_query = FALSE,
                           multi_query = FALSE, significant = TRUE, exclude_iea = FALSE,
                           measure_underrepresentation = FALSE, evcodes = FALSE,
                           user_threshold = 0.05, correction_method = "false_discovery_rate",
                           domain_scope = "annotated", custom_bg = NULL,
                           numeric_ns = "", as_short_link = FALSE,
                           sources="GO:BP")
```

https://f1000research.com/articles/9-709

- Graficar en un *Manhattan plot* no interactivo

```
#GO PLOT
  p <- gostplot(x_s, capped = F, interactive = FALSE)
  p
```

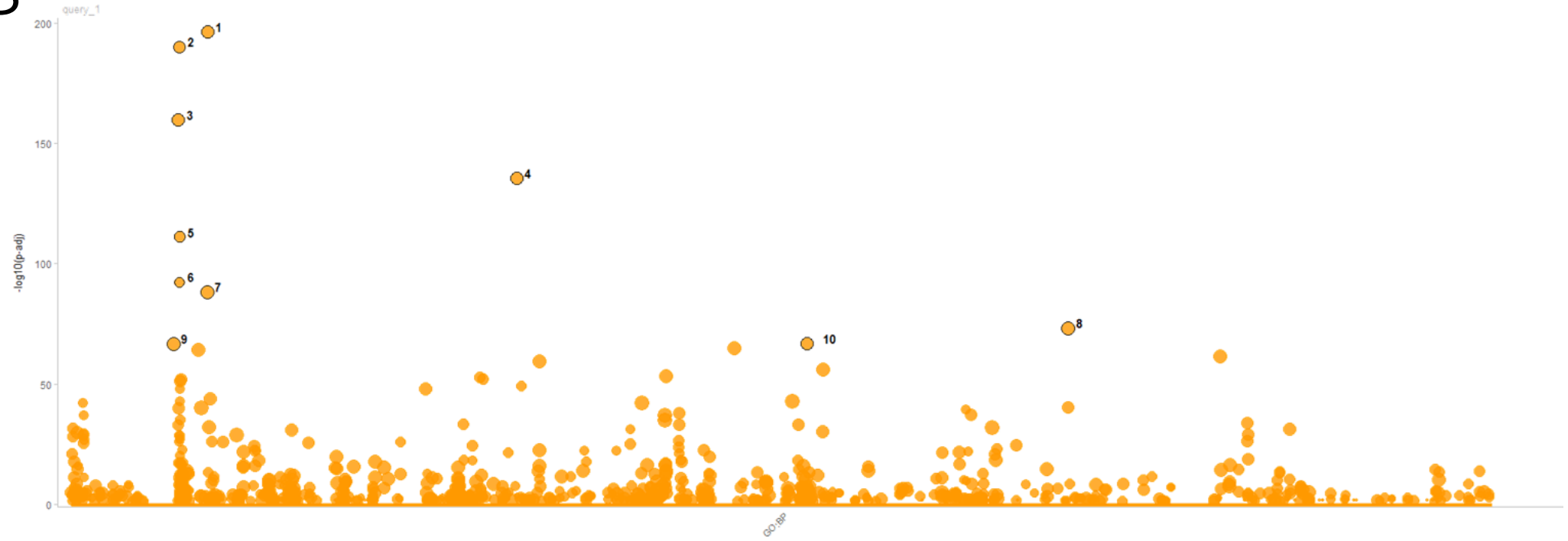
- Graficar en un *Manhattan plot* no interactivo el top diez de GO obtenidos

```
#GO PLOT with the top five of GO
  publish_gostplot(p, highlight_terms = x_s$result$term_id[1:10])
```

View(x_s$result)

	query	significant	p_value	term_size	query_size	intersection_size	precision	recall	term_id	source	term_name	effective_domain_size	source_order	parents
458	query_1	TRUE	2.200238e-04	21	166	4	0.024096386	0.19047619	GO:0000002	GO:BP	mitochondrial genome maintenance	18123	2	GO:0007005
363	query_1	TRUE	1.782347e-05	1515	166	33	0.198795181	0.02178218	GO:0000003	GO:BP	reproduction	18123	3	GO:0008150
384	query_1	TRUE	3.311130e-05	13	166	4	0.024096386	0.30769231	GO:0000012	GO:BP	single strand break repair	18123	5	GO:0006281
90	query_1	TRUE	2.274247e-21	136	166	23	0.138554217	0.16911765	GO:0000018	GO:BP	regulation of DNA recombination	18123	7	c("GO:0006310", "GO:0051052")
446	query_1	TRUE	1.611159e-04	7	166	3	0.018072289	0.42857143	GO:0000019	GO:BP	regulation of mitotic recombination	18123	8	c("GO:0000018", "GO:0006312")
859	query_1	TRUE	4.193197e-02	19	166	2	0.012048193	0.10526316	GO:0000028	GO:BP	ribosomal small subunit assembly	18123	15	c("GO:0022618", "GO:0042255", "GO
467	query_1	TRUE	3.178246e-04	185	166	9	0.054216867	0.04864865	GO:0000070	GO:BP	mitotic sister chromatid segregation	18123	32	c("GO:0000819", "GO:0140014", "GO
46	query_1	TRUE	3.694259e-32	196	166	34	0.204819277	0.17346939	GO:0000075	GO:BP	cell cycle checkpoint signaling	18123	35	c("GO:0035556", "GO:1901988")
259	query_1	TRUE	6.615812e-08	17	166	6	0.036144578	0.35294118	GO:0000076	GO:BP	DNA replication checkpoint signaling	18123	36	GO:0031570
58	query_1	TRUE	1.536764e-28	136	166	28	0.168674699	0.20588235	GO:0000077	GO:BP	DNA damage checkpoint signaling	18123	37	c("GO:0031570", "GO:0042770")
373	query_1	TRUE	2.456541e-05	98	166	8	0.048192771	0.08163265	GO:0000079	GO:BP	regulation of cyclin-dependent protein serine/threonin...	18123	39	c("GO:0071900", "GO:1904029")
104	query_1	TRUE	2.557733e-18	260	166	26	0.156626506	0.10000000	GO:0000082	GO:BP	G1/S transition of mitotic cell cycle	18123	41	c("GO:0044772", "GO:0044843")
522	query_1	TRUE	1.363356e-03	34	166	4	0.024096386	0.11764706	GO:0000083	GO:BP	regulation of transcription involved in G1/S transition o...	18123	42	c("GO:0000082", "GO:0006355", "GO
172	query_1	TRUE	6.015003e-12	262	166	20	0.120481928	0.07633588	GO:0000086	GO:BP	G2/M transition of mitotic cell cycle	18123	45	c("GO:0044772", "GO:0044839")
790	query_1	TRUE	3.149738e-02	1	166	1	0.006024096	1.00000000	GO:0000160	GO:BP	phosphorelay signal transduction system	18123	69	GO:0035556
670	query_1	TRUE	1.282726e-02	155	166	6	0.036144578	0.03870968	GO:0000187	GO:BP	activation of MAPK activity	18123	78	c("GO:0032147", "GO:0043406")
227	query_1	TRUE	4.741069e-09	343	166	19	0.114457831	0.05539359	GO:0000209	GO:BP	protein polyubiquitination	18123	100	GO:0016567
539	query_1	TRUE	2.072091e-03	663	166	16	0.096385542	0.02413273	GO:0000226	GO:BP	microtubule cytoskeleton organization	18123	103	c("GO:0007010", "GO:0007017")

El resultado: ¿Qué se puede inferir del gráfico?



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0006974	cellular response to DNA damage stimulus	894	6.0e-197
2	GO:BP	GO:0006281	DNA repair	586	5.7e-191
3	GO:BP	GO:0006259	DNA metabolic process	976	1.8e-160
4	GO:BP	GO:0033554	cellular response to stress	2142	3.2e-136
5	GO:BP	GO:0006289	nucleotide-excision repair	107	1.0e-111
6	GO:BP	GO:0006283	transcription-coupled nucleotide-excision repair	72	4.4e-93
7	GO:BP	GO:0006950	response to stress	4321	5.4e-89
8	GO:BP	GO:0090304	nucleic acid metabolic process	5234	6.5e-74
9	GO:BP	GO:0006139	nucleobase-containing compound metabolic process	5752	1.8e-67
10	GO:BP	GO:0051276	chromosome organization	1270	2.4e-67

g:Profiler (biit.cs.ut.ee/gprofiler)

Paquete topGO

Tareas

- Cargar librerías
- Cargar un archivo CSV
- Descargar los términos GO asociados a la lista de genes usando ENSEMBL

Código

```
#cargar librerias
  require(gprofiler2);library(biomaRt);library(topGO);
require(clusterProfiler);require(GOsummaries)

#cargar ejemplo desde CSV
  url_file = "https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv"
  x <- read.csv(url_file,header = T)
```

```
#cargar ejemplo desde CSV
  url_file = "https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv"
  x <- read.csv(url_file,header = T)
```

```
#Obtener los GO desde ENSEMBL
db= biomaRt::useMart('ENSEMBL_MART_ENSEMBL',dataset='hsapiens_gene_ensembl', host="www.ensembl.org")
go_ids= biomaRt::getBM(attributes=c('go_id', 'external_gene_name', 'namespace_1003'),
                       filters='external_gene_name',
                       values=x[,1],
                       mart=db)
```

Tarea

- Obtener el formato necesario de la lista de genes para usarse en topGO y remover genes sin anotaciones
- Obtener el objeto a ser usado en el análisis

Código

```
# remove any candidate genes without GO annotation
keep = x[,1] %in% go_ids[,2]
keep =which(keep==TRUE)
candidate_list=x[,1][keep]
geneList=factor(as.integer(x[,1] %in% candidate_list),levels = c(0,1))
names(geneList)= x[,1]
```

```
# Create the class topGOdata
GOdata=new('topGOdata', ontology='BP', allGenes = geneList, annot = annFUN.gene2GO, gene2GO = gene_2_GO)
```

Tarea

- Correr pruebas de hipótesis y otros algoritmos
- Guardar los resultados de los análisis en una única tabla y solo conservar el top diez
- Obtener los valores de las pruebas de hipótesis
- Graficar el grafo con los tres resultados más significantes usando el método elim

Código

```
#Run stastical tests
resultFisher <- runTest(GOdata, algorithm = "classic", statistic = "fisher")
resultKS <- runTest(GOdata, algorithm = "classic", statistic = "ks")
resultKS.elim <- runTest(GOdata, algorithm = "elim", statistic = "ks")

#summarize in a table
```

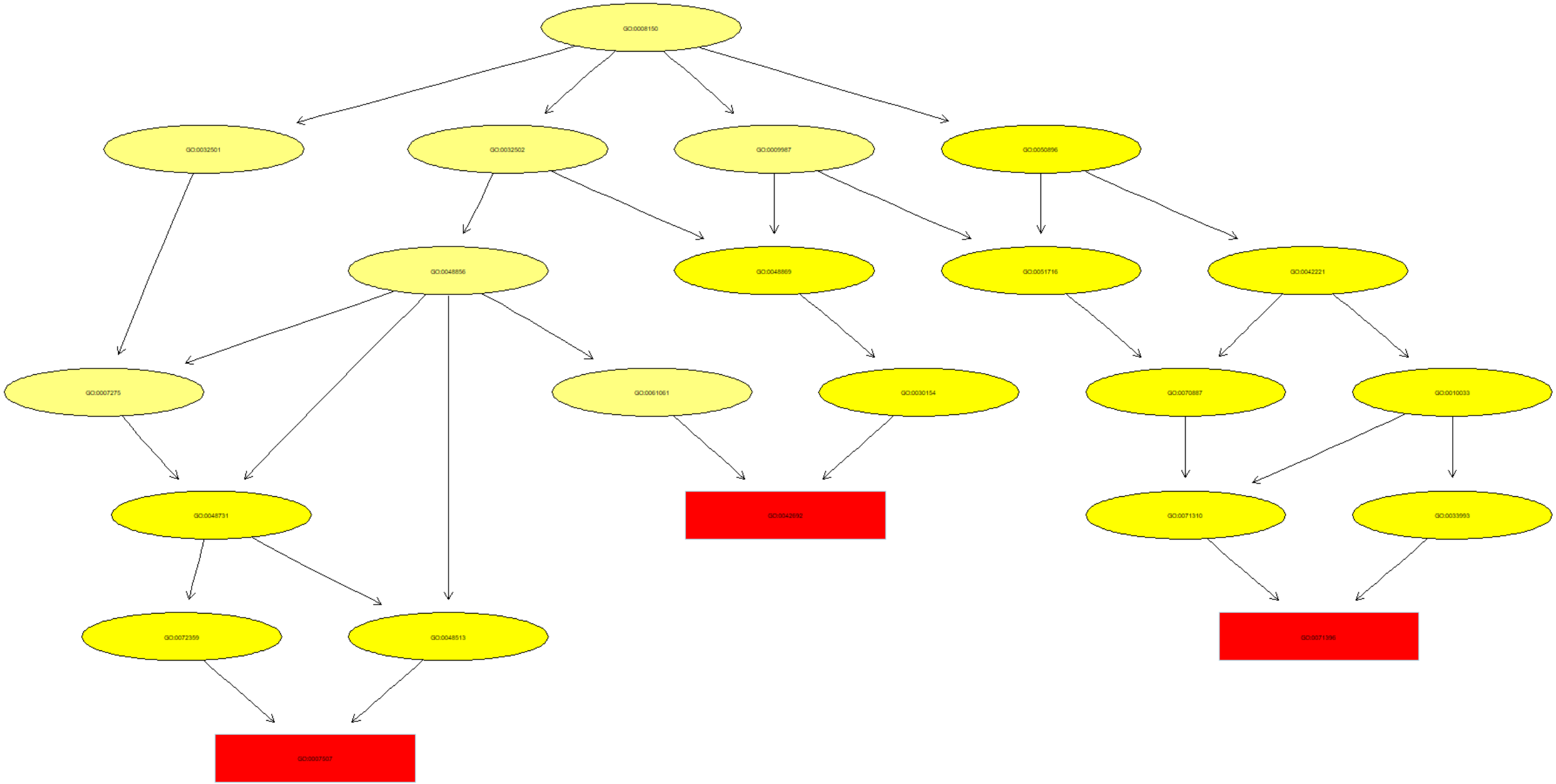
```
#summarize in a table
allRes <- topGO::GenTable(GOdata, classicFisher = resultFisher,
                          classicKS = resultKS, elimKS = resultKS.elim,
                          orderBy = "elimKS", ranksOf = "classicFisher", topNodes = 10)
```

```
#diversas formas de obtener p valores
pValue.classic <- score(resultKS)
pValue.elim <- score(resultKS.elim)[names(pValue.classic)]
gstat <- topGO::termStat(GOdata, names(pValue.classic))
```

```
par(cex = 1)
showSigOfNodes(GOdata, score(resultKS.elim), firstSigNodes = 3,useInfo = "def")
```

View(allRes)

	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
1	GO:0042692	muscle cell differentiation	5	5	5	1	1	0.0019	0.0019
2	GO:0071396	cellular response to lipid	8	8	8	2	1	0.0021	0.0021
3	GO:0007507	heart development	4	4	4	3	1	0.0027	0.0027
4	GO:0051240	positive regulation of multicellular org...	24	24	24	4	1	0.0040	0.0040
5	GO:0022411	cellular component disassembly	7	7	7	5	1	0.0046	0.0046
6	GO:2000026	regulation of multicellular organismal d...	23	23	23	6	1	0.0063	0.0063
7	GO:0031124	mRNA 3'-end processing	4	4	4	7	1	0.0070	0.0070
8	GO:0034504	protein localization to nucleus	4	4	4	8	1	0.0075	0.0075
9	GO:0006611	protein export from nucleus	3	3	3	9	1	0.0077	0.0077
10	GO:2001251	negative regulation of chromosome organi...	12	12	12	10	1	0.0081	0.0081


GO:0008150
GO:0032501
GO:0032502
GO:0009987
GO:0050896
GO:0048856
GO:0048869
GO:0051716
GO:0042221
GO:0007275
GO:0061061
GO:0030154
GO:0070887
GO:0010033
GO:0048731
GO:0042692
GO:0071310
GO:0033993
GO:0072359
GO:0048513
GO:0071396
GO:0007507

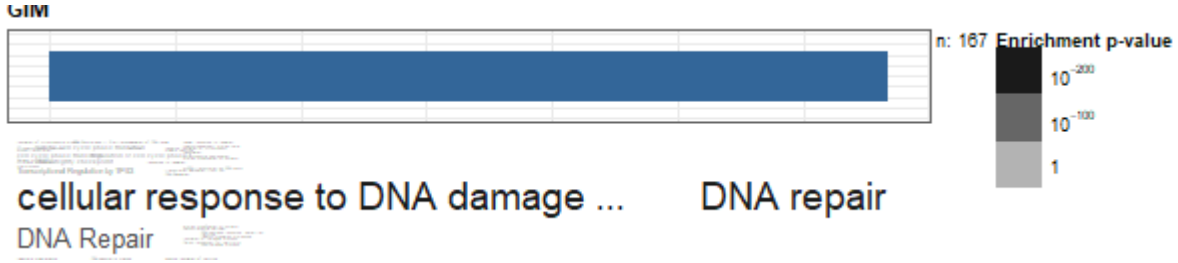
Paquete GOSummaries

Tareas

- Listar los genes candidatos
- Usar g:Profiler para realizar el análisis de enriquecimiento
- Graficar

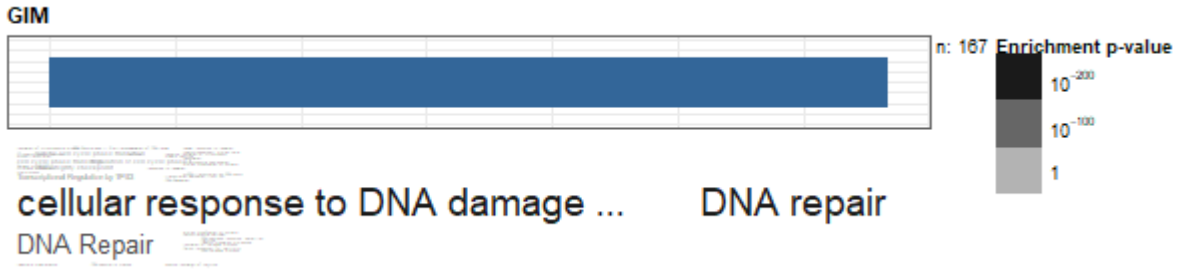
Código

```
#alternativa usando gosummaries
gl = list(GIM = x[,1]) # Two lists per component
gs = GOsummaries::gosummaries(gl)
plot(gs, fontsize = 8)
```



View(gs[[1]]WCDwcd1)

	Term	Score
1	cellular response to DNA damage stimulus	1.62e-210
2	DNA repair	6.98e-208
3	DNA Repair	3.48e-138
4	Transcriptional Regulation by TP53	2.86e-43
5	DNA integrity checkpoint	4.70e-35
6	cell cycle phase transition	1.60e-31
7	mitotic cell cycle phase transition	1.00e-30
8	regulation of cell cycle phase transition	7.54e-29
9	response to radiation	1.03e-27
10	Cell Cycle	9.23e-26
11	telomere maintenance	2.13e-25
12	DNA Replication	5.44e-25
13	DNA geometric change	2.28e-21
14	regulation of DNA repair	6.37e-19
15	positive regulation of cell cycle	2.28e-17
16	transcription elongation from RNA polymerase II promoter	2.89e-17
17	protein modification by small protein removal	4.80e-17
18	regulation of chromosome organization	3.57e-15
19	RNA Polymerase II Pre-transcription Events	2.35e-14
20	meiotic cell cycle	1.33e-13



Análisis de enriquecimiento funcional por categorías y entre dos especies usando R

Paquete gProfiler2

Tareas

- Cargar librerías
- Cargar un archivo CSV
- Convirtiendo cada columna en una lista de genes

Código

```r
#cargar librerias
  require(gprofiler2);library(biomaRt);library(topGO);
require(clusterProfiler);require(GOsummaries)

#cargar ejemplo desde csv
  url_file = "https://raw.githubusercontent.com/ccsosa/R_Examples/master/GIM.csv"
  x <- read.csv(url_file,header = T)
```

```r
url_file = "https://raw.githubusercontent.com/ccsosa/TALLER_OMICAS/master/Hallmarks_of_Cancer_AT.csv"
x_group <- read.csv(url_file)
#leyendo archivo
x_group[,1] <- NULL
#definiendo categorias
CH <- c("AID","AIM","DCE","ERI","EGS","GIM","IA","RCD","SPS","TPI")
```

```r
#preparacion de los datos
x_Hsap <- lapply(seq_len(length(CH)), function(i){
  x_unique <- unique(na.omit(x_group[,i]))
  x_unique <- x_unique[which(x_unique!="")]
  x_unique <- as.list(x_unique)
  return(x_unique)
})

names(x_Hsap) <- CH
```

https://f1000research.com/articles/9-709

Tarea

- Correr el análisis de enriquecimiento funcional para cada una de las listas de genes
- Guardar el resultado en un objeto
- Obtendremos el número de términos enriquecidos por lista de genes

Código

```
#Correr enriquecimiento funcional para todas las listas de genes
x_s_group <-  gprofiler2::gost(query = x_Hsap,
                               organism = "hsapiens", ordered_query = FALSE,
                               multi_query = FALSE, significant = TRUE, exclude_iea = FALSE,
                               measure_underrepresentation = FALSE, evcodes = T,
                               user_threshold = 0.05, correction_method = "fdr",
                               domain_scope = "annotated", custom_bg = NULL,
                               numeric_ns = "", sources = "GO:BP", as_short_link = FALSE)
```

```
res_group <- x_s_group$result
```

```
#Obtener numero de GO enriquecidos por lista de genes
tapply(res_group$query,res_group$query,length)
```

Tarea

- Esta vez haremos el gráfico interactivo
- Obtener el top tres de GO enriquecidos por categoría

- Obtener la tabla de frecuencia para la grafica

Código

```
p <- gprofiler2::gostplot(x_s_group, capped = F, interactive = TRUE)
```

```
#Obtener top tres GO por categoria
x_res <- list()
for(i in 1:length(CH)){
  x_res[[i]] <- res_group[which(res_group$query==CH[[i]]),]
  x_res[[i]] <- x_res[[i]][c(1:3),]
}
x_res <- do.call(rbind, x_res)
```

```
#obtener la tabla de frecuencias para graficar
x_res$query <- factor(x_res$query,levels = CH)
other_table <- table( x_res$query,x_res$term_name)
```

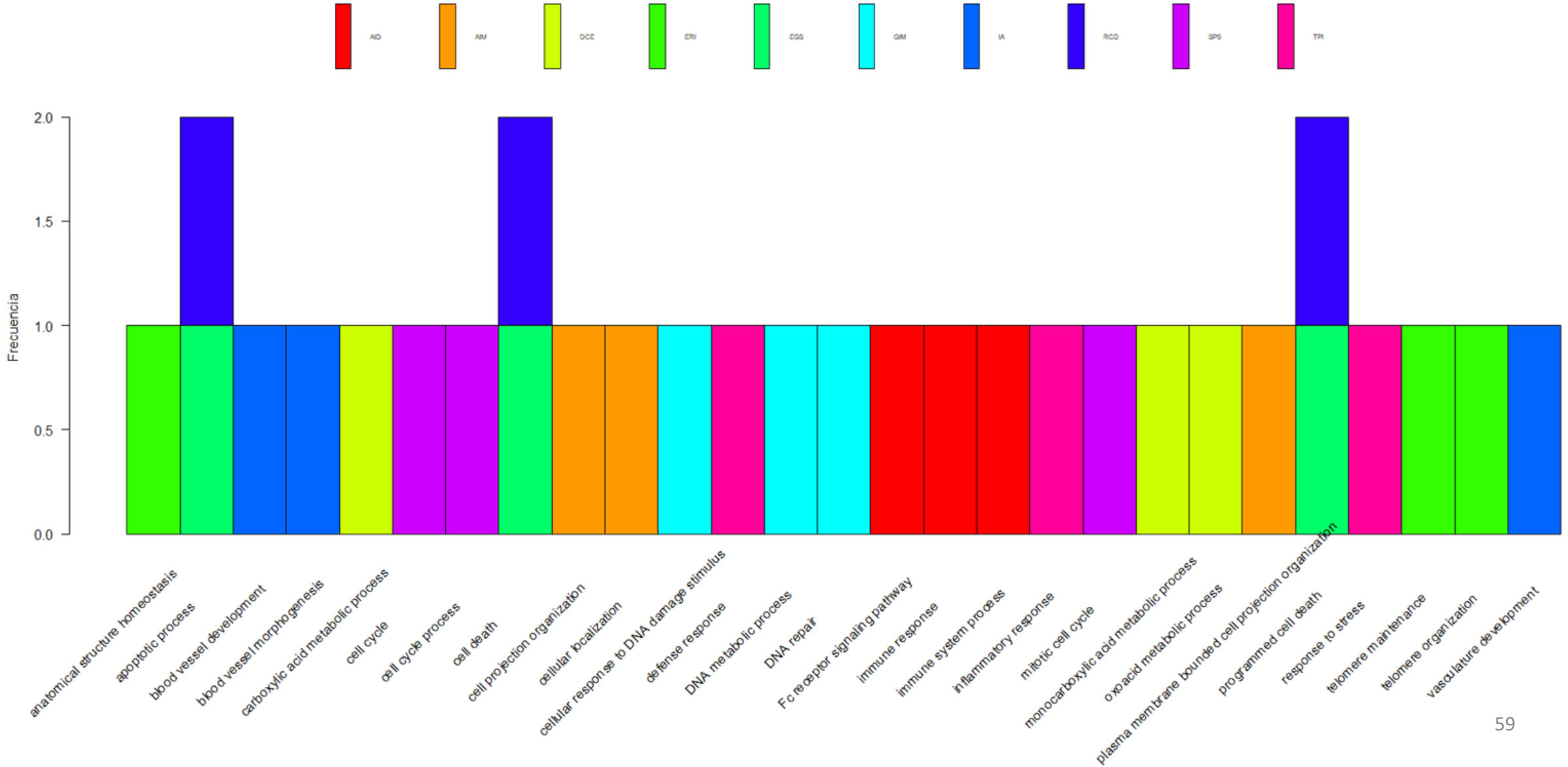
Vamos a graficar el top tres de GO enriquecidos

Código

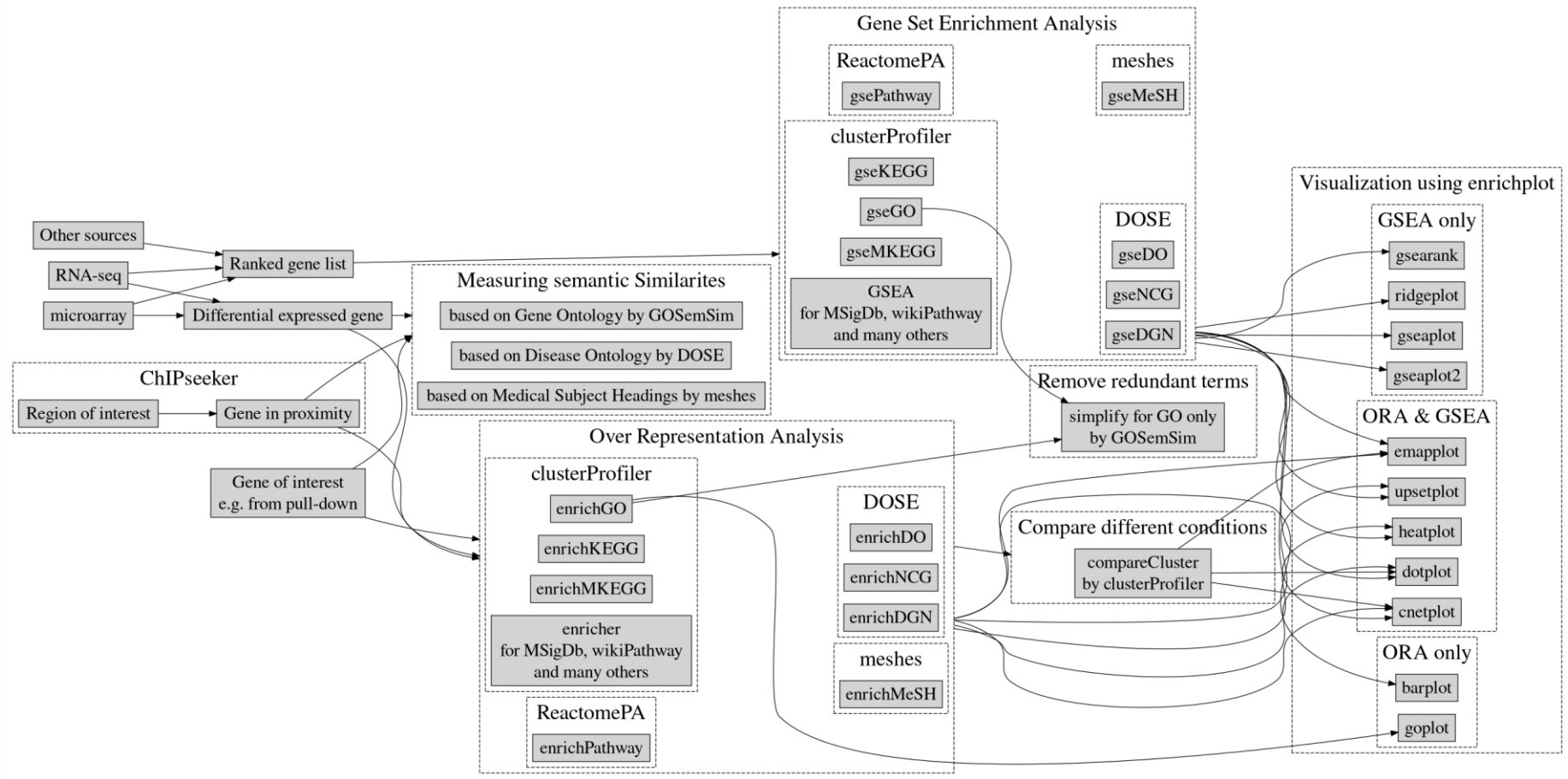
```
par(mar = c(13, 4, 11, 1) + 0.2) #add room for the rotated labels
#par(mar=c(1,1,1,1))

bar <- barplot(other_table,
        main = "",
        xlab = "", ylab = "Frecuencia",
        col = rainbow(10),
        xaxt="n",
        axes=T,
       # legend.text = rownames(other_table),
        beside = F,
        las=2,horiz = F,
        space=0,cex.names = 0.8)

labs <- paste(colnames(other_table))
text(cex=1, x=bar-1, y=-.52,labs, xpd=TRUE, srt=45)
#axis(2, at = 0:5, labels = 0:5)
legend("top", rownames(other_table), fill = rainbow(10), bty = "n",horiz = T,inset = c(0,-0.5),xpd = T,
       cex = 0.5)
```

Top tres de términos GO por *cancer hallmark*

```
> tapply(res_group$query,res_group$query,length)
 AID  AIM  DCE  EGS  ERI  GIM   IA  RCD  SPS  TPI
1102 1605 1067  351  717  874  822 1784 1492  803
```



clusterProfiler



Tarea

- Cambiar ids de HUGO a ENTREZ
- Correr enriquecimiento
- Obtener resultados de enriquecimiento
- Graficar en un *dotplot*

Código

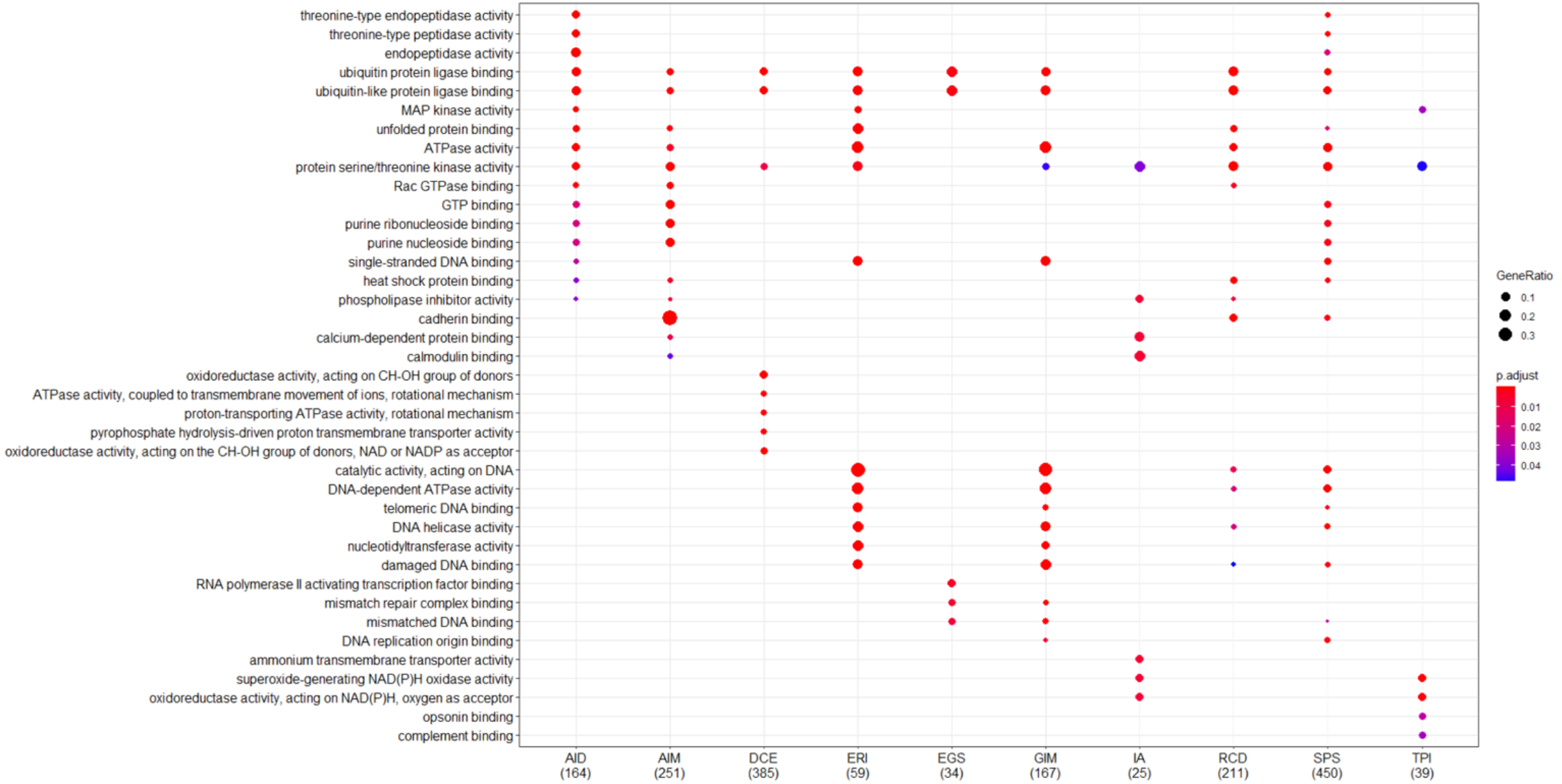
```
#Organizar la lista de genes para usarse en clusterProfiler
x_Hsap_2 <- list()
for(i in 1:length(x_Hsap)){
  x_Hsap_2[[i]] <- clusterProfiler::bitr(as.character(unlist(x_Hsap[[i]])),
                                         fromType = "SYMBOL",
                                         toType = c("ENTREZID"),
                                         OrgDb = "org.Hs.eg.db")[,2]
}
names(x_Hsap_2) <-CH
```

```
x_compare <- clusterProfiler::compareCluster(geneClusters=x_Hsap_2,enrichGO, OrgDb = org.Hs.eg.db)
```

```
clust_results <- x_compare@compareClusterResult
tapply(clust_results$Cluster,clust_results$Cluster,length)
```

```
#Graficar en un dotplot
enrichplot::dotplot(x_compare)
```

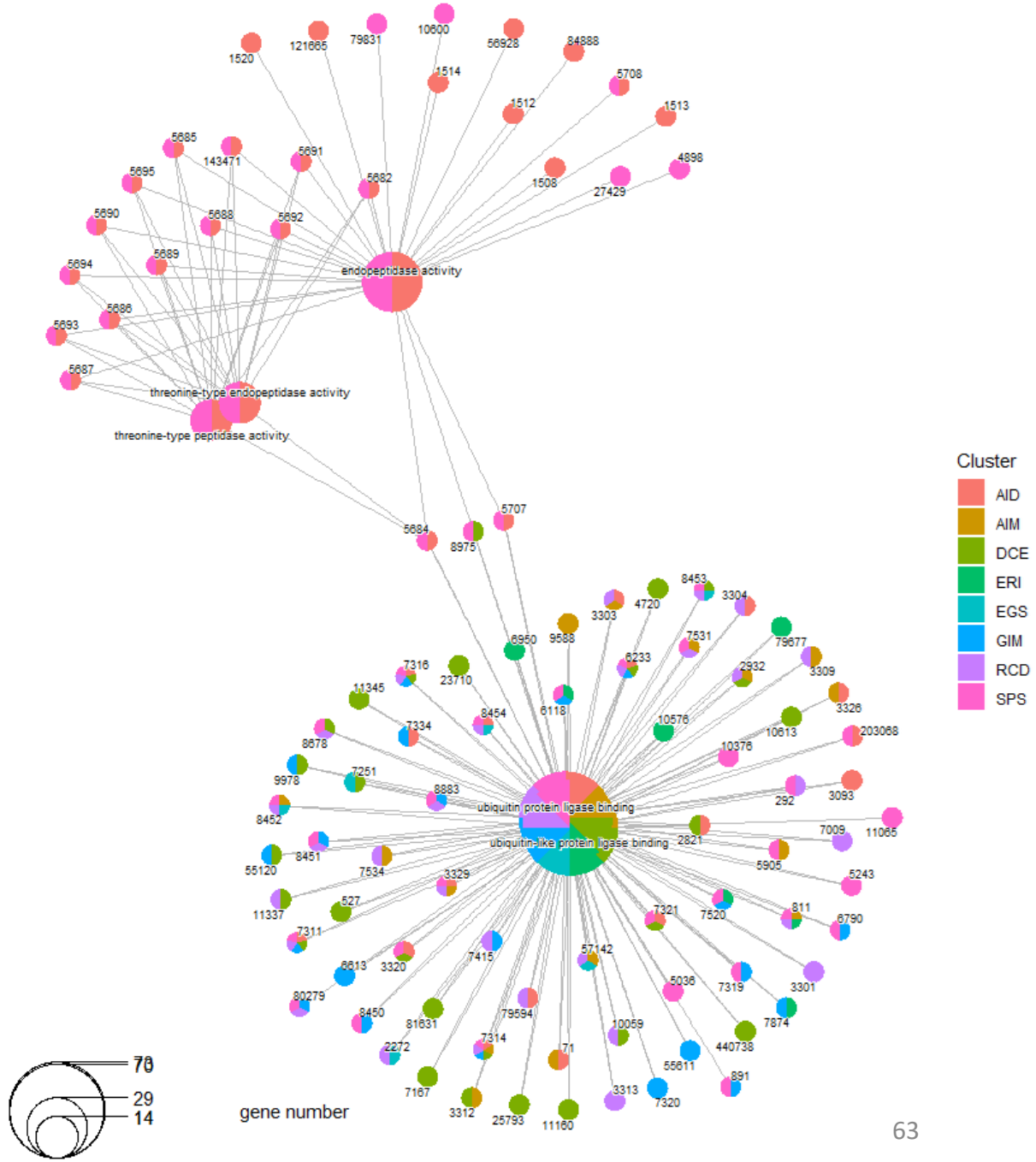
```
> tapply(clust_results$Cluster,clust_results$Cluster,length)
AID AIM DCE ERI EGS GIM  IA RCD SPS TPI
 66  81 141  64   9 108  15  75 112  11
```



CNETPLOT

- Graficar una red de interacciones

```
#Graficar una red
enrichplot::cnetplot(x_compare)
```



Visualización de gprofiler en clusterprofiler

Tarea

- Obtener los resultados de gProfiler
- Calcular proporción de GO enriquecidos y alistar para ver los resultados de gprofiler en clusterprofiler
- Convertir de formato gprofiler a clusterprofiler

Código

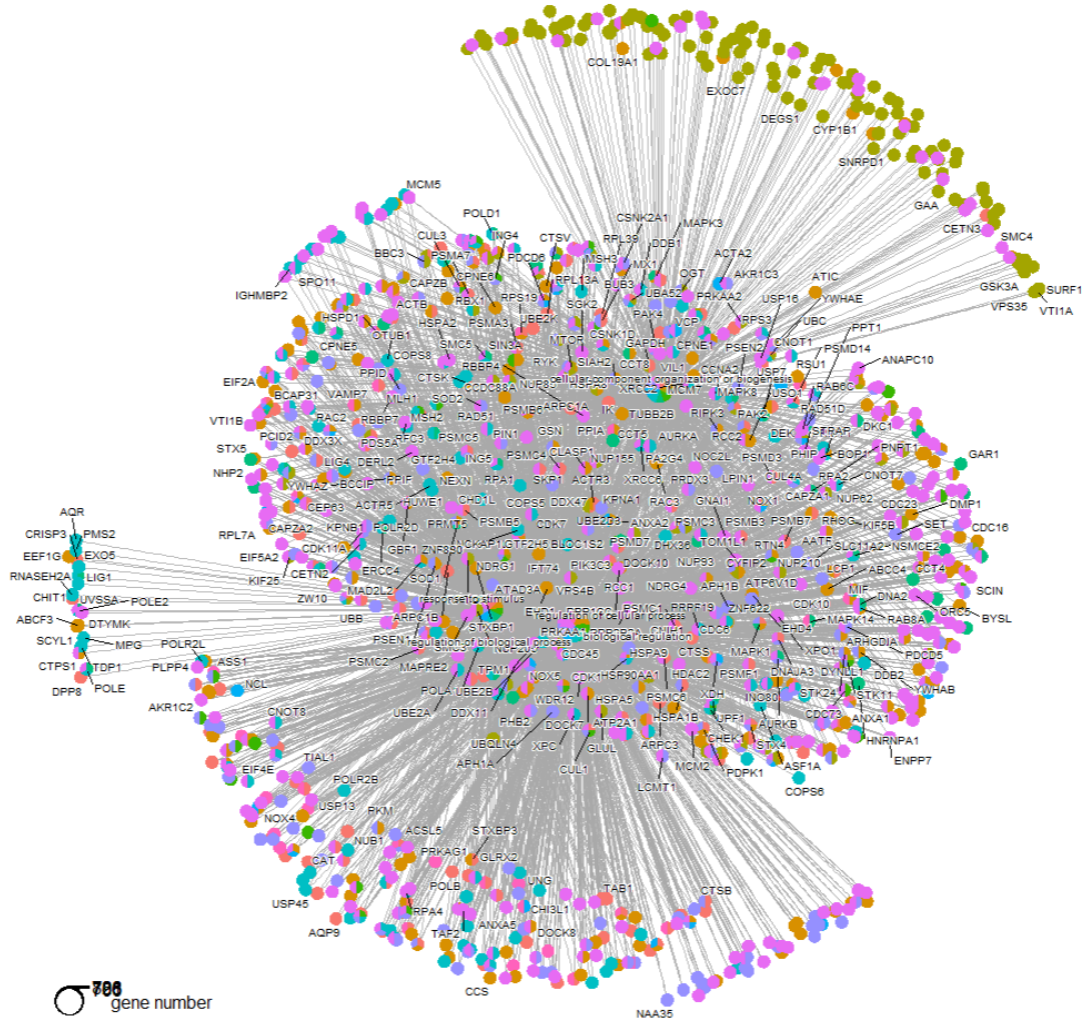
```
# Modificar archivo de resultados para correr enrichplot
gp_mod =              res_group[,c("query",
                                   "source",
                                   "term_id",
                                   "term_name",
                                   "p_value",
                                   "query_size",
                                   "intersection_size",
                                   "term_size",
                                   "effective_domain_size",
                                   "intersection")]
```

```
gp_mod$GeneRatio = paste0(gp_mod$intersection_size, "/", gp_mod$query_size)
gp_mod$BgRatio = paste0(gp_mod$term_size, "/", gp_mod$effective_domain_size)
names(gp_mod) = c("Cluster", "Category", "ID", "Description", "p.adjust",
                  "query_size", "Count", "term_size", "effective_domain_size",
                  "geneID", "GeneRatio", "BgRatio")
gp_mod$geneID = gsub(",", "/", gp_mod$geneID)
gp_mod$Cluster <- factor(gp_mod$Cluster)
```

```
# Convertir de gprofiler a clusterprofiler
gp_mod_cluster = new("compareClusterResult", compareClusterResult = gp_mod)
gp_mod_enrich = new("enrichResult", result = gp_mod)
```

Visualización de gprofiler en clusterprofiler

Visualización de gprofiler en clusterprofiler



GOSummaries

Tarea

- Ajustar el formato de genes a GOSummaries

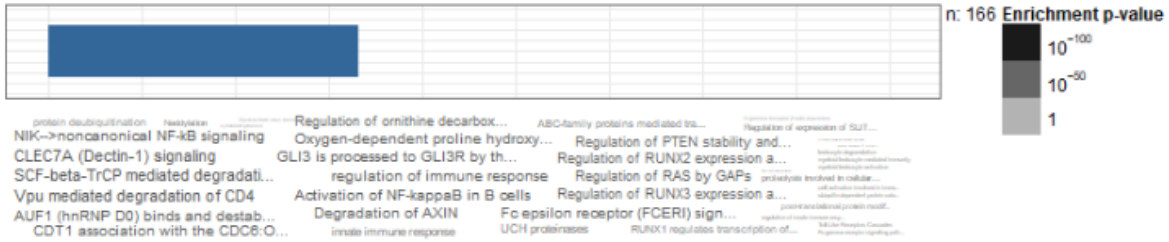
- Correr el análisis via g:Profiler y graficar

Código

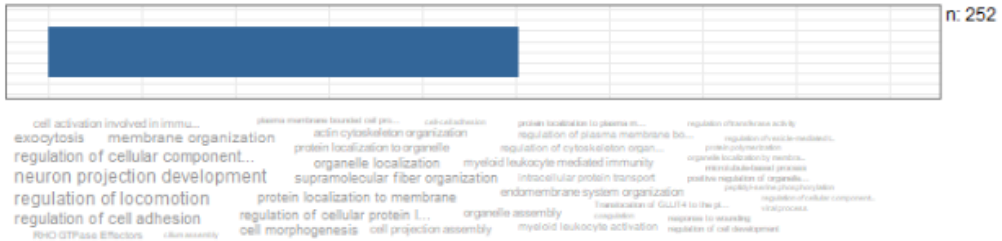
```
#Ajustar el formato a GOSummaries
x_Hsap3 <- lapply(seq_len(length(CH)), function(i){
  x_unique <- as.character(x_Hsap[[i]])
  return(x_unique)
})

names(x_Hsap3) <- CH
```

```
#Correr el analisis y graficar
gs1 = gosummaries(x_Hsap3)
plot(gs1[1:5])
plot(gs1[6:10])
```

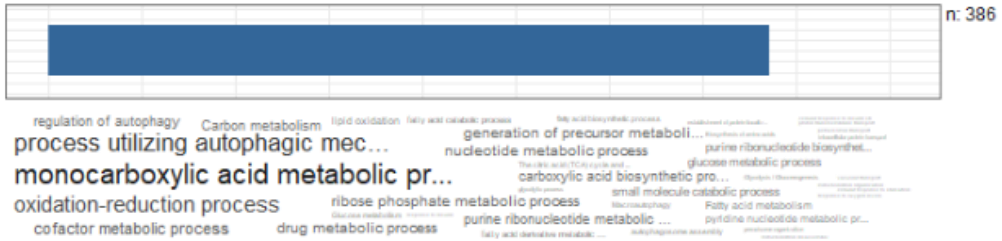
AID



AIM

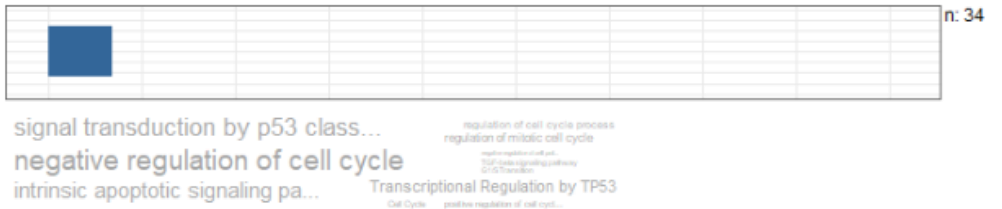


DCE

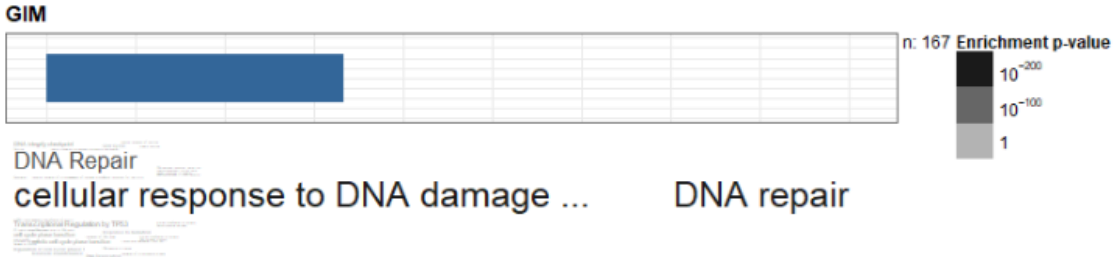


ERI

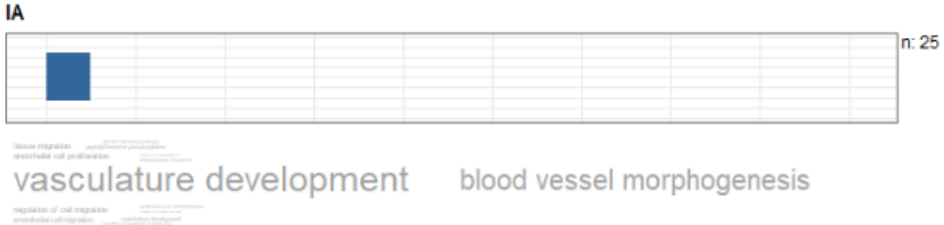
EGS



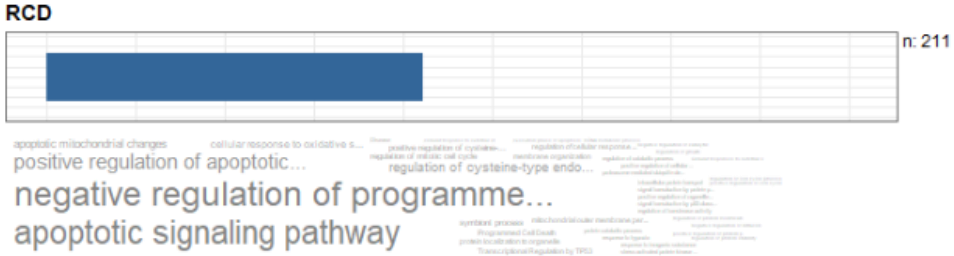
GIM



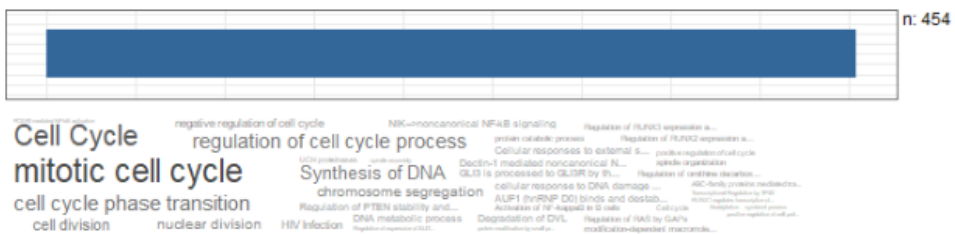
IA



RCD



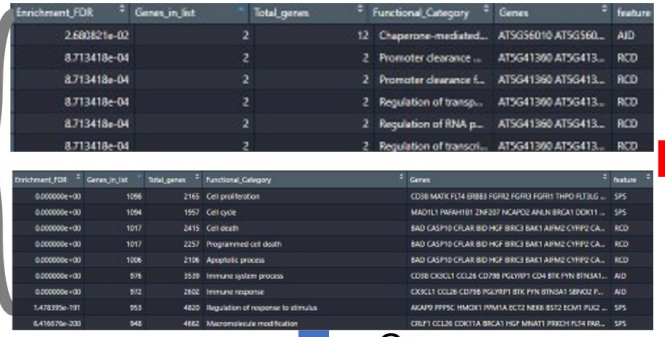
SPS



TPI

GOCompare

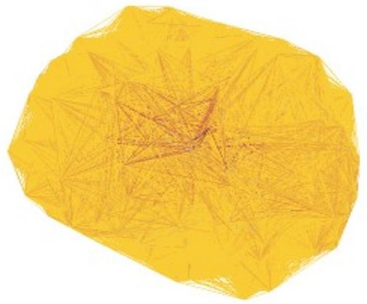
Inputs



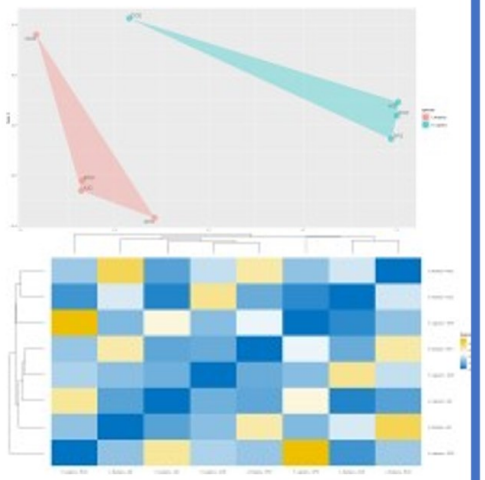
two species

One species

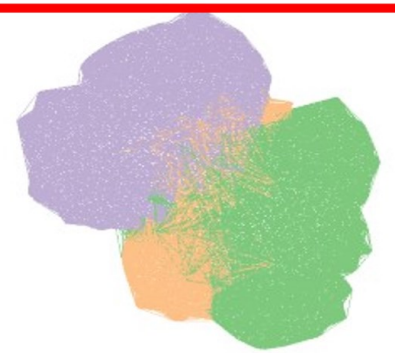
graphGOspecies



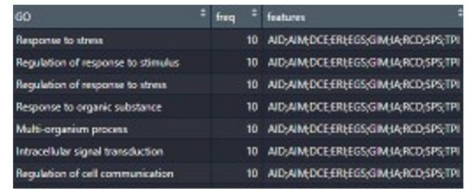
compareGOspecies



graph_two_GOspecies

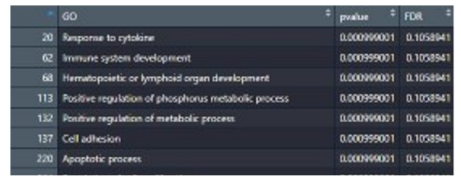


mostFrequentGOs



GO	freq	features
Response to stress	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Regulation of response to stimulus	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Regulation of response to stress	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Response to organic substance	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Multi-organism process	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Intracellular signal transduction	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Regulation of cell communication	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI

evaluateGO_species



	GO	pvalue	FDR
20	Response to cytokine	0.000999001	0.1058941
62	Immune system development	0.000999001	0.1058941
68	Hematopoietic or lymphoid organ development	0.000999001	0.1058941
113	Positive regulation of phosphorus metabolic process	0.000999001	0.1058941
132	Positive regulation of metabolic process	0.000999001	0.1058941
137	Cell adhesion	0.000999001	0.1058941
220	Apoptotic process	0.000999001	0.1058941

Lista de funciones actuales implementadas

Función	Foco	Descripción
mostFrequentGOs	Una especie	procesos biológicos más frecuentes por categoría
graphGOspecies	Una especie	Conversión una lista procesos biológicos y categorías a grafos
compareGOspecies	Dos especies	Comparación de dos listas de procesos biológicos y categorías usando PCoA y distancias de Jaccard
evaluateGO_species	Dos especies	Evaluación de la importancia de un proceso biológico para n categorías usando pruebas Chi cuadrado
graph_two_Gospecies	Dos especies	Conversión una lista procesos biológicos y categorías en dos especies a grafos

Las ideas detrás

- El enriquecimiento funcional se usa para identificar posibles funciones asociadas a una lista de genes especificada.
- Los posibles procesos biológicos asociados se representan como términos de ontología de genes (*GO terms*).
- No hay una manera sencilla de comparar *GO terms* entre especies y categorías.
- Los *GO terms* se pueden representar como grafos no dirigidos tambien.
- Se pueden otorgar pesos para cada nodo de acuerdo a cuantos GO *terms* están compartidos entre categorías de listas de genes.

Paquete GOCompare

Tareas

- Cargar los datos a ser usados

H. Sapiens y *A. thaliana*

- Definir la columna que tiene la info de los GO enriquecidos

- Nombrar las especies

- Comparar GO enriquecidos entre especies usando estadística multivariada

Código

```
#Cargar datos de ejemplo (cuatro cancer hallmarks)
data(H_sapiens_compress)
data(A_thaliana_compress)
```

```
#Definir la columna que tiene la info de los GO enriquecidos
GOterm_field <- "Functional_Category"
```

```
#Nombrar las especies
species1 <- "H. sapiens"
species2 <- "A. thaliana"
```

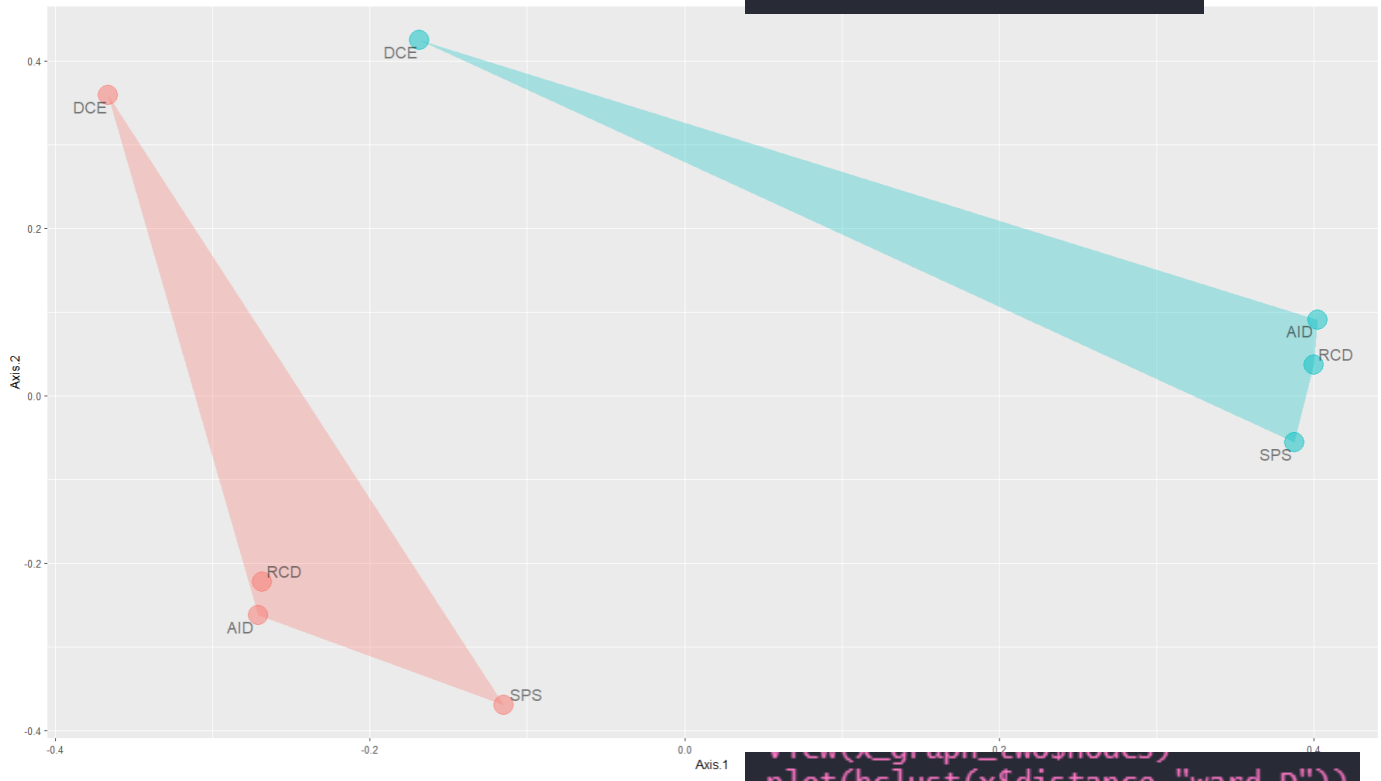
```
x <- compareGOspecies(df1=H_sapiens_compress,
                      df2=A_thaliana_compress,
                      GOterm_field=GOterm_field,
                      species1=species1,
                      species2=species2)
```

```
View(x$unique_GO_list)
View(x$shared_GO_list)
```

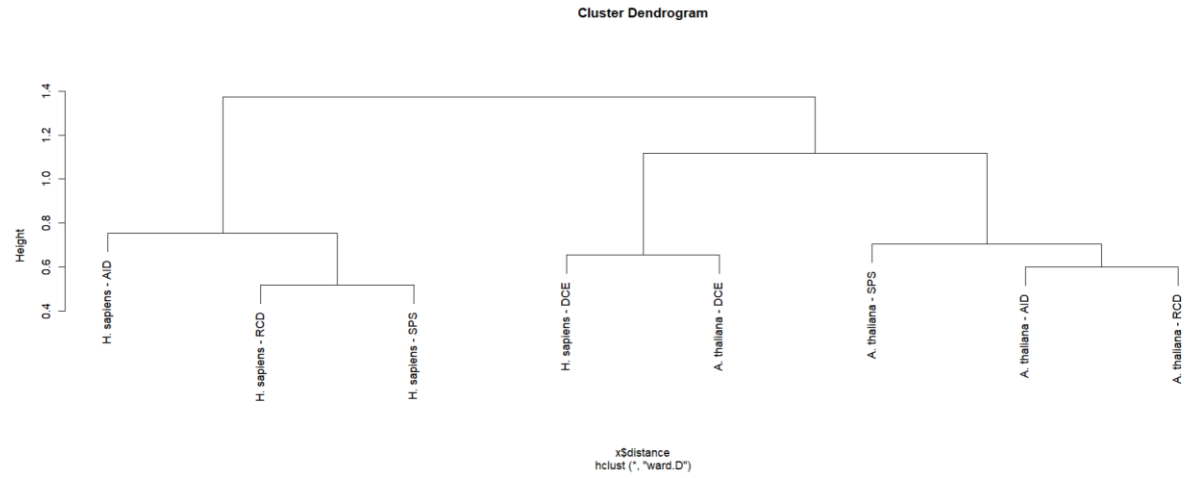
feature	GO	species
AID	Regulation of defense response	H. sapiens
AID	Immune effector process	H. sapiens
AID	Activation of immune response	H. sapiens
AID	Regulation of immune system process	H. sapiens
AID	Positive regulation of immune system process	H. sapiens
AID	Regulation of response to stimulus	H. sapiens
AID	Positive regulation of response to stimulus	H. sapiens
AID	Regulation of immune response	H. sapiens
AID	Positive regulation of immune response	H. sapiens
AID	Immune response-activating signal transduction	H. sapiens
AID	Immune response-regulating signaling pathway	H. sapiens

feature	GO	species
AID	Response to stress	Shared
AID	Defense response	Shared
AID	Immune response	Shared
AID	Immune system process	Shared
AID	Innate immune response	Shared
AID	Response to external stimulus	Shared
AID	Response to biotic stimulus	Shared
AID	Response to other organism	Shared
AID	Response to external biotic stimulus	Shared
AID	Response to organic substance	Shared
AID	Cellular response to organic substance	Shared

```
#graficar el PCoA
x$graphics
```




```
plot(hclust(x$distance,"ward.D"))
```



¿Qué GO enriquecidos son más importantes para una sola especie?

Tareas

- Comparar GO enriquecidos entre categorías para una sola especie y extraer importancia

```
#Extraer pesos para GO enriquecidos
x_graph <- graphGOspecies(df=H_sapiens_compress,
                          GOterm_field=GOterm_field,
                          option = "GO",
                          numCores=2,
                          saveGraph=FALSE,
                          outdir = NULL)
```

View(x_graph$nodes)

View(x_graph$nodes)

	GO	GO_WEIGHT
1	Response to stress	1.587908
11	Regulation of response to stimulus	1.587908
19	Regulation of response to stress	1.587908
26	Response to external stimulus	1.587908
34	Response to organic substance	1.587908
36	Cellular response to organic substance	1.587908
37	Cellular response to chemical stimulus	1.587908
41	Multi-organism process	1.587908
47	Positive regulation of cell communication	1.587908
48	Positive regulation of signaling	1.587908
49	Positive regulation of signal transduction	1.587908
57	Intracellular signal transduction	1.587908
59	Regulation of cell communication	1.587908
60	Regulation of signaling	1.587908
79	Regulation of protein metabolic process	1.587908
81	Positive regulation of intracellular signal transduction	1.587908
84	Positive regulation of protein metabolic process	1.587908
86	Interspecies interaction between organisms	1.587908

¿Qué GO enriquecidos son más importantes entre dos especies?

Tareas

- Comparar GO enriquecidos entre categorías para una sola especie y extraer importancia

```
#Extraer pesos para GO enriquecidos entre dos especies y categorias
x_graph_two <- graph_two_GOspecies(x=x,
                                   species1=species1,
                                   species2=species2,
                                   GOterm_field=GOterm_field,
                                   numCores=2,
                                   saveGraph = FALSE,
                                   option= "GO",
                                   outdir = NULL)
```

```
View(x_graph_two$nodes)
```

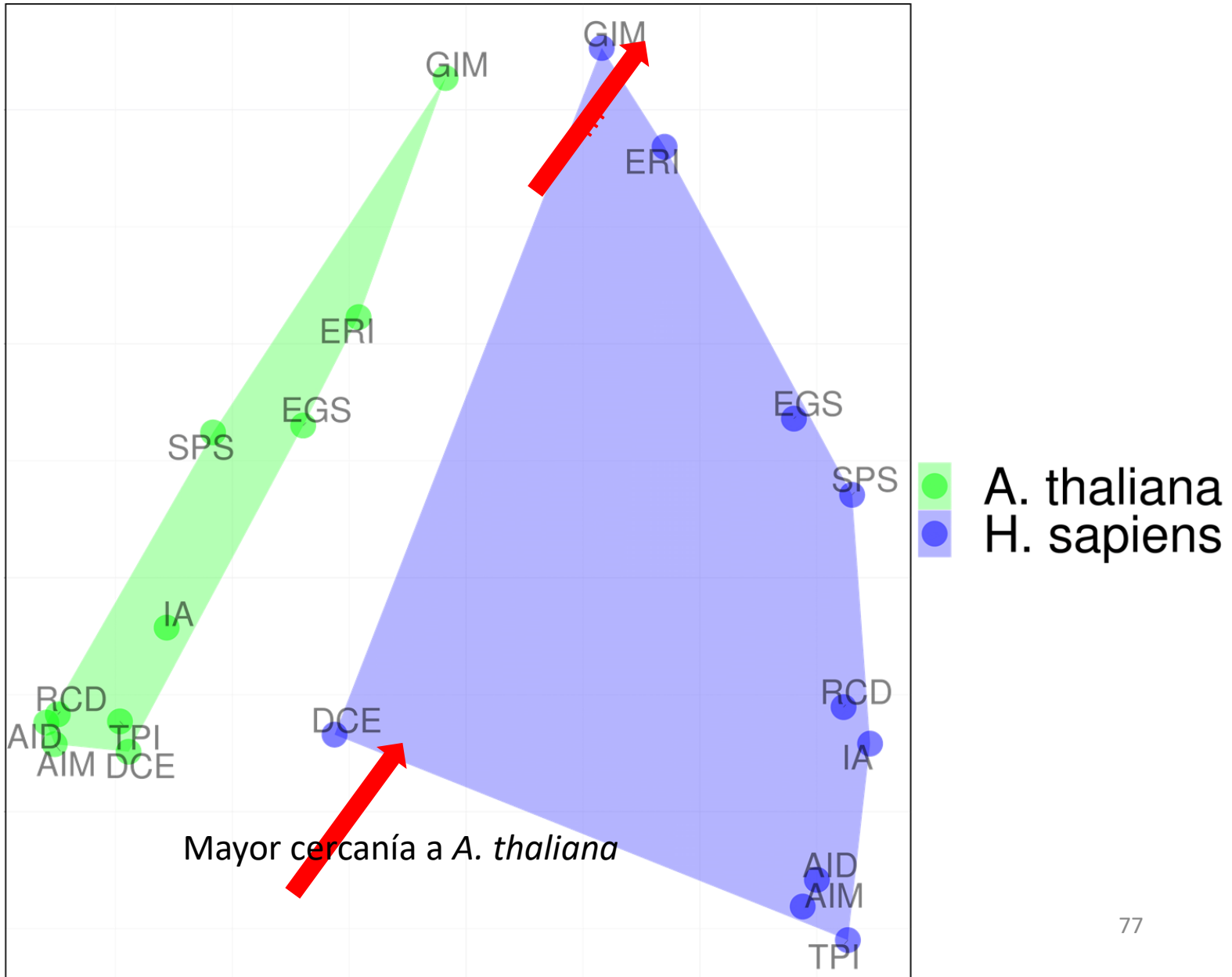
Resultado

	GO	GO_WEIGHT
123	Organic substance catabolic process	1.818718
217	Establishment of localization in cell	1.818718
121	Organic substance transport	1.795058
138	Nitrogen compound transport	1.795058
222	Intracellular transport	1.795058
225	Intracellular protein transport	1.795058
279	Response to temperature stimulus	1.795058
283	Response to heat	1.795058
285	Organonitrogen compound catabolic process	1.795058
297	Response to acid chemical	1.795058
78	Positive regulation of phosphorylation	1.779604
82	Cellular catabolic process	1.670305
84	Catabolic process	1.670305
190	Macromolecule localization	1.670305
193	Establishment of protein localization	1.670305

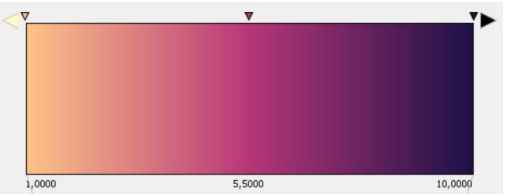
Un caso de la vida real:
¿Se puede usar *A. thaliana* para estudiar cáncer?

Caso de estudio: Comparación de genes asociados a cáncer y sus ortólogos en *A. thaliana*

- 5494 genes de *H. sapiens*
- 2223 genes ortólogos en *A. thaliana*
- ¡Muchos *GO terms* únicos!
- Alta incidencia de procesos biológicos asociados a respuesta al estrés y ciclo celular
- Cuatro CH priorizados:
  - ***Activating Invasion Motility (AIM)***
  - ***Deregulating Cellular Energetics (DCE)***
  - ***Resisting Cell Death (RCD)***
  - ***Sustaining Proliferative Signaling (SPS)***



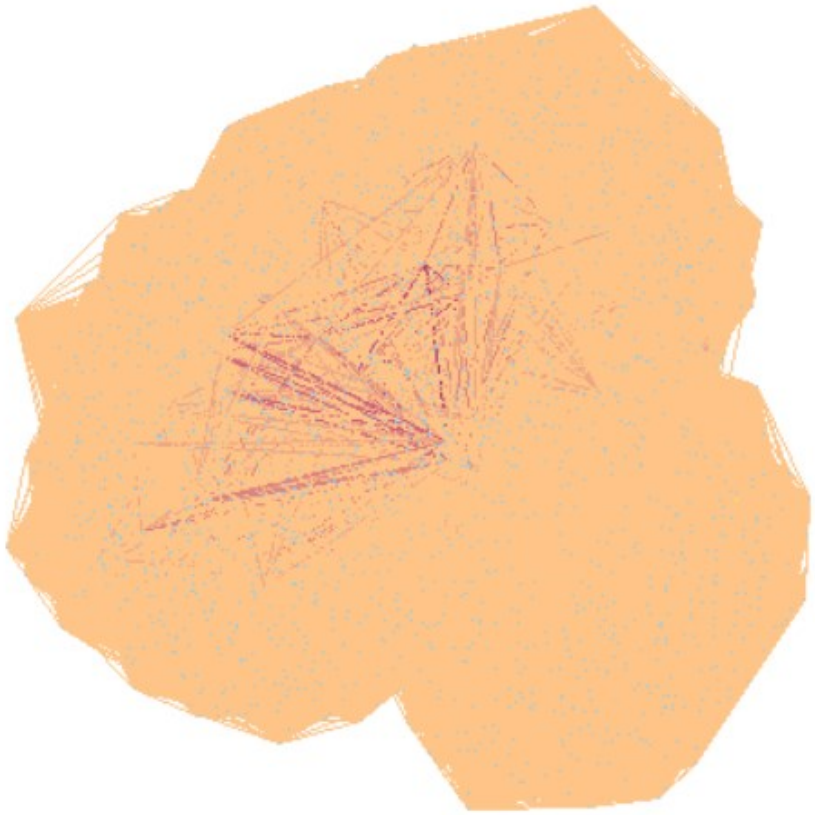
Comparación de procesos biológicos entre categorías para humano



Categorías compartidas por BP

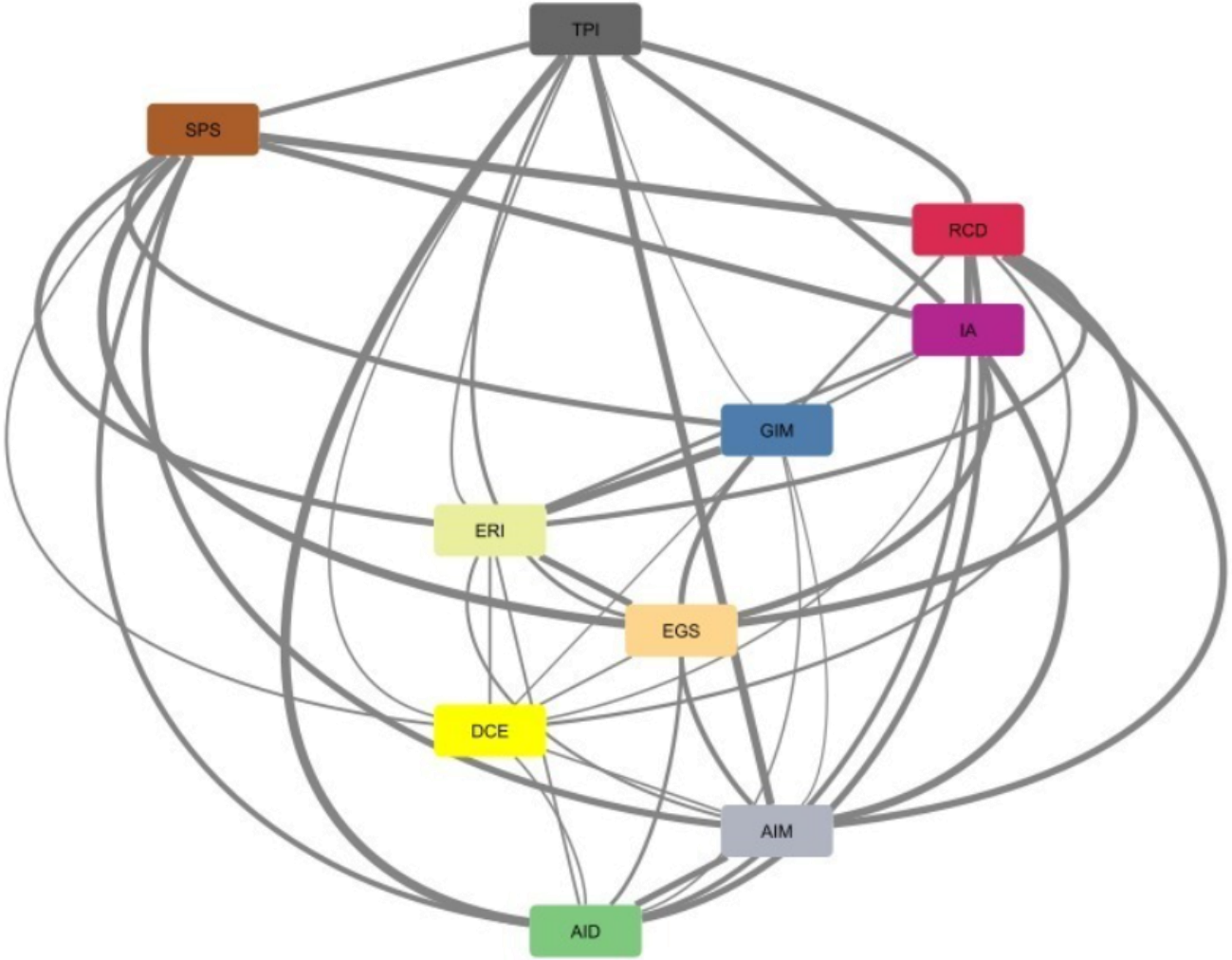
- Nodos: 2001 procesos biológicos
- Aristas:888508 interacciones
- Alta incidencia de procesos biológicos asociados a respuesta a estrés y ciclo celular

shared name	name	GO_WEIGHT
Response to stress	Response to ...	2.49375312343828
Regulation of response to stimulus	Regulation o...	2.49375312343828
Regulation of response to stress	Regulation o...	2.49375312343828
Response to organic substance	Response to ...	2.49375312343828
Multi-organism process	Multi-organis...	2.49375312343828
Intracellular signal transduction	Intracellular ...	2.49375312343828
Regulation of cell communication	Regulation o...	2.49375312343828
Regulation of signaling	Regulation o...	2.49375312343828
Regulation of protein metabolic process	Regulation o...	2.49375312343828
Positive regulation of protein metabolic process	Positive regu...	2.49375312343828
Phosphorylation	Phosphorylat...	2.49375312343828
Positive regulation of cellular protein metabolic process	Positive regu...	2.49375312343828
Regulation of intracellular signal transduction	Regulation o...	2.49375312343828
Regulation of cellular protein metabolic process	Regulation o...	2.49375312343828
Positive regulation of metabolic process	Positive regu...	2.49375312343828
Positive regulation of cellular metabolic process	Positive regu...	2.49375312343828
Regulation of phosphorylation	Regulation o...	2.49375312343828



Comparación de categorías como nodos

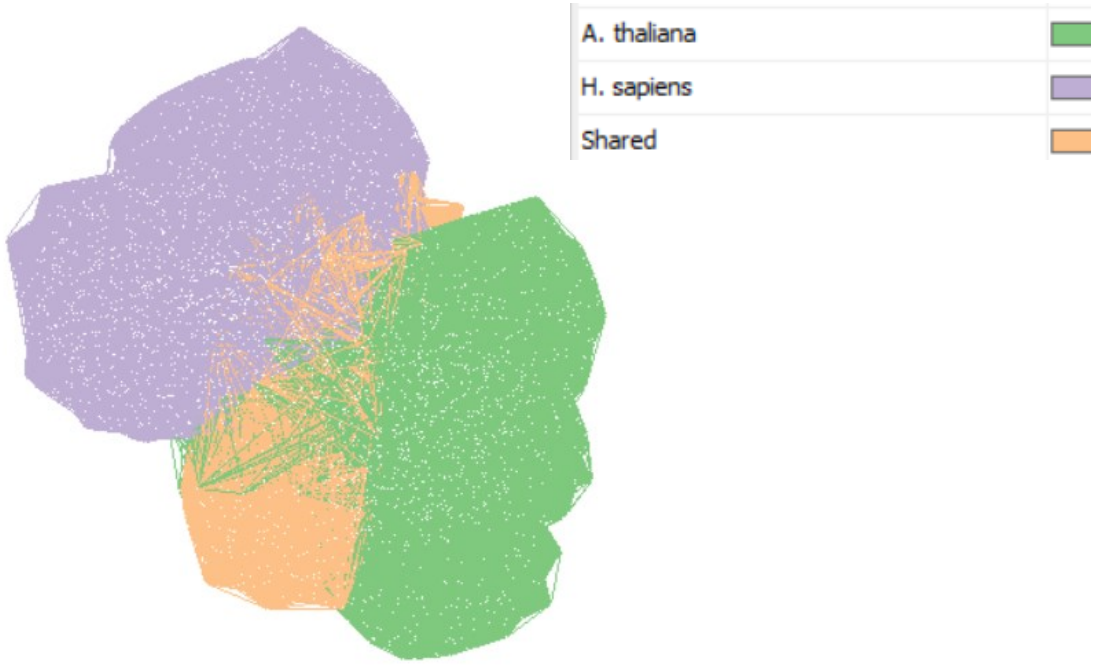
- Nodos:10 categorías
- Aristas: 52 interacciones entre categorías
- Hay un gran numero de procesos biológicos compartidos entre categorías, en especial para TPI, IA,RCD y AIM



Comparación de procesos biológicos entre especies

- Nodos:2863 procesos biológicos
- Aristas: 870003 interacciones entre categorías
- Los pesos se calculan para cada subgrafo (es decir por especie y si se comparten)
- Alta incidencia de procesos biológicos asociados a respuesta a estrés y regulación metabólica

$$K_w(U) = \sum_{1}^{k} \sum_{1}^{n} w(U, V)$$



shared name	GO_WEIGHT	name
Positive regulation of catalytic activity	2.57658916083916	Positive regu…
Positive regulation of metabolic process	2.5451013986014	Positive regu…
Positive regulation of cellular metabolic process	2.5451013986014	Positive regu…
Positive regulation of nitrogen compound metabolic process	2.5451013986014	Positive regu…
Regulation of response to stress	2.51456381118881	Regulation o…
Cellular component assembly	2.44154241904418	Cellular comp…
Cellular component biogenesis	2.44154241904418	Cellular comp…
Positive regulation of protein modification process	2.4270506993007	Positive regu…
Positive regulation of molecular function	2.4270506993007	Positive regu…
Positive regulation of cellular protein metabolic process	2.4248513986014	Positive regu…
Positive regulation of macromolecule metabolic process	2.39556293706294	Positive regu…

El mensaje a llevar a casa

- **El análisis de enriquecimiento funcional es un recurso vital en la bioinformática moderna pero debe usarse con cuidado**
- **Aunque los análisis son relativamente fáciles de realizar, se requiere un conocimiento previo de los genes para no caer en trampas**
- **R es bastante útil a la hora de realizar análisis bioinformáticos**
- **Se presenta a consideración una nueva metodología que puede ser interesante para el campo de la genómica comparativa**

ómicas

Aliados

CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

cenicaña

FEDEARROZ
FEDERACIÓN NACIONAL DE ARROCEROS

NIAB

intelecto
Soluciones y Tecnología

HITECH
AUTOMATIZACION S.A.

Caltech

GHENT UNIVERSITY

東京大学
THE UNIVERSITY OF TOKYO

ILLINOIS

Pontificia Universidad JAVERIANA Cali

IES Ancla

UNIVERSIDAD DEL QUINDÍO

Universidad de Ibagué

UNILLANOS
UNIVERSIDAD DE LOS LLANOS

Universidad de los Andes
Colombia

UNIVERSIDAD ICESI

Pontificia Universidad JAVERIANA Bogotá

Apoyan

El futuro es de todos
Gobierno de Colombia

COLOMBIA CIENTÍFICA
Conocimiento Global para el Desarrollo