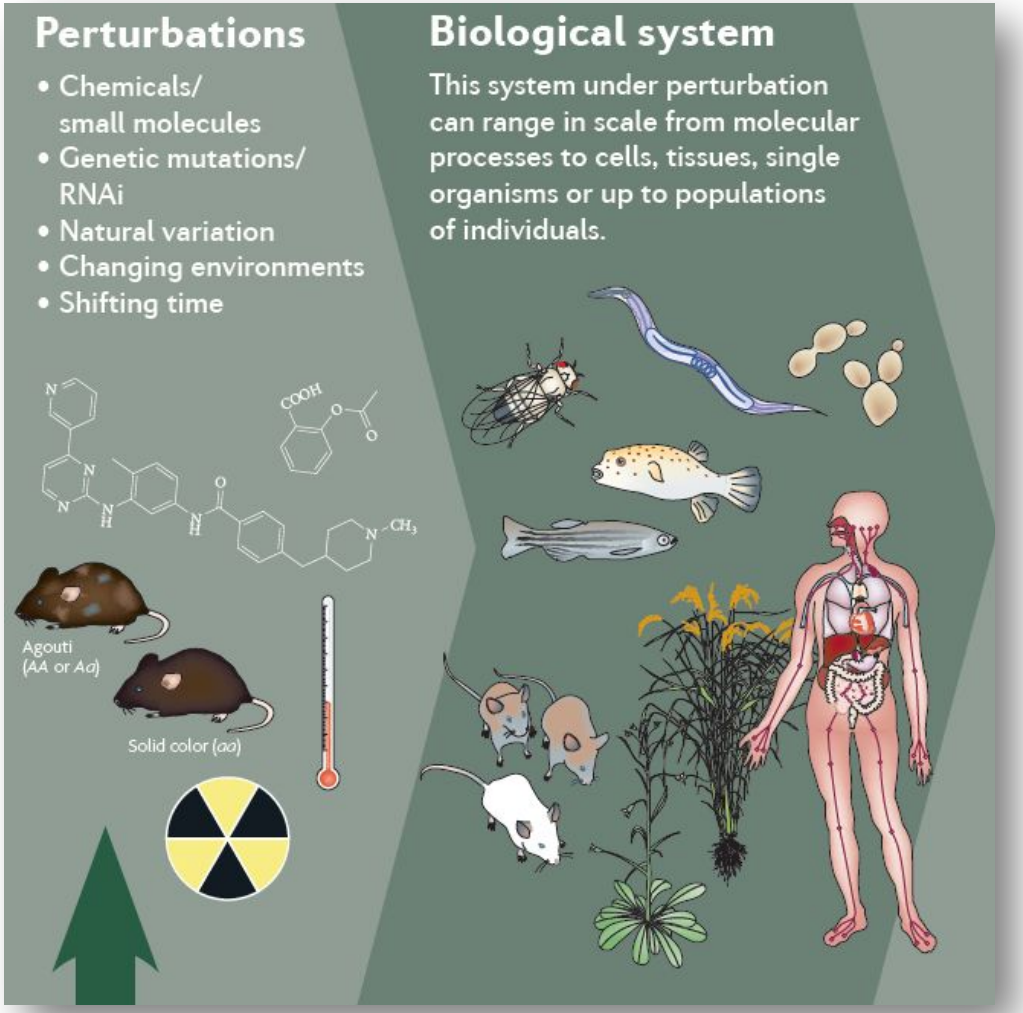
Redes. Clusters. Biologia.

v1.2

Contexto: Biología de Sistemas Integrativa

Estudio de sistemas biológicos mediante

- Análisis de respuestas globales ante perturbaciones sistemáticas
- Integración de datos en modelos predictivos



Ideker & Bandyopadhyay *Integrative Systems Biology* Nature Genetics 2010

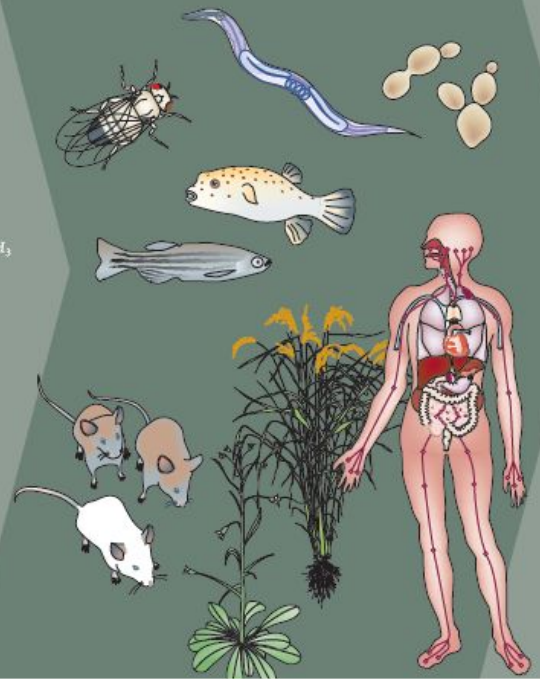
Biología de Sistemas Integrativa

Perturbations

- Chemicals/ small molecules
- Genetic mutations/ RNAi
- Natural variation
- Changing environments
- Shifting time

Biological system

This system under perturbation can range in scale from molecular processes to cells, tissues, single organisms or up to populations of individuals.



Molecular state measurements (nodes)

	Large-scale dataset types	Technologies
Genome	Whole-genome DNA sequences, SNPs and CNVs	DNA sequencing, genotyping microarrays
Epigenome	Chromatin modifications and structure	ChIP-seq, methyl-seq, DHS-chip
Transcriptome	Transcript abundances, translation rate and microRNAs	DNA microarrays, RNA-seq, CAGE, GRO-seq, ribosome profiling
Proteome	Protein abundances and modifications	NMR, mass spectrometry, multiparameter FACS
Metabolome	Metabolite profiling	Mass spectrometry, liquid chromatography

Molecular interaction measurements (edges)

	Data types	Technologies
Physical	Protein–protein	Immuno-precipitation (IP), co-affinity purification, yeast two-hybrid, protein arrays, kinase-substrate measurements
	Protein–DNA, protein–RNA	Genome-wide chromatin immuno-precipitation (ChIP), DNA binding arrays
	Protein–small molecule, reaction fluxes	Isotope labeling, mass spectrometry
Genetic and functional	Synthetic lethality, epistasis	Synthetic genetic arrays (SGA) combinatorial RNAi, population genetics
	Cause–effect relationships	Genetic perturbation (gene knockout, RNAi) followed by phenotyping (microarrays, cellular imaging); *trans* eQTLs (expression quantitative trait loci)

Ideker & Bandyopadhyay *Integrative Systems Biology* Nature Genetics 2010

Biología de Sistemas Integrativa

Molecular state measurements (nodes)

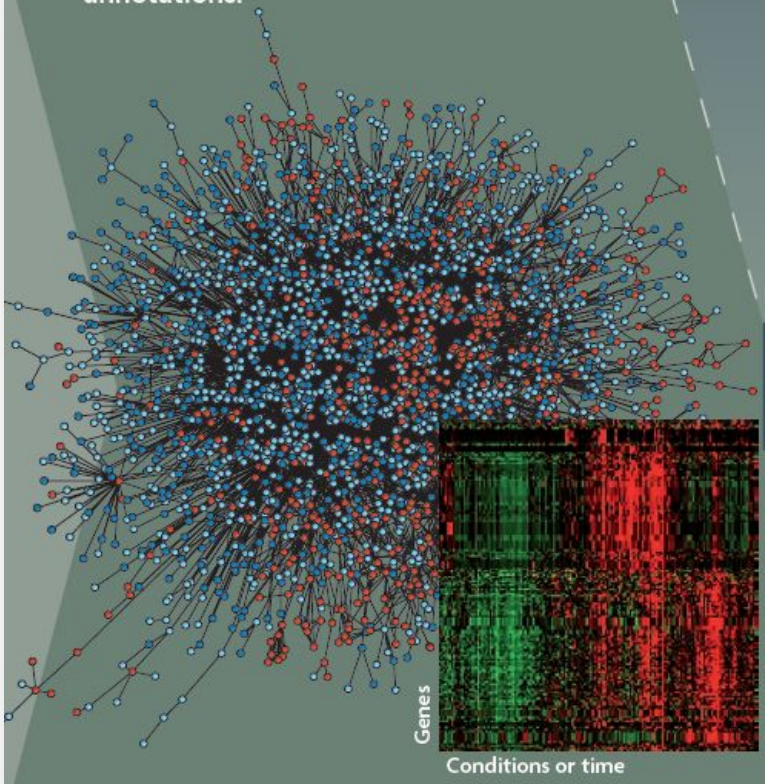
	Large-scale dataset types	Technologies
Genome	Whole-genome DNA sequences, SNPs and CNVs	DNA sequencing, genotyping microarrays
Epigenome	Chromatin modifications and structure	ChIP-seq, methyl-seq, DHS-chip
Transcriptome	Transcript abundances, translation rate and microRNAs	DNA microarrays, RNA-seq, CAGE, GRO-seq, ribosome profiling
Proteome	Protein abundances and modifications	NMR, mass spectrometry, multiparameter FACS
Metabolome	Metabolite profiling	Mass spectrometry, liquid chromatography

Molecular interaction measurements (edges)

	Data types	Technologies
Physical	Protein–protein	Immuno-precipitation (IP), co-affinity purification, yeast two-hybrid, protein arrays, kinase-substrate measurements
	Protein–DNA, protein–RNA	Genome-wide chromatin immuno-precipitation (ChIP), DNA binding arrays
	Protein–small molecule, reaction fluxes	Isotope labeling, mass spectrometry
Genetic and functional	Synthetic lethality, epistasis	Synthetic genetic arrays (SGA) combinatorial RNAi, population genetics
	Cause–effect relationships	Genetic perturbation (gene knockout, RNAi) followed by phenotyping (microarrays, cellular imaging); *trans* eQTLs (expression quantitative trait loci)

Molecular databases

New measurements are stored alongside existing data, including functional annotations.



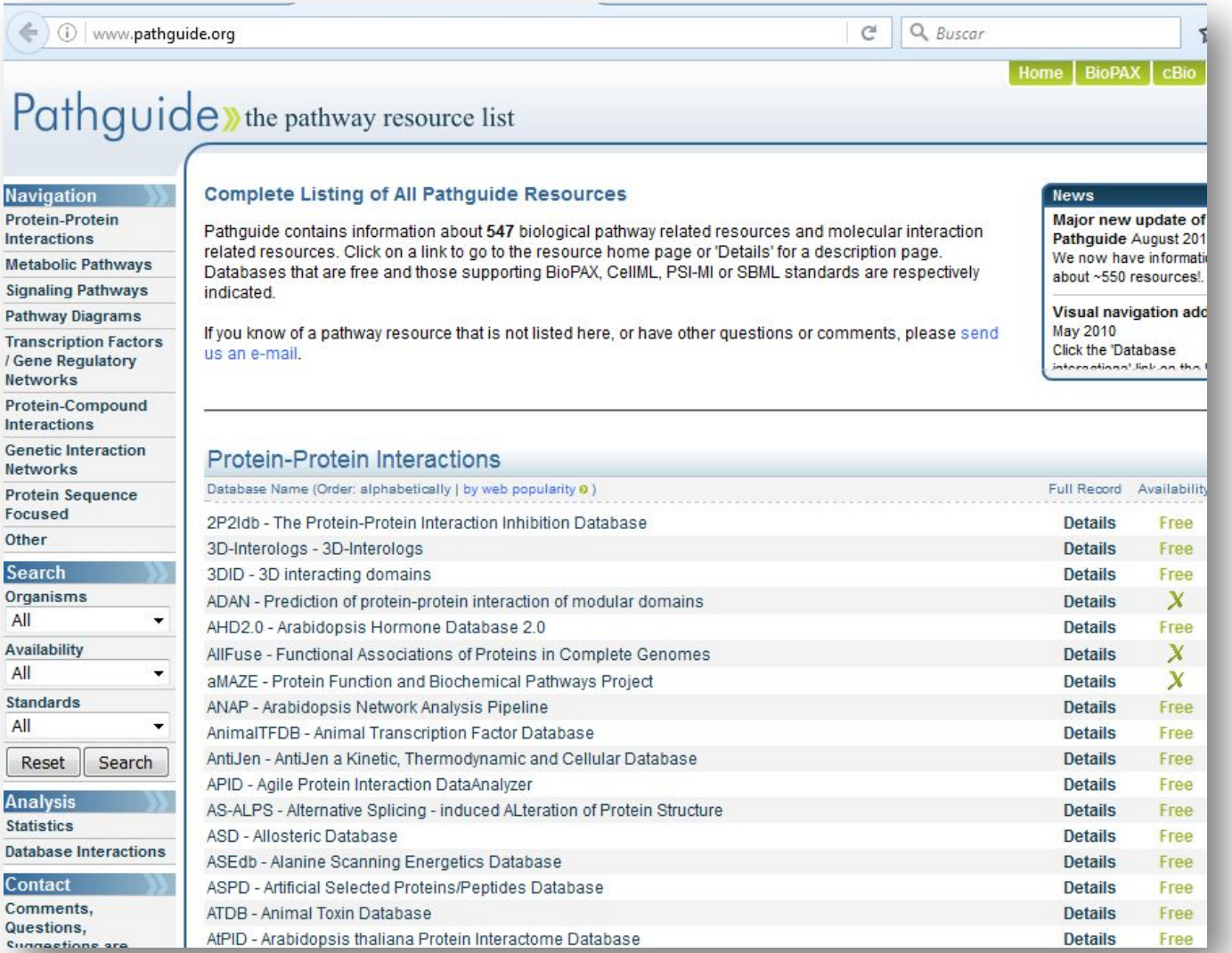
	Annotations	States	Interactions
Selected molecular databases	GO, KEGG, REACTOME, NetPath, UCSC Genome Browser	Genbank, GEO, ArrayExpress, Proteome Commons	BioGRID, HPRD, IntAct, TRANSFAC, STRING, iHOP, functionalnet.org

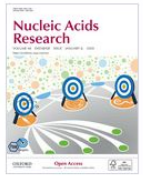
Conocimiento Biológico

	Interaction type(s)	Detection methodologies	Databases
Physical	Protein-protein	Yeast-2-hybrid (Y2H)[1-3], co-immuno-precipitation (Co-IP)[4], mass spectroscopy[5,6]}, affinity purification coupled with mass spectroscopy (AP/MS)[7,8]	BioGRID[9], IntAct[10], APID[11], STRING[12], MINT[13], DIP[14], HPRD[15], MIPS-MPPI[16], Netpath[17], DroiD[18]
	Protein-DNA (e.g., regulatory networks)	Yeast-1-hybrid (Y1H)[19], chromatin immuno-precipitation based methods (CHIP-CHIP)[20], DNA-footprinting[21]	TRANSFAC[22], UniProbe[23], DroiD[18], BioGRID[9], TcoF-DB[24], BIPA[25], hPDI[26], EDGEdb[27], NPIDB[28,29]
	Protein-RNA	RNA electro-mobility shift (RNA-EMSA)[30], RNA-pull down[31]	PRID[32], BIPA[25] NPInter[33], RBPDB[34], PRIDB[35], StarBase[36]
	Metabolic (e.g., enzyme-substrate, ligand-receptor)	Mass spectroscopy based selective reaction monitoring (SRM)[6,37], NMR[38], affinity purification[8], co-IP[3], fluorescence spectroscopy[39]	Reactome[40], KEGG [41], BioCyc and MetaCyc[42], HMDB[43,44], EcoCyc[45], HumanCyc[46], ConsensusPathDB[47]
	Protein/gene-compound (e.g., drug-target, chemical-protein)	Chemical structure[48,49], forward or reverse chemo-genomic/proteomic profiling[50-52], in silico predictions[53]	SuperTarget[54], Matador[54], DrugBank[55] ChemProt[56], STITCH[57], AffinDB[58], MatrixDB[59], PSMDB[60], PDB-Ligand[61], ChEMBL[62], ConsensusPathDB[47]
Functional	Genetic (gene-gene)	Synthetic genetic array (SGA)[63], Epistatic Miniarray Profiling (E-MAP)[64], co-expression profiling[65,66]	BioGRID[9], DRYGIN[67], CYGD[68], DroID[18], ConsensusPathDB[47]
	Gene-Disease	Literature curation, clinical and sequence information	OMIM[69], HuDiNe[70], Diseasome[71]
Omics data type		**Detection methodologies**	**Databases**
Transcriptomics		Microarray[72], RNASeq[73]	GEO[74], SMD[75], TCGA[76], GXD[77], ONCOMINE[78], ArrayExpress[79],
RNAi (phenomics)		RNAi interference assay[80]	RNAiDB[81], GenomeRNAi[82], siRecords[83]
Epigenomics		Methylation profiling[84]	DAnCER[85], DiseaseMeth[86], PubMeth[87], MethDB[88], MethCancerDB[89], MethyCancer[90]
Mutation / SNP		SNP Array[91], genome sequencing[92]	TCGA[76], dbSNP[93], dbQSNP[94], GWAS Central [95], OMIM[69]
Proteomics		CHIP [11,2,96], Mass Spectrometry[5,6]	PDB[97], ExPASy[98], InterPro[99], World-2DPage[100], JASPAR[101]
Phosphorylation profile		Mass Spectrometry[5], literature curation	PhosphoGRID[102], PhospoELM[103], PHOSIDA[104]

Mitra 2013 Nat Review Integrative approaches for finding modular structure in biological networks

Portal de Conocimiento Biológico





Volume 48, Issue D1
08 January 2020
Cover image

ISSN 0305-1048
EISSN 1362-4962

Volume 48, Issue D1, 08 January 2020

EDITORIAL

The 27th annual Nucleic Acids Research database issue and molecular biology database collection

Daniel J Rigden, Xosé M Fernández

Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D1–D8, https://doi.org/10.1093/nar/gkz1161

Abstract ▾ View article

MAJOR MULTI-DATABASE RESOURCES

Database resources of the National Center for Biotechnology Information

Eric W Sayers, Jeff Beck, J Rodney Brister, Evan E Bolton, Kathi Canese ...

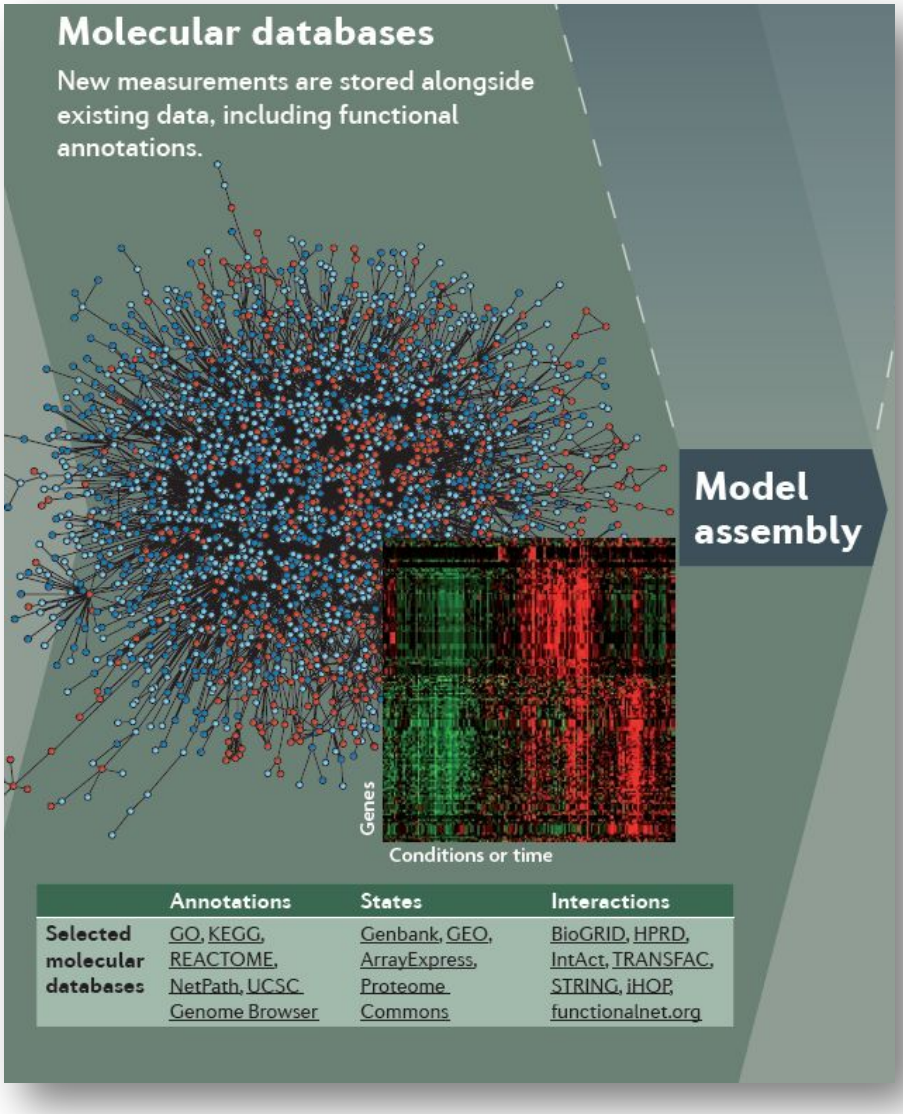
Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D9–D16, https://doi.org/10.1093/nar/gkz899

Abstract ▾ View article

The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences

Charles E Cook, Oana Stroe, Guy Cochrane, Ewan Birney, Rolf Apweiler

Biología de Sistemas Integrativa



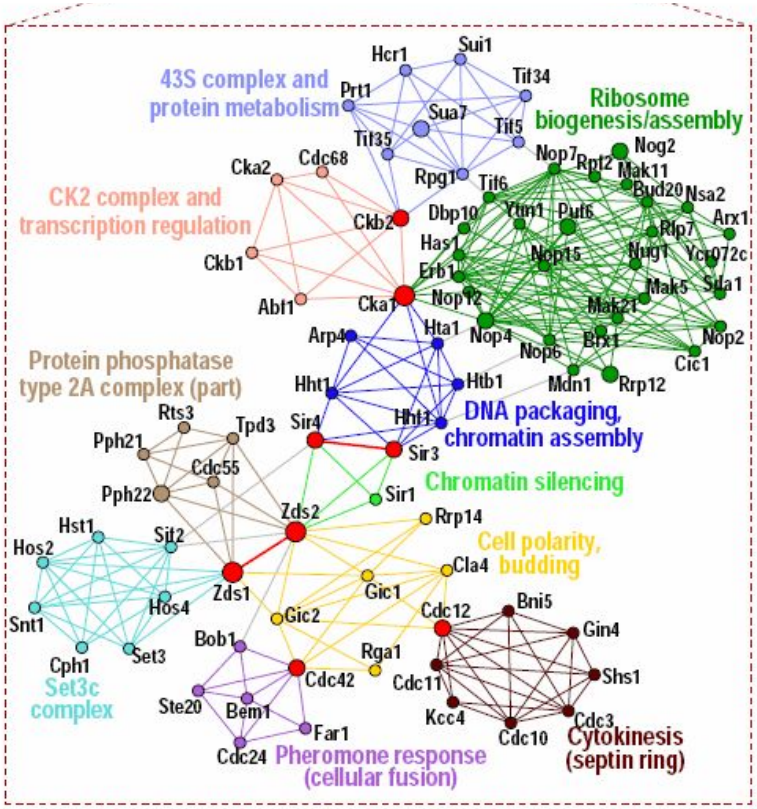
Análisis estadístico

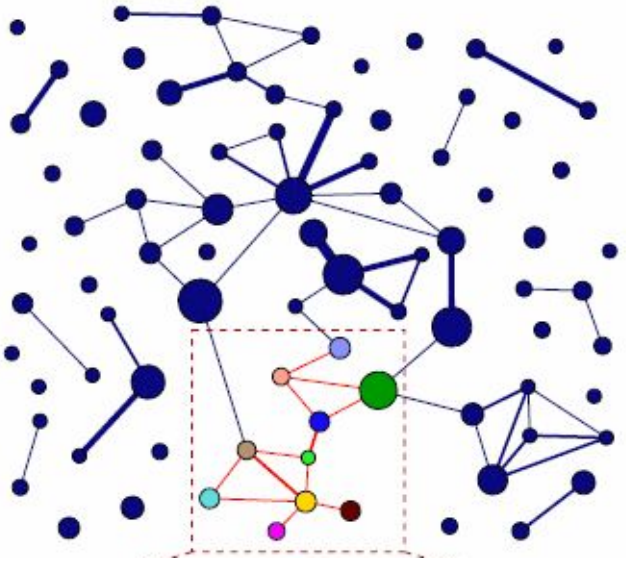
- Clustering de datos
- Overlaps y correlaciones en datasets
- Enriquecimiento funcional

Modelado

- Inferencia estadística de interacciones a partir de datos
- Modelado de sistemas dinamicos y simulaciones biofísicas

Enriquecimiento funcional





- **MsigDB**
- **Gene Ontology**
- **DAVID**

Molecular **Sig**natures **D**ata**B**ase

MSigDB
Molecular Signatures Database

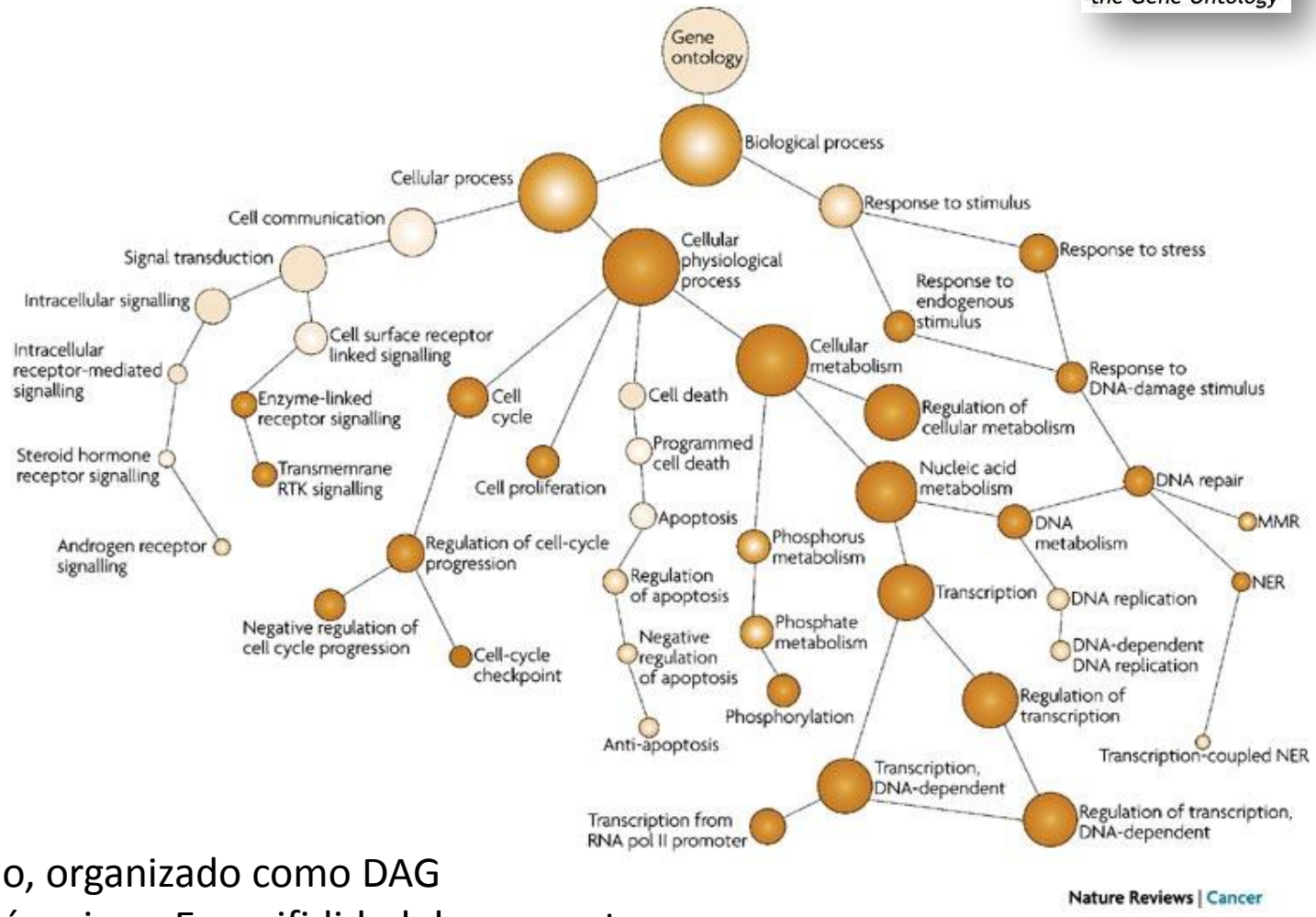
- Conjunto de **bolsas de genes** armadas bajo algún criterio relevante desde el punto de vista biológico.
- Conocimiento biológico se encuentra embebido en la composición de cada conjunto

The MSigDB gene sets are divided into 8 major collections:

H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

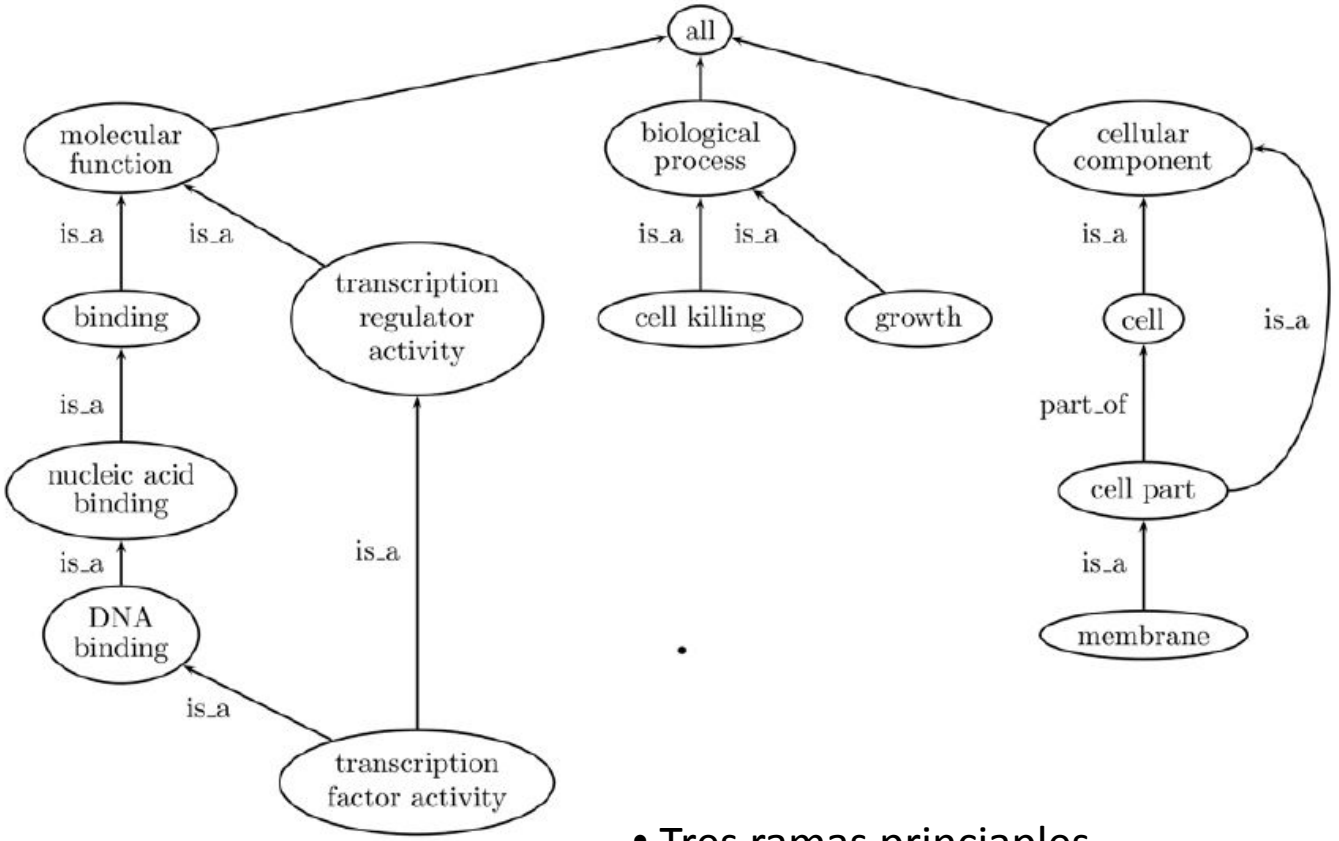
http://software.broadinstitute.org/gsea/msigdb

Gene **O**ntology



- Vocabulario controlado, organizado como DAG
- Organización cuasi-jerárquica – Especifididad de conceptos
- Productos génicos son mapeados a un (o más) nodos específicos y automáticamente son tambíen anotados en nodos ancestros de la ontología

www.geneontology.org



- Tres ramas principales
  - **M**olecular **F**unction (Qué hace?)
  - **C**ellular **C**omponent (Dónde lo hace?)
  - **B**iological **P**rocess (Por qué lo hace?)

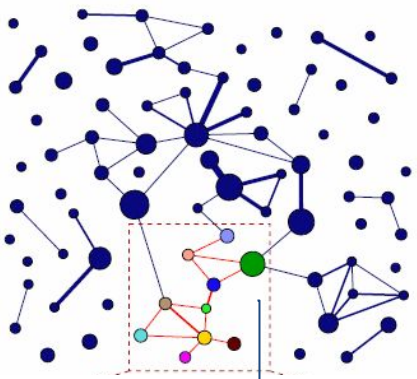
Analisis de sobre-representación

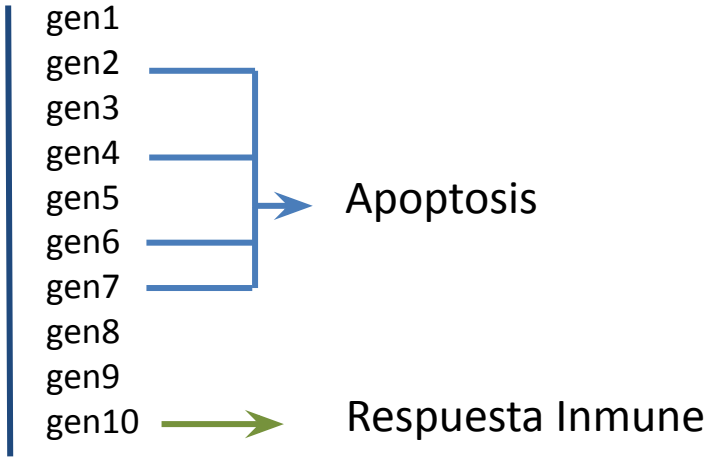
Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerepresentado en alguna comunidad o cluster detectado en mis datos?

Existe un **vinculo** entre pertenecer a una **cjto de genes identificado en mi experimento** y estar anotado dentro de un **tema biologico**?

Test de Fisher Exacto

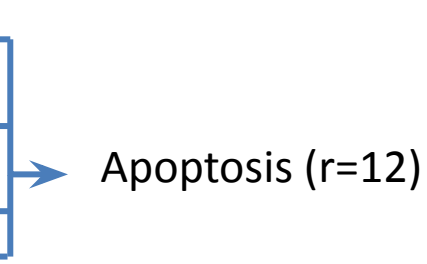
Que tan probable es observar un dado numero de hits en la lista solo por azar

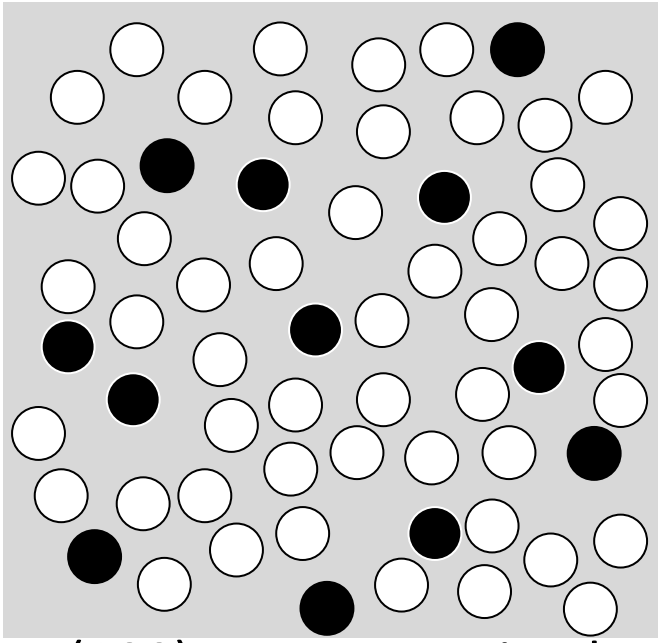




ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10





N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar

O, alternativamente…

existe una relacion en el sentido estadistico entre pertenecer a la lista de genes detectados y pertenecer a la categoria apoptosis?

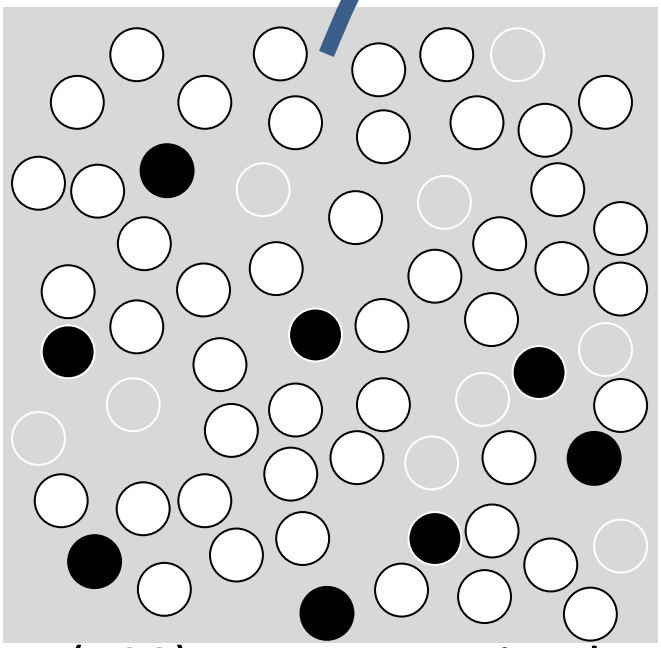
ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

Apoptosis (r=12)

Test de Fisher Exacto

Que tan probable es observar un dado número de hits en la lista **sólo por azar**



N (100) genes examinados

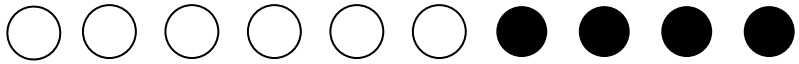
ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

Apoptosis (r=12)

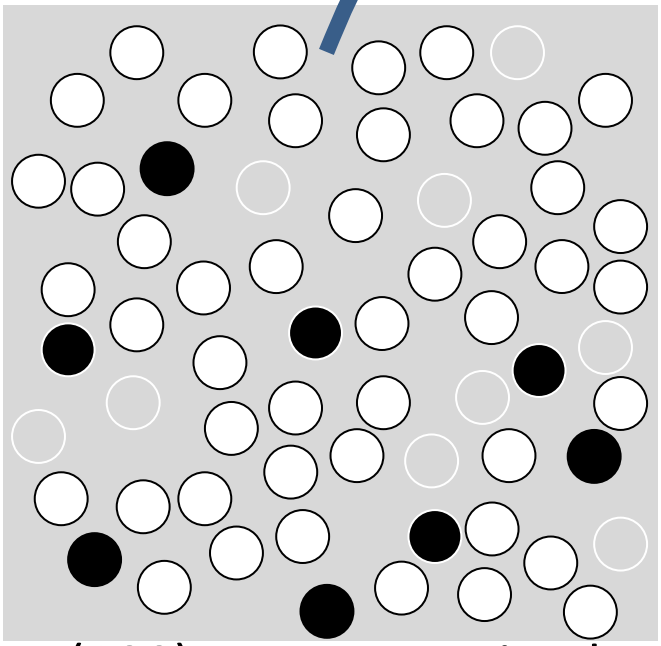
Test de Fisher Exacto

Que tan probable es observar un dado número de hits en la lista **sólo por azar**



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica



N (100) genes examinados

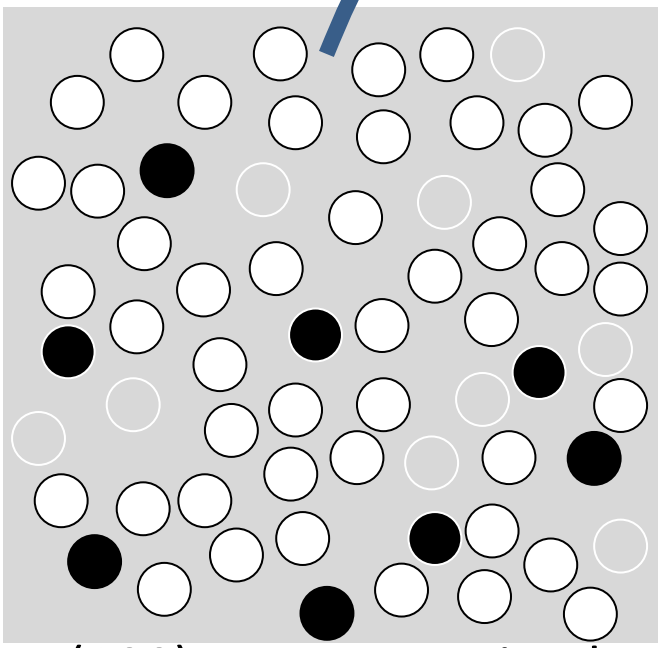
	categoria bio	$\overline{\text{categoria bio}}$	
cluster	4	6	10
$\overline{\text{cluster}}$	8	82	90
	12	88	100

	categoria bio	$\overline{\text{categoria bio}}$	
cluster	**m**	k-m	**k**
$\overline{\text{cluster}}$	r-m	N-k+m-r	N-k
	r	N-r	**N**

ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

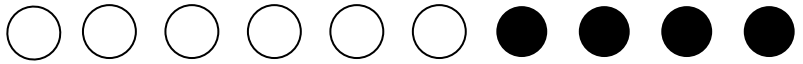
Apoptosis (r=12)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado número de hits en la lista **sólo por azar**



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

	categoria bio	$\overline{\text{categoria bio}}$	
cluster	**m**	k-m	**k**
$\overline{\text{cluster}}$	r-m	N-k+m-r	N-k
	r	N-r	**N**

Cual es la probabilidad de ver esta tabla o alguna **más extrema** si saco las bolas al azar? (i.e. no hay conexión alguna entre pertenencia al cluster y a la categoría biológica?)

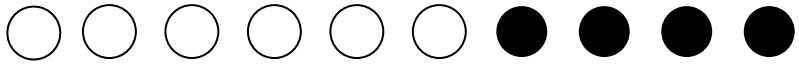
ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

Apoptosis (r=12)

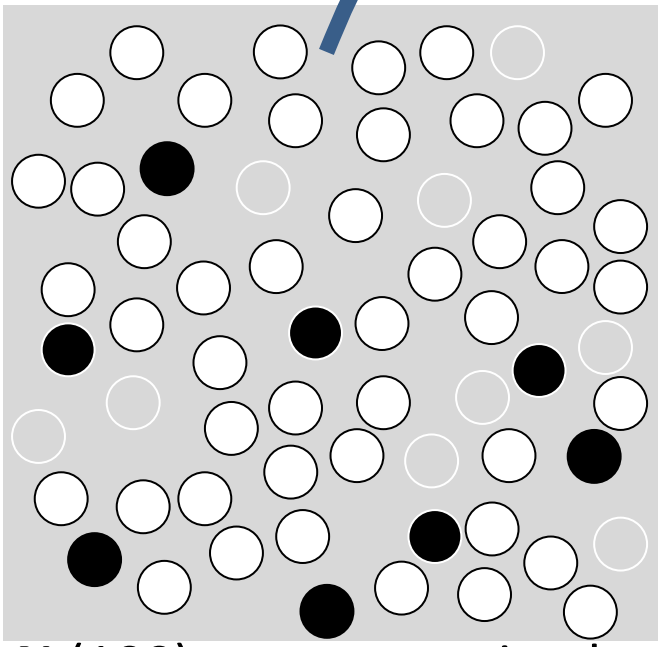
Test de Fisher Exacto

Que tan probable es observar un dado número de hits en la lista **sólo por azar**



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica



N (100) genes examinados

	categoria bio	$\overline{\text{categoria bio}}$	
cluster	**m**	k-m	**k**
$\overline{\text{cluster}}$	r-m	N-k+m-r	N-k
	r	N-r	**N**

Cual es la probabilidad de ver esta tabla?

de maneras de elegir las bolas negras

$$p = \frac{\binom{r}{m}\binom{N-r}{k-m}}{\binom{N}{r}}$$

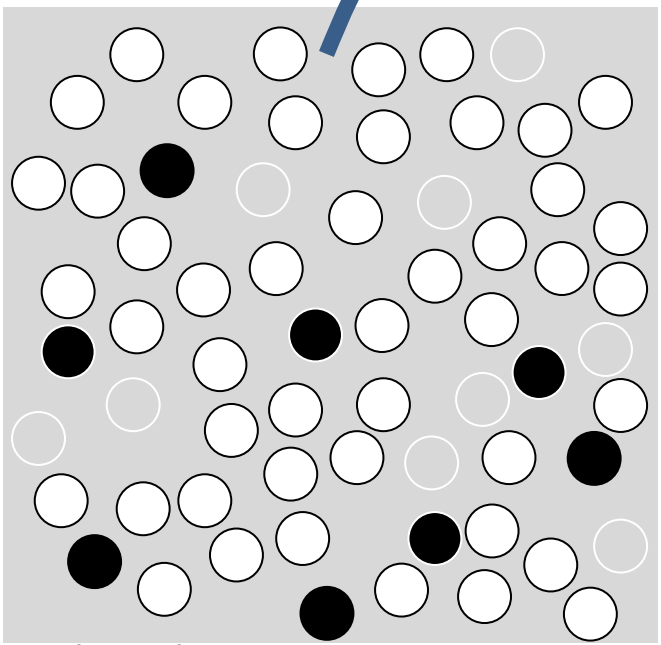
de maneras de elegir las bolas blancas

maneras de elegir *r* elementos de un total de *N*

ORA - Test de Fisher

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

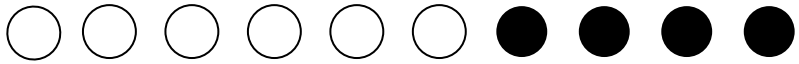
Apoptosis (r=12)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado número de hits en la lista **sólo por azar**



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

	categoria bio	$\overline{\text{categoria bio}}$	
cluster	**m**	k-m	**k**
$\overline{\text{cluster}}$	r-m	N-k+m-r	N-k
	r	N-r	**N**

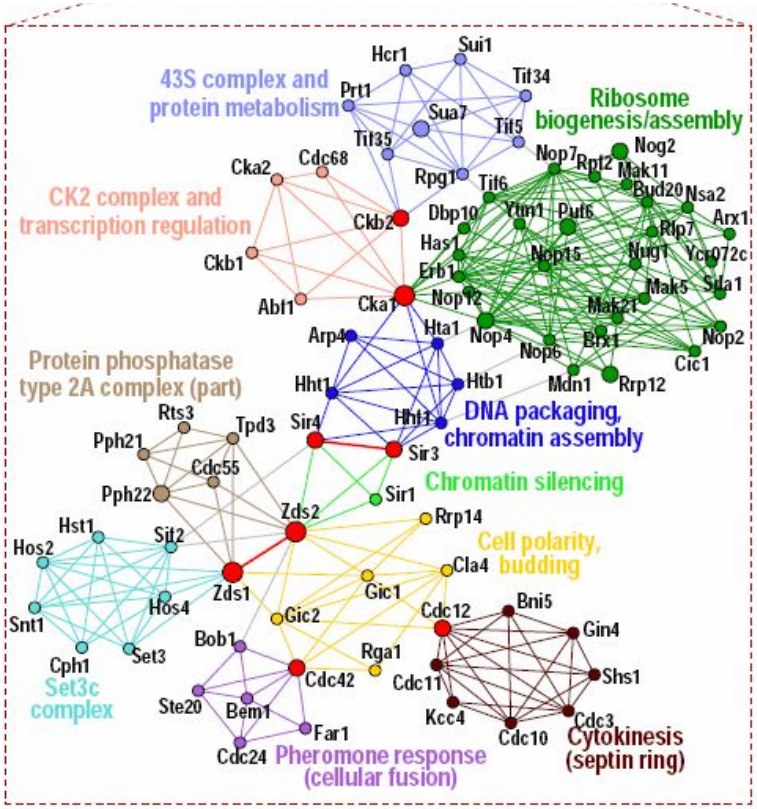
Cual es la probabilidad de ver esta tabla u **otra más extrema**?

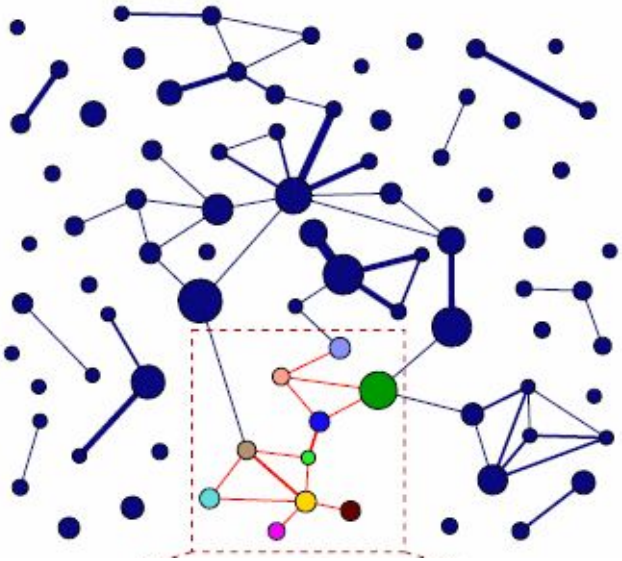
$$p_{valor} = \sum_{m'=m}^{r} \frac{\binom{r}{m}\binom{N-r}{k-m}}{\binom{N}{r}}$$

Es chica esta cantidad?

$$p_v < p_{umbral}$$

Enriquecimiento funcional

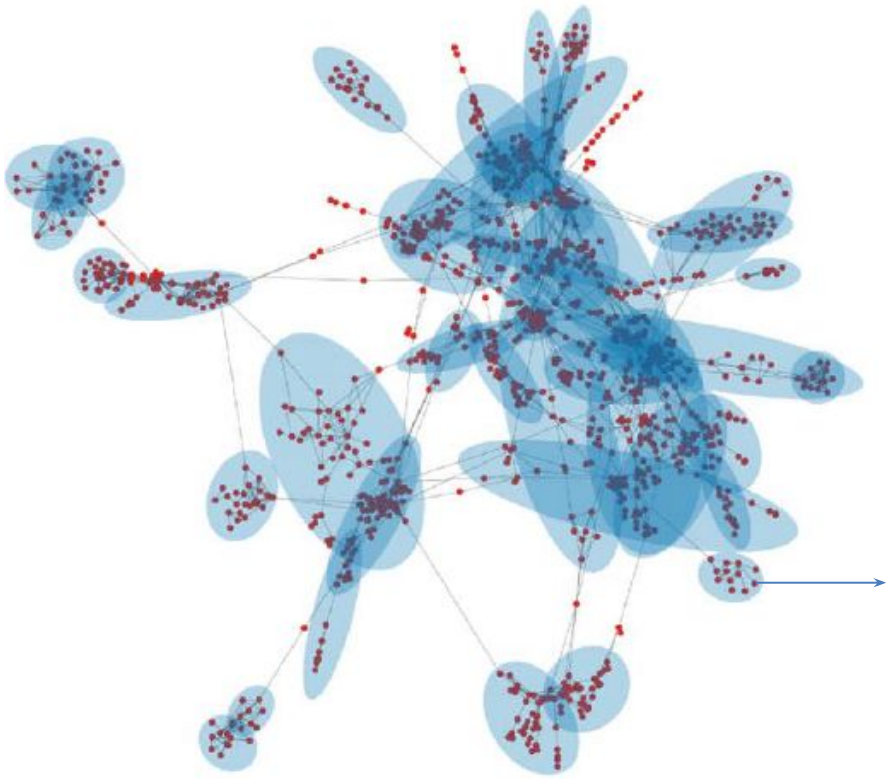


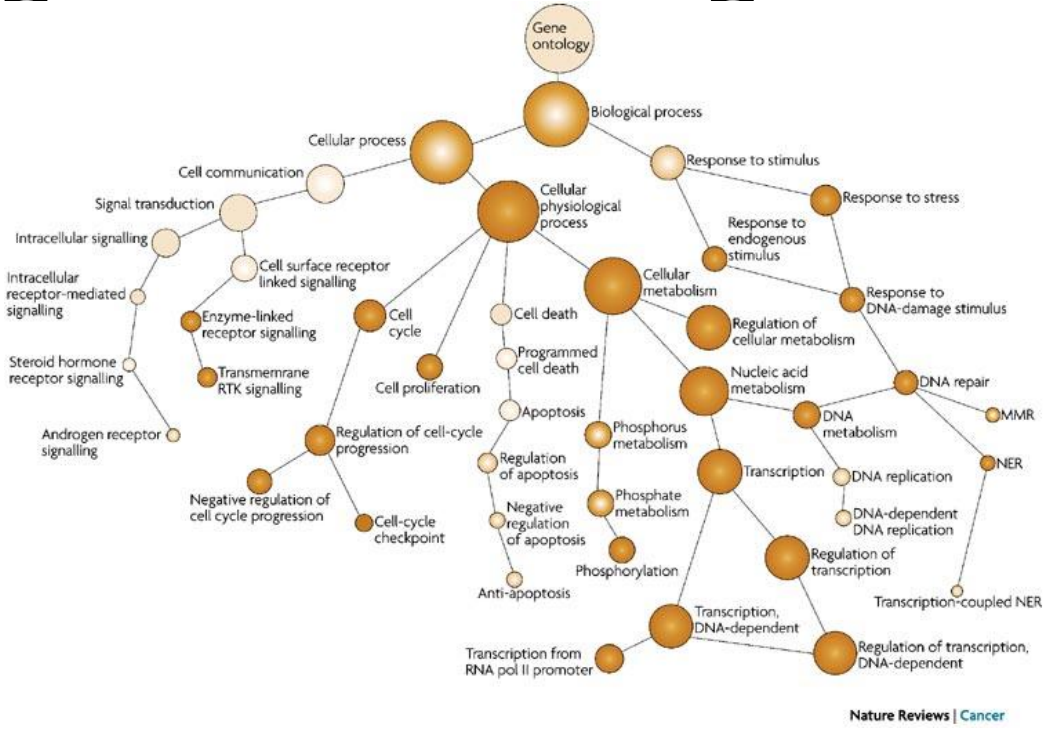


- **MsigDB**
- **Gene Ontology**
- **DAVID**

Indice de Homogeneidad Biológica

Validación de comunidades utilizando **conocimiento externo**



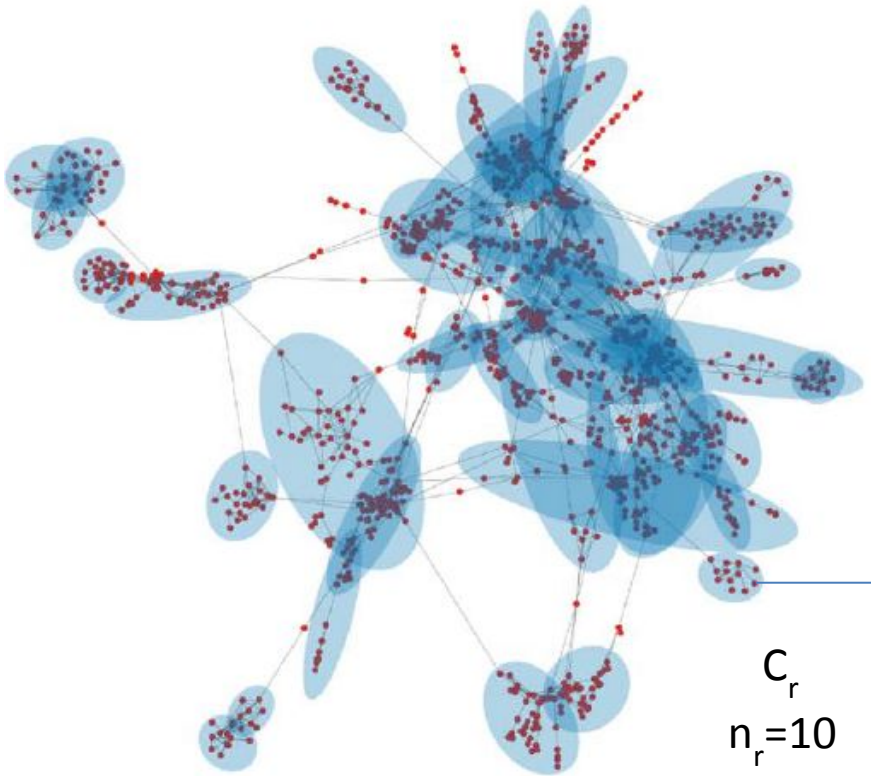


gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

Grupos de genes/proteinas **cercanos** en el espacio de conocimiento capturado por la **red**, qué tan **compactos/homogéneos** son en **otro espacio de conocimiento** relevante?

Indice de Homogeneidad Biológica

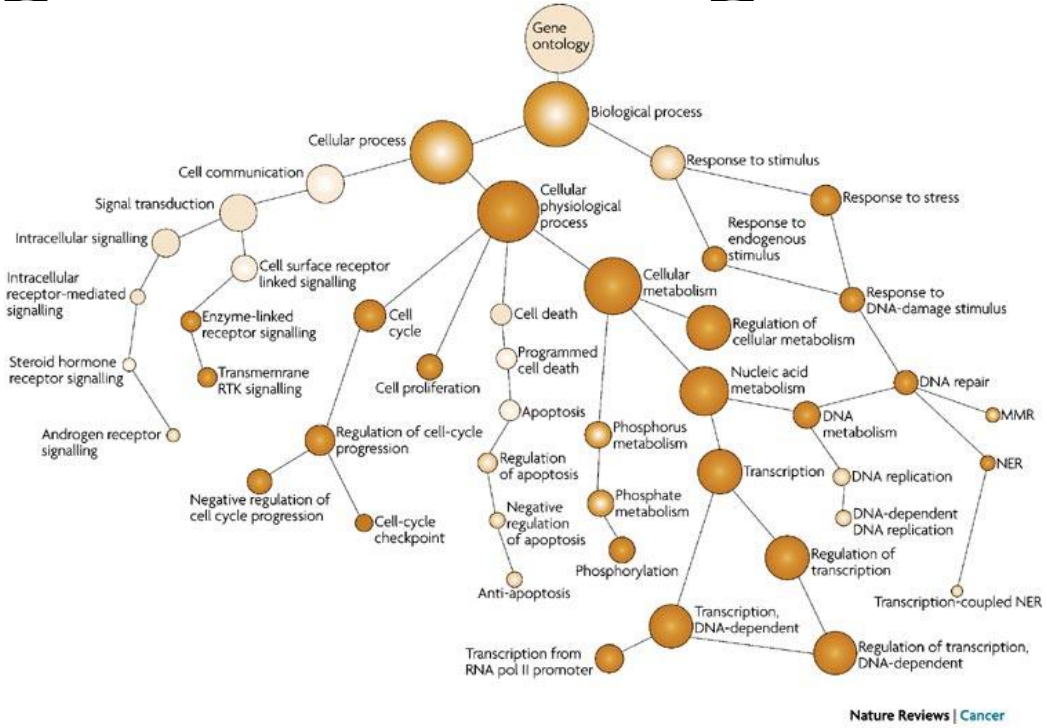
Validación de comunidades utilizando **conocimiento externo**



C_r

$n_r=10$

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



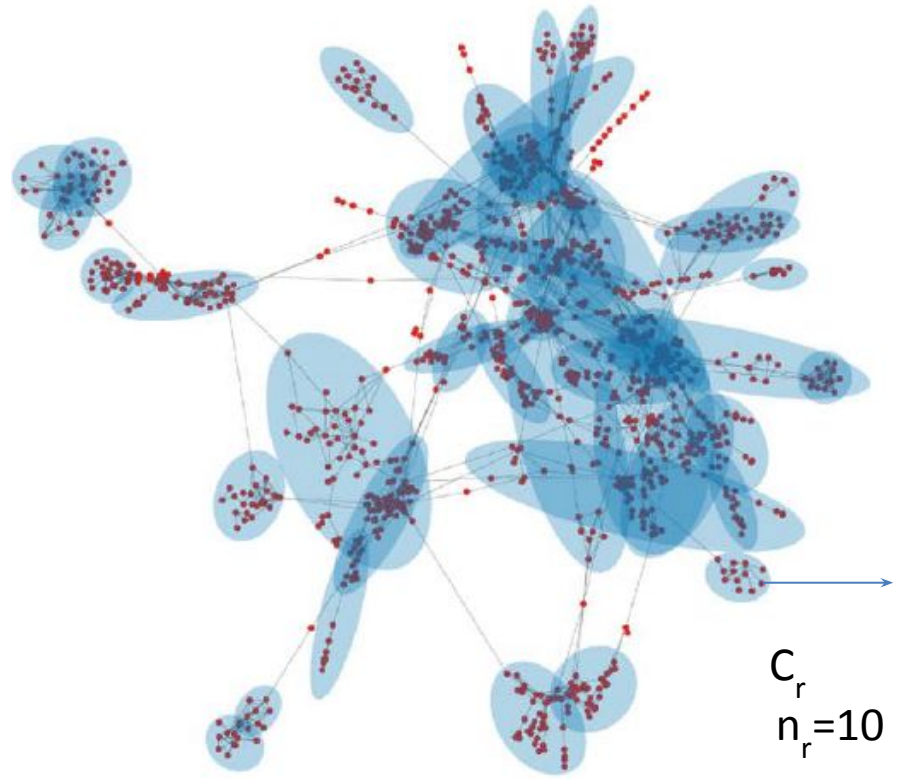
$$BHI_r = \frac{2}{n_r(n_r-1)} \sum_{i \neq j \in C_r} \delta_{GO_i,GO_j}$$

$$BHI = \frac{1}{k} \sum_{r=1}^{k} BHI_r$$

criterio binario de coincidencia en espacio GO

Indice de Homogeneidad Biológica

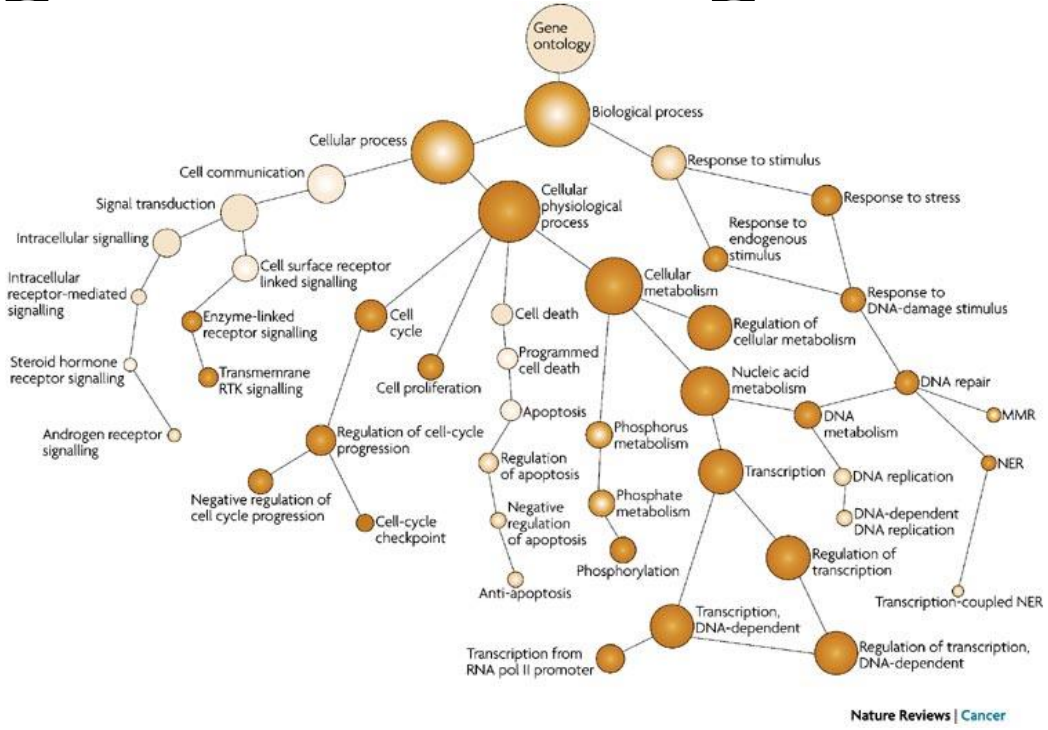
Validación de comunidades utilizando **conocimiento externo**



C_r

$n_r=10$

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



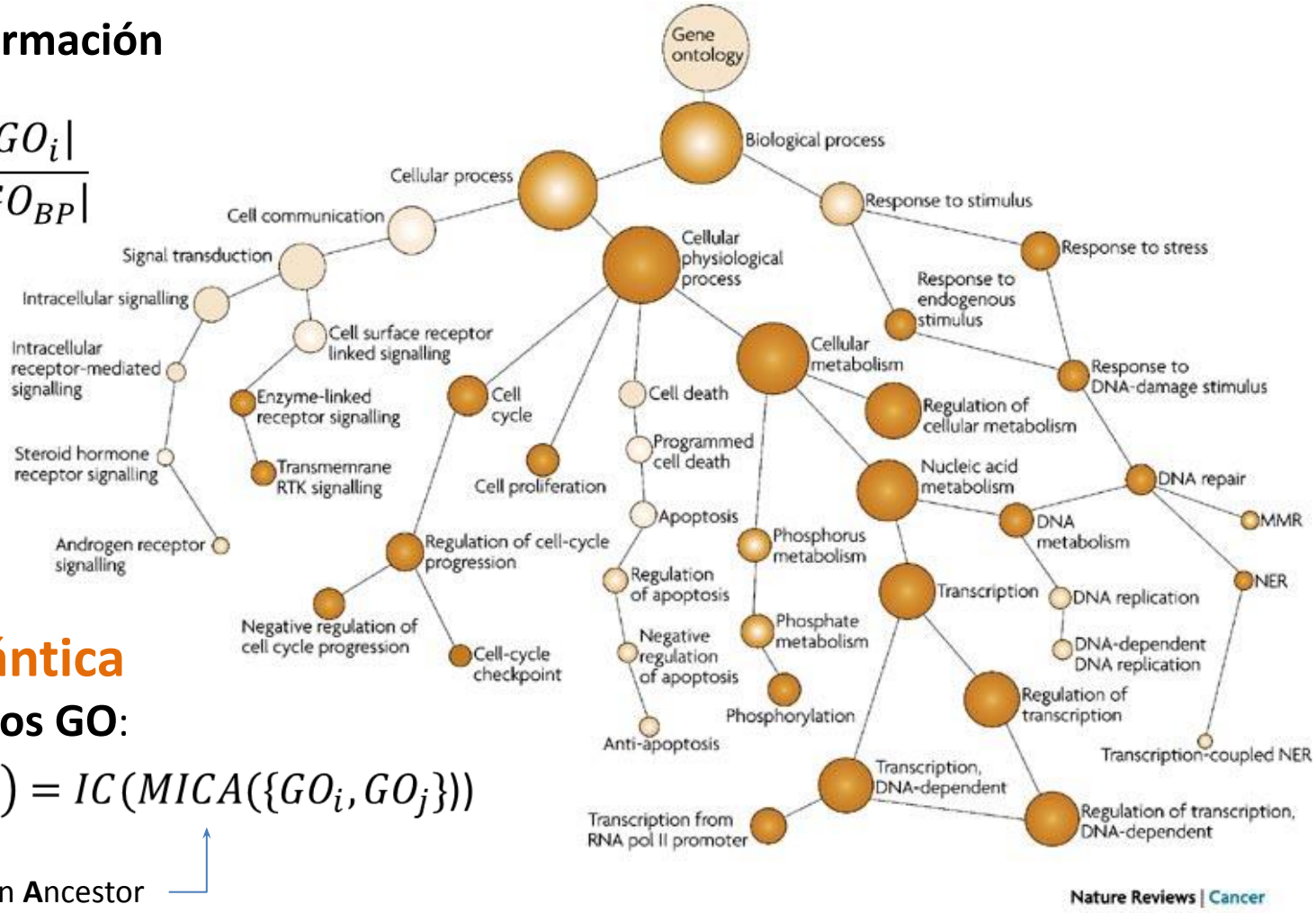
Así como es posible inferir una métrica a partir de la red tambíen es posible inferir una **métrica** a partir de la estructura de **GO** que permita cuantificar el nivel de **similaridad** entre proteinas en un **espacio de conocimiento biológico.**

Hacia una métrica GO

Cuanta información gano si me entero que una proteina está anotada en una clase GO_i?

Contenido de información de un concepto GO:

$$IC(GO_i) = -\log \frac{|GO_i|}{|GO_{BP}|}$$



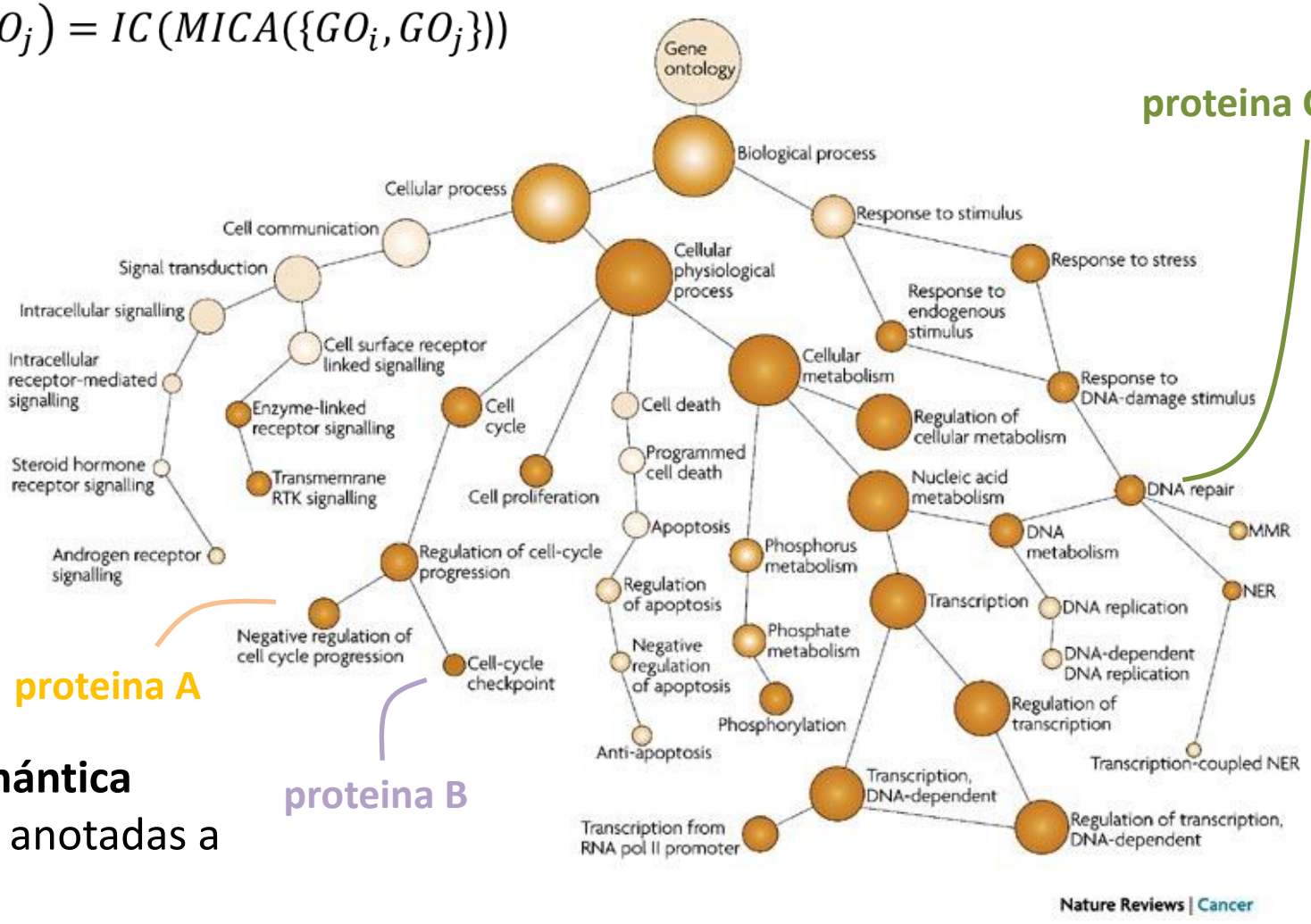
Similaridad semántica entre dos **conceptos GO**:

$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$

Most **I**nformative **C**ommon **A**ncestor

Similaridad semántica

$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$



proteina C

proteina A

proteina B

Similaridad semántica entre **proteinas** anotadas a conceptos GO.

$$sim(prot_A, prot_B) = f(\{GO_i\}_A, \{GO_i\}_B)$$

Pesquita 2009 PLoS Comp Biom, Semantic Similarity

Similaridad semántica

Similaridad semántica entre **conceptos**

$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$

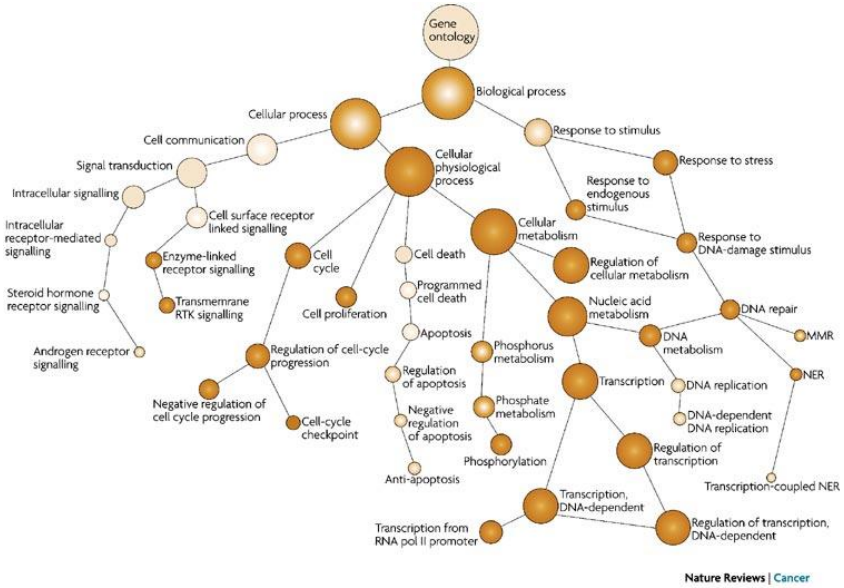
Similaridad semántica entre **proteinas**

$prot_A \rightarrow \{GO_i\}_A \qquad prot_B \rightarrow \{GO_i\}_B$

$GO_3\ GO_5\ GO_{223}\ GO_{542}\ GO_{23}$

GO_1
GO_{12}
GO_{23}

Matriz de sim GO´s

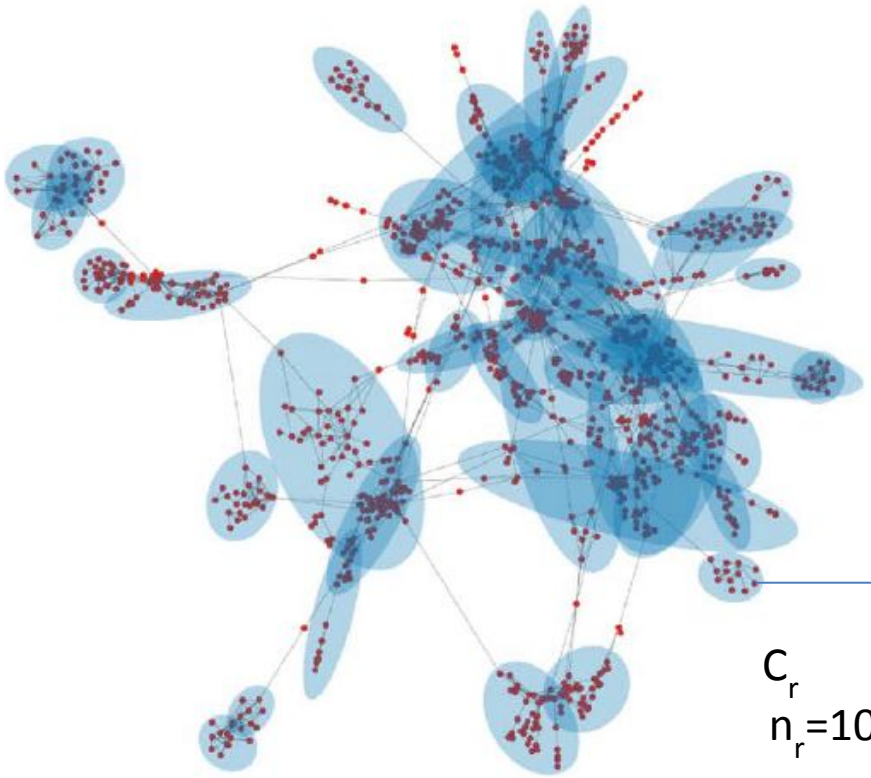


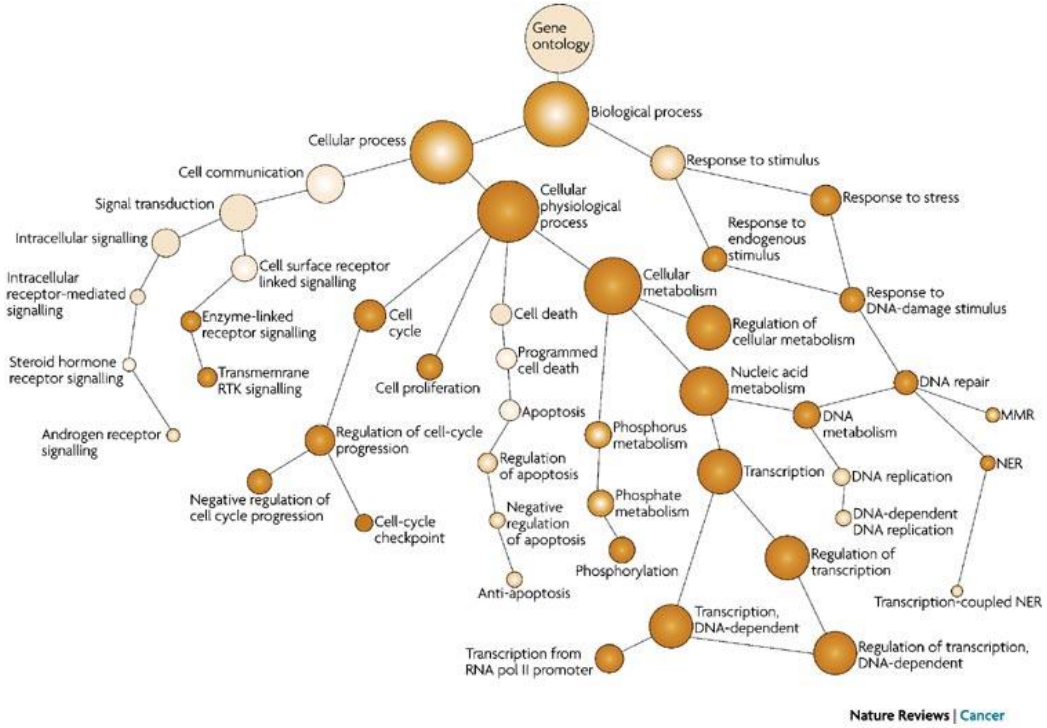
$$sim(prot_A, prot_B) = f(\{GO_i\}_A, \{GO_i\}_B)$$

Pesquita 2009 PLoS Comp Biom, Semantic Similarity

Indice de Homogeneidad Biológica **Reloaded**

Validación de comunidades utilizando **conocimiento externo**





gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

C_r
$n_r=10$

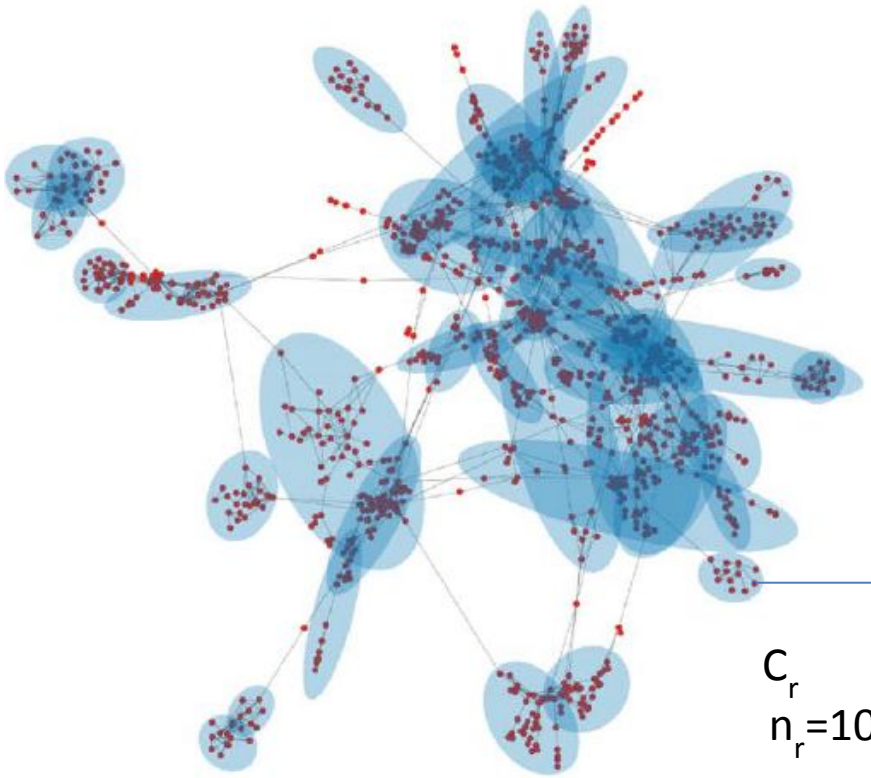
$$BHI_r = \frac{2}{n_r(n_r-1)} \sum_{i \neq j \in C_r} \delta_{GO_i,GO_j}$$

$$BHI = \frac{1}{k} \sum_{r=1}^{k} BHI_r$$

criterio binario de coincidencia en espacio GO

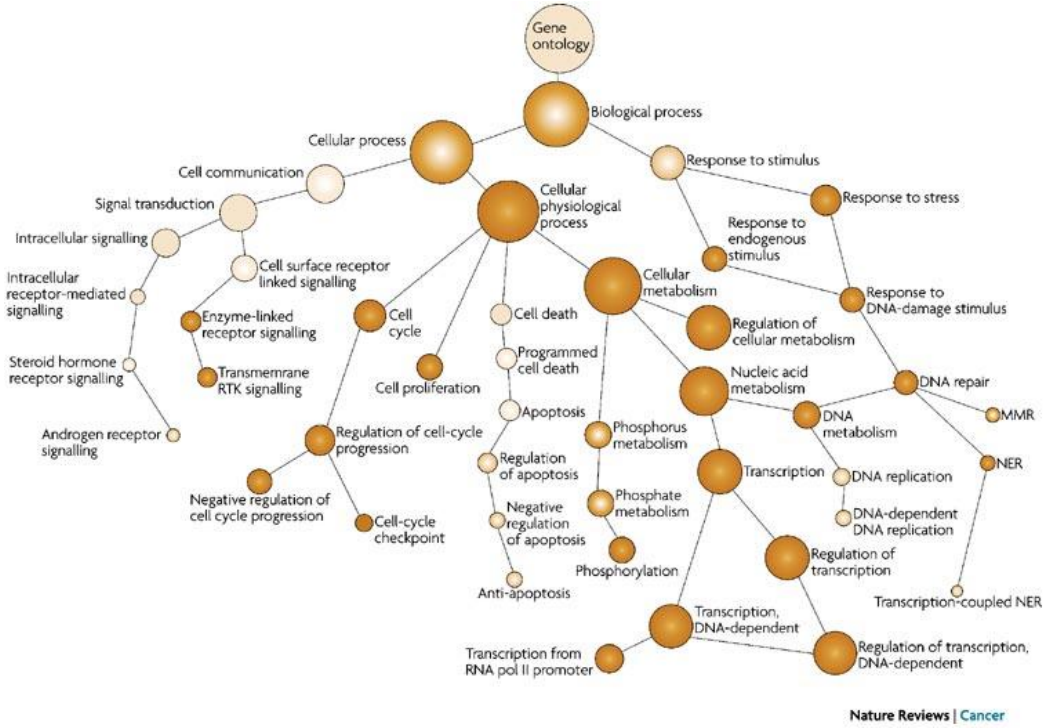
Indice de Homogeneidad Biológica **Reloaded**

Validación de comunidades utilizando **conocimiento externo**



C_r
n_r=10

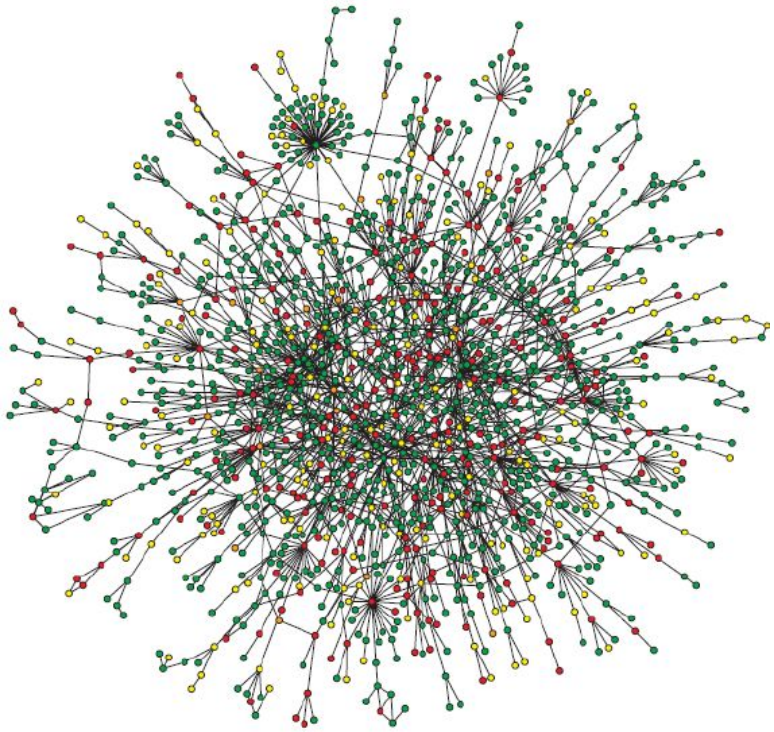
gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



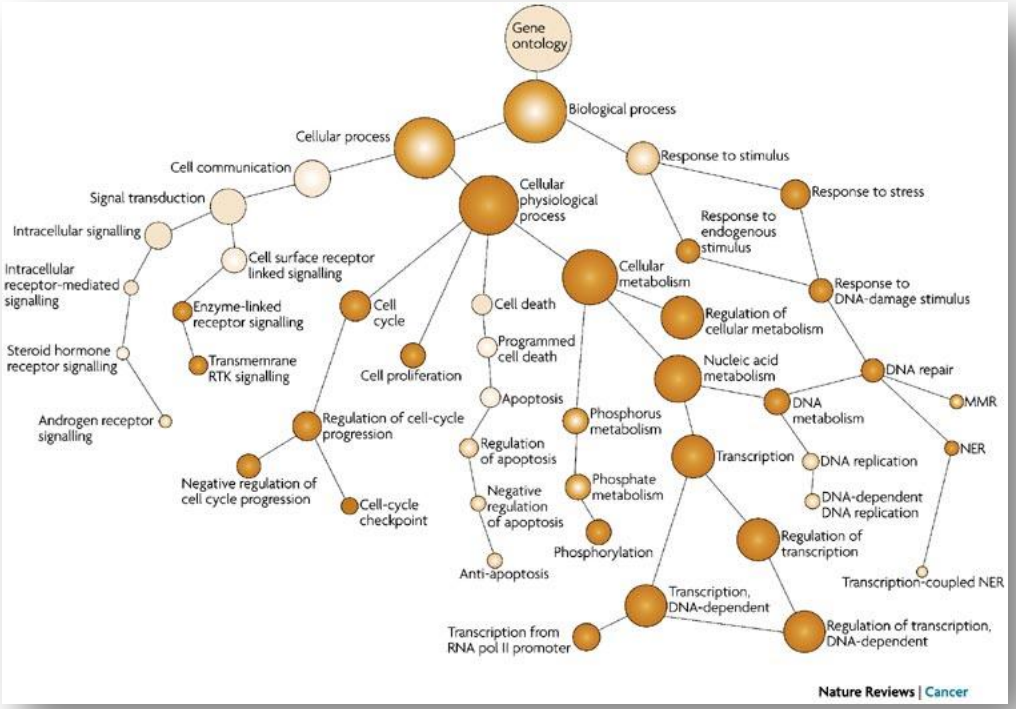
$$BHI_r = \frac{2}{n_r(n_r-1)} \sum_{i \neq j \in C_r} sim_{GO}(i,j)$$

$$BHI = \frac{1}{k} \sum_{r=1}^{k} BHI_r$$

Integrando métricas



Gene Ontology

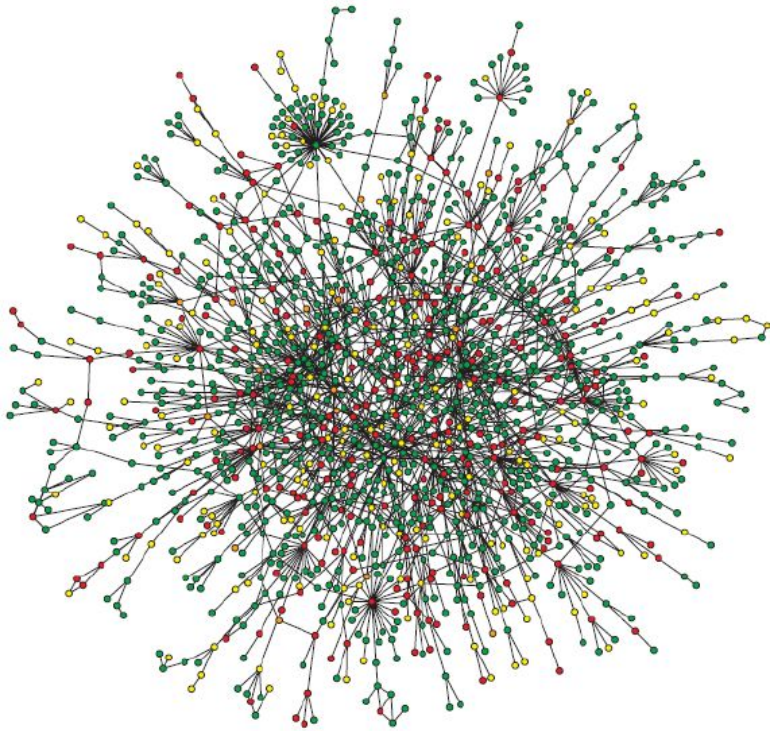


d_{red} — $d_{GO} = d_{semantic}$

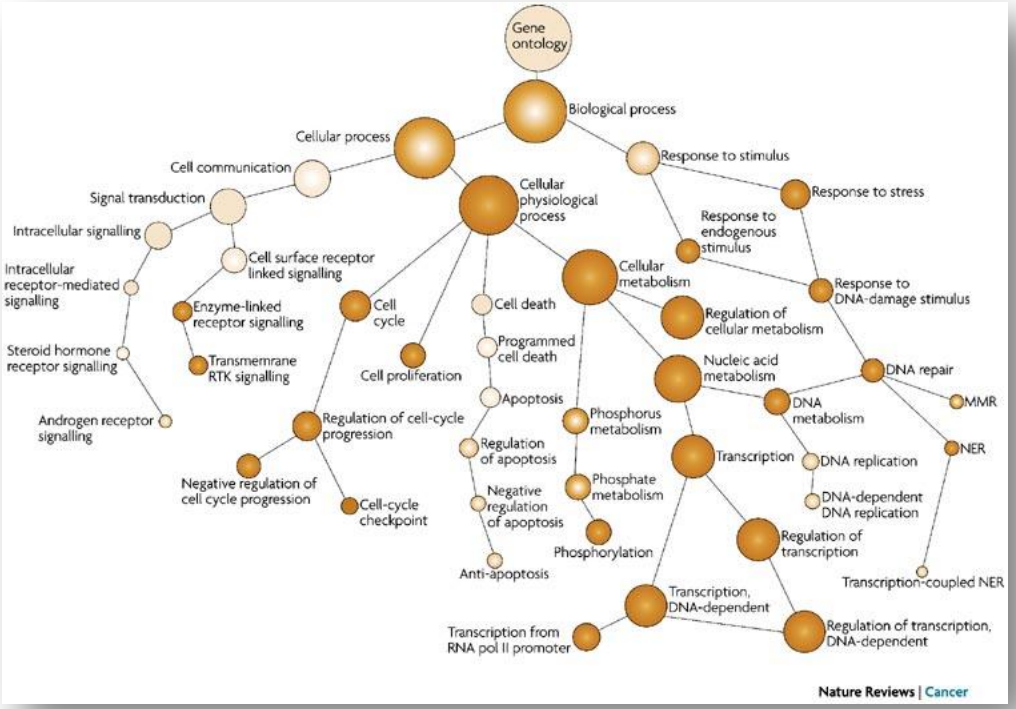
Mixed Metrics

Buscamos estructuras **similares** en cuanto a la información embebida en la topología de la red **Y** el espacio de conocimiento biológico

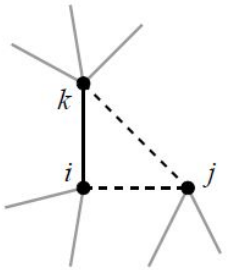
Integrando métricas



Gene Ontology



Por ejemplo…recordemos…



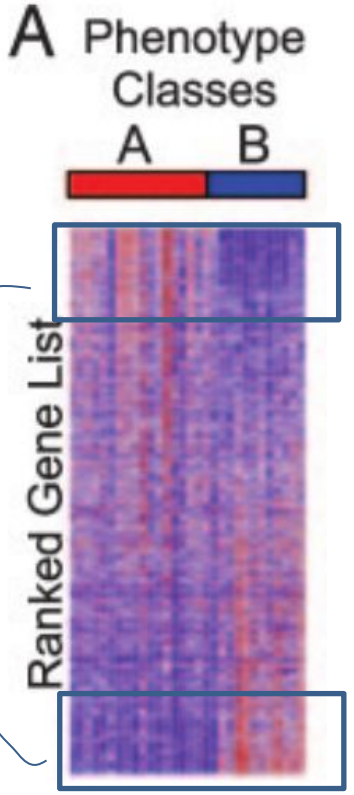
Equivalencia estructural **extendida**:
Dos vértices, *i* y *j* son similares si *i* tiene como vecino a *k*, quien a su vez es similar a *j*.

$$\sigma_{ij} = \alpha \sum_{k} A_{ik}\sigma_{kj} + s^{ext}{}_{ij}$$

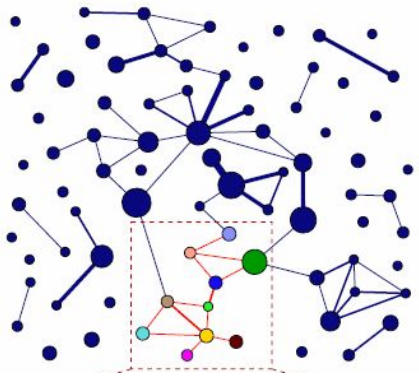
SAMBA + CLICK

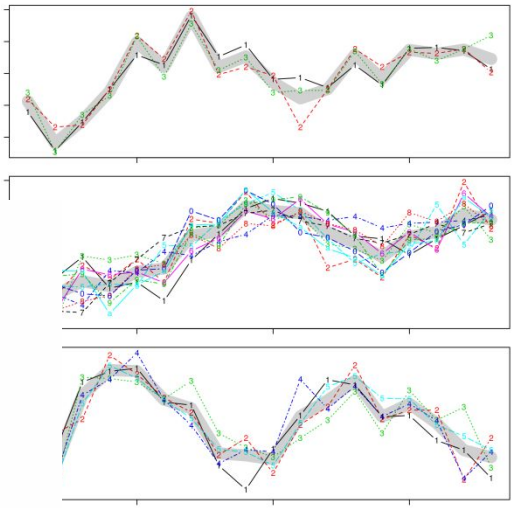
Analisis de sobre-representacion

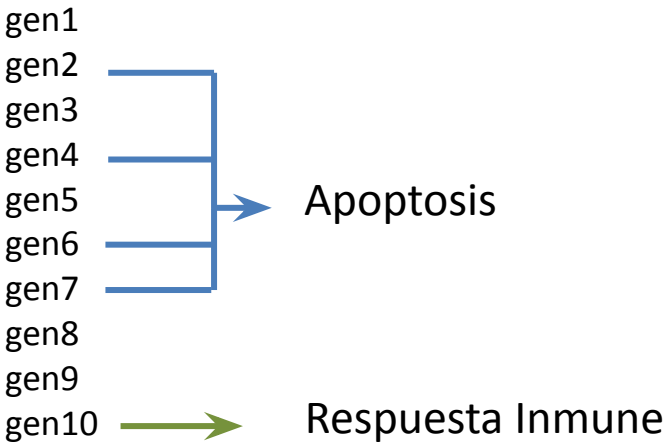
Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerepresentado en los **genes diferencialmente expresados** que encuentro en mi experimento?



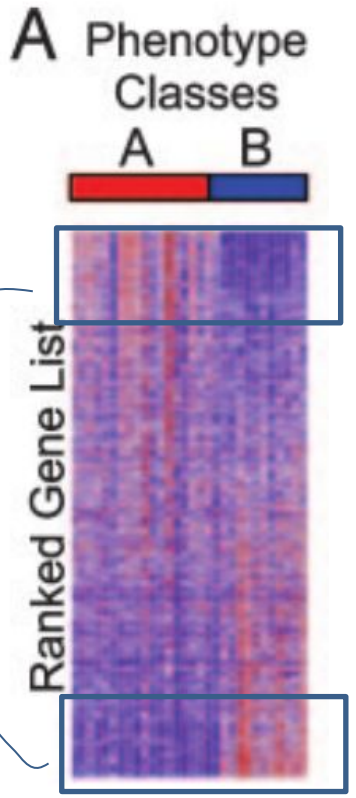
Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerepresentado en alguna comunidad o cluster detectado en mis datos?





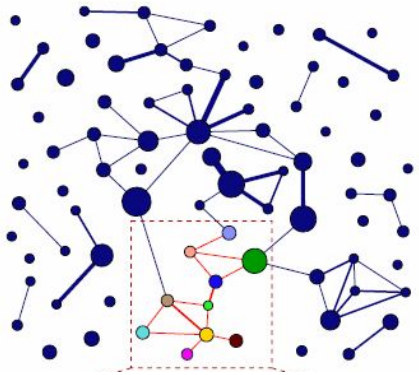


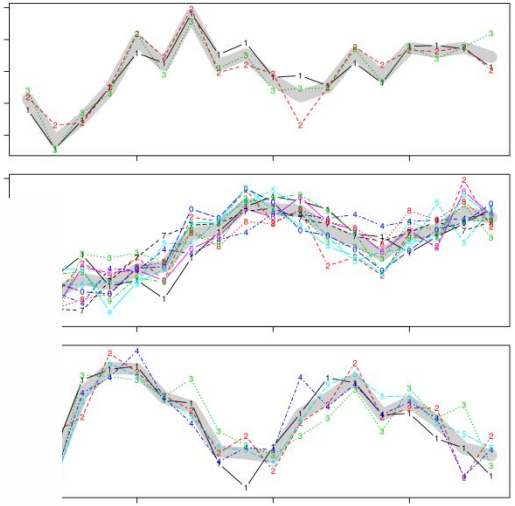
Analisis de sobre-representacion



Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerepresentado en los **genes diferencialmente expresados** que encuentro en mi experimento?

Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerepresentado en alguna comunidad o cluster detectado en mis datos?





Existe un **vinculo** entre pertenecer a una **cjto de genes identificado en mi experimento** y estar anotado dentro de un **tema biologico**?

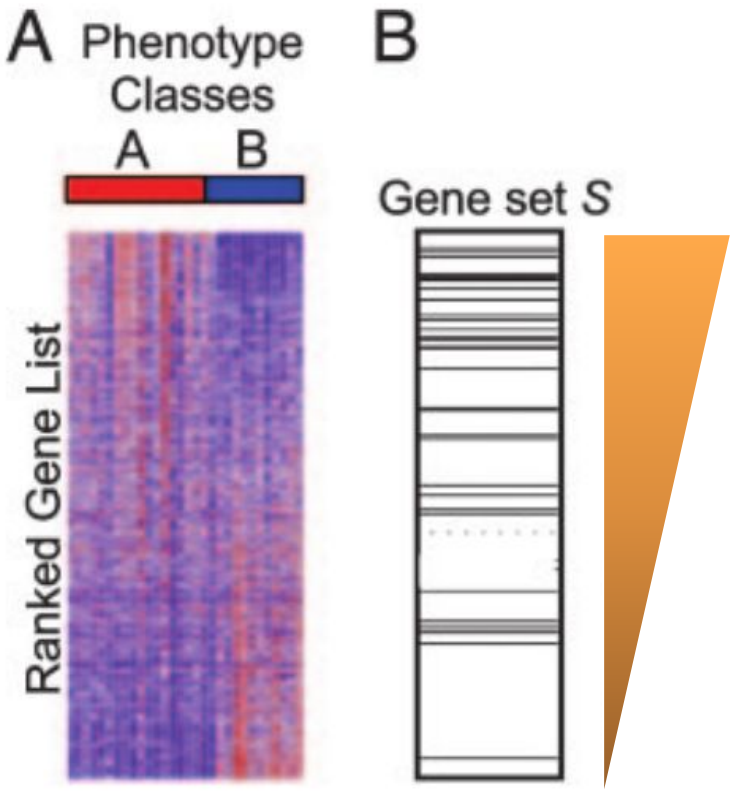
Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar

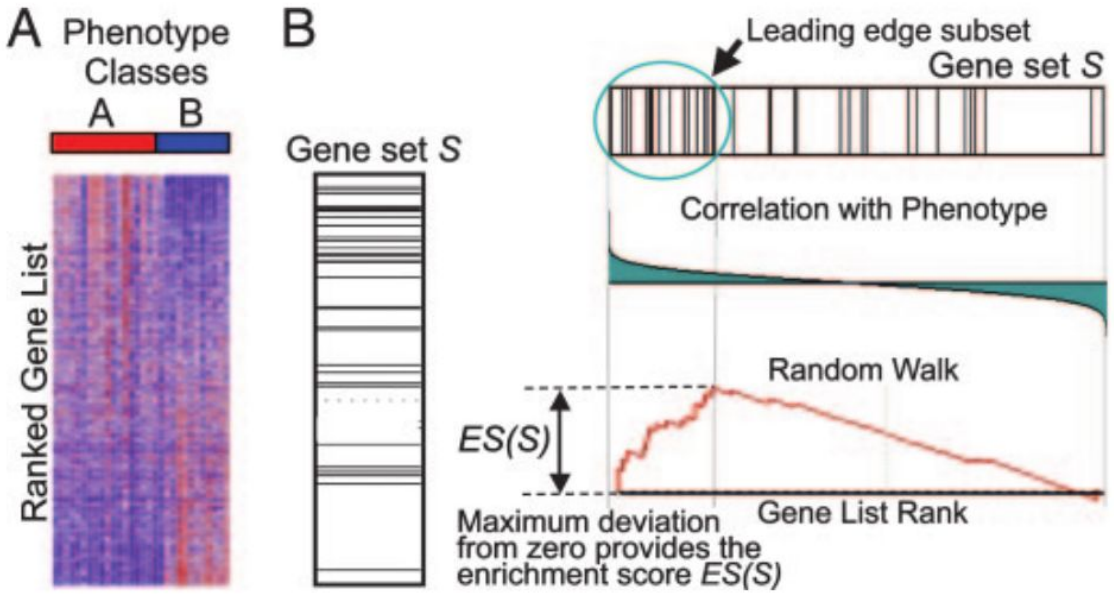
Gene **S**et **E**nrichment **A**nalysis

- Mootha et al., Nature Genetics 34, 267–273 (2003)
- Subramanian et al., PNAS 102(43), 15545–15550 (2005).

- Estimé la expresion de un conjunto de genes y los rankeo segun la correlacion que presenten con algun fenotipo de interes
- Con estos datos, me pregunto si el fenotipo estudiado se relaciona con una dada categoria biologica de interes (CBI) previamente asociada un dado cjto de genes (por ej una categoria GO)
- Si hay una asociacion positiva espero que los genes de la CBI *esten rankeados arriba* en el ranking…como cuantifico esto?



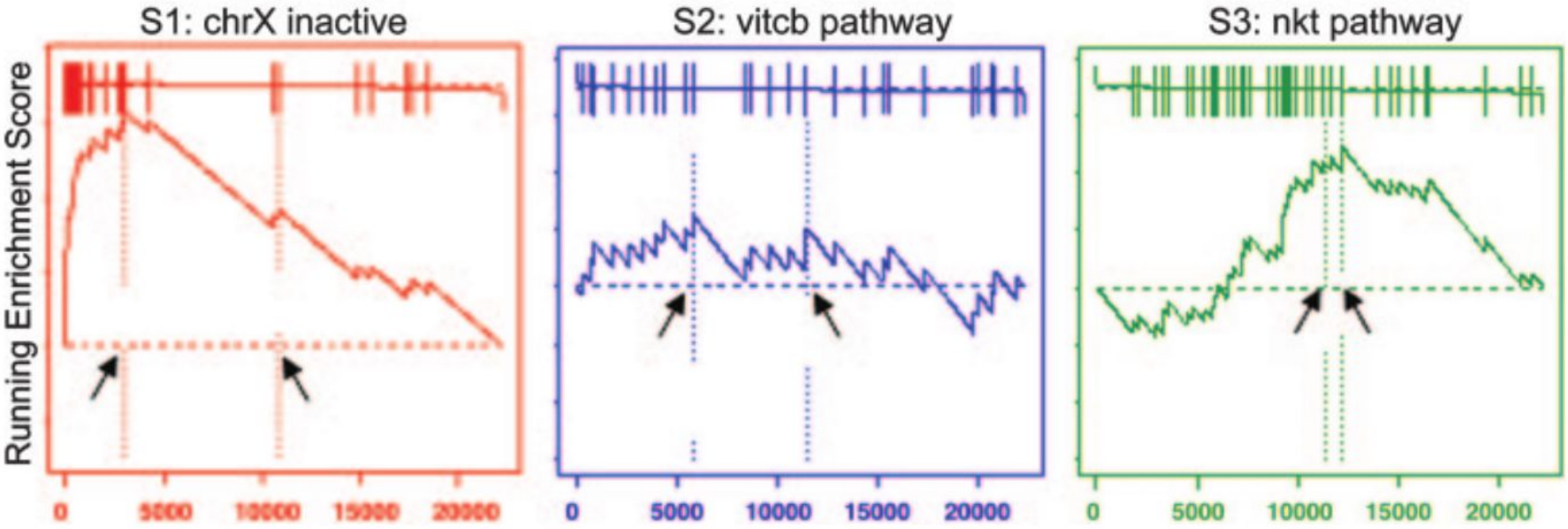
GSEA



- Computo suma acumulativa sobre el ranking: Sumo un poroto cuando encuentro un gen del CBI, sino resto uno (la magnitud del incremento depende de la correlacion del gen con el fenotipo)

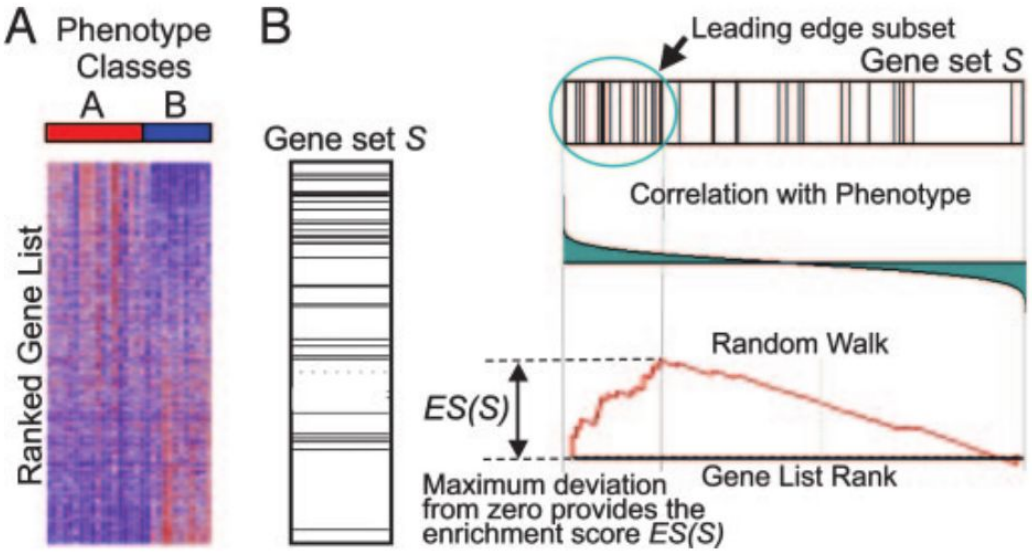
- La maxima desviacion respecto a cero es el Enrichment Score (ES) asociado al CBI

GSEA



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

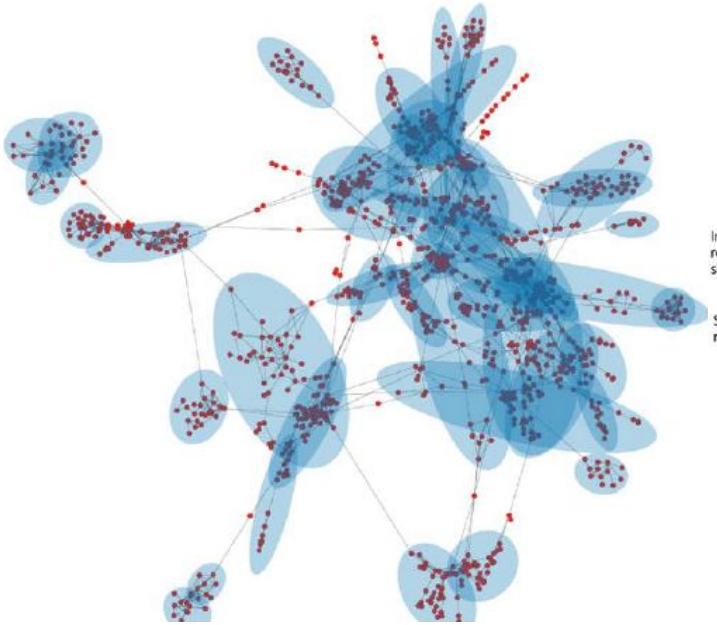
GSEA

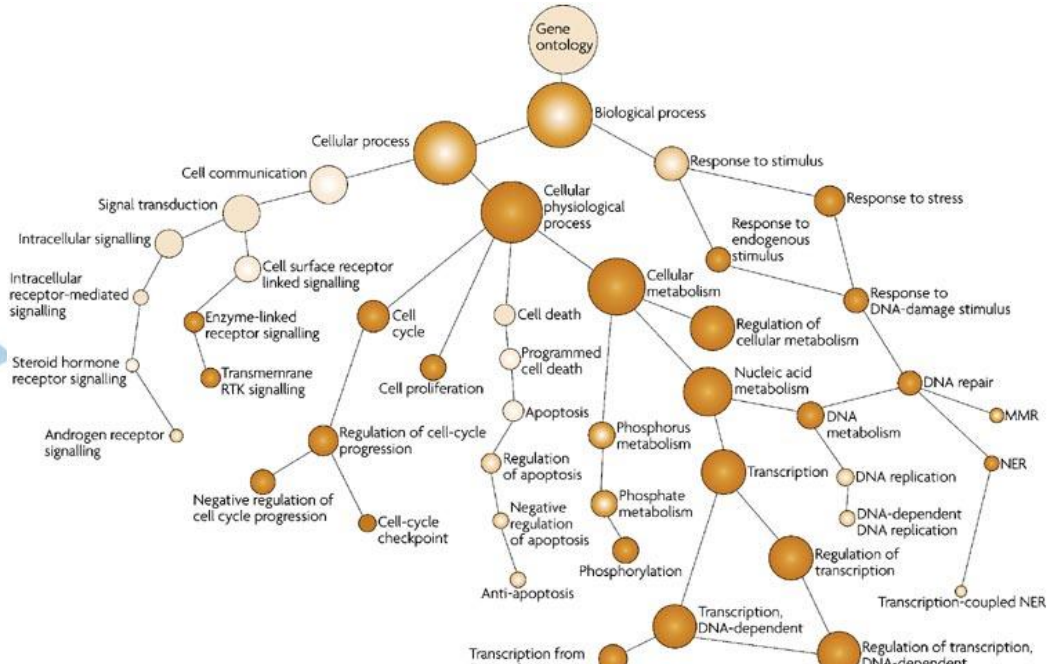


- Para asignar significancia: se permutan las etiquetas de fenotipos 1000 veces y se computa el ES para cada permutacion.

- Se reporta adicionalmente un Score Normalizado (NES) por la media observada en el dataset random, para alivianar la dependencia con el tamanio del CBI

El problema de los grandes datos: testeo de múltiples hipótesis





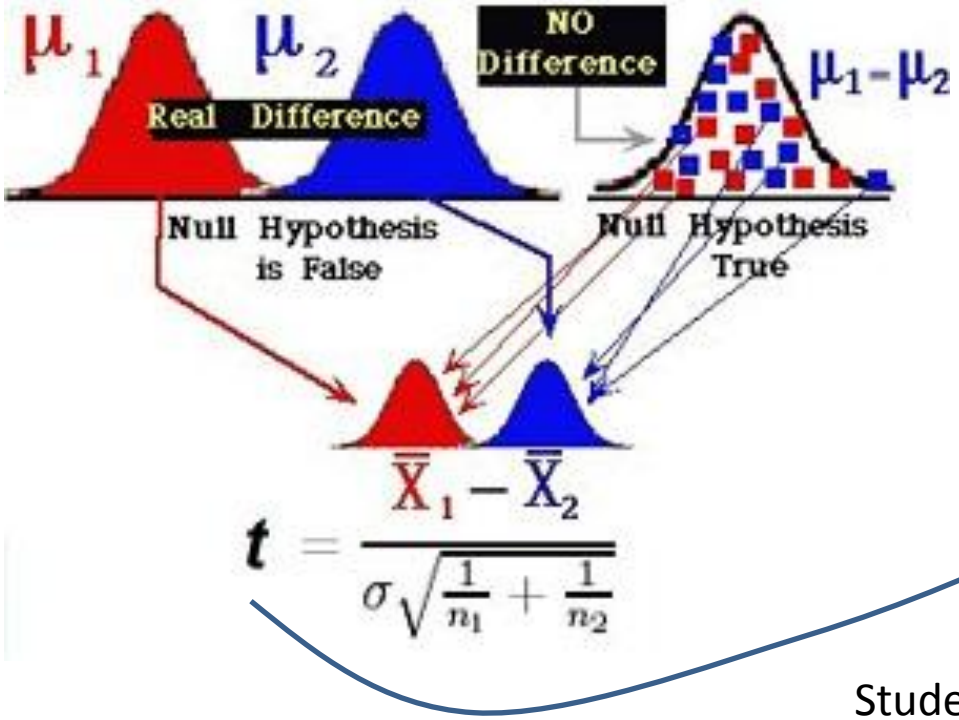
- Para una dada comunidad/cluster voy a realizar un **test sobre TODO posible concepto** GO para buscar sobrerepresentación…..son muchos tests (!)

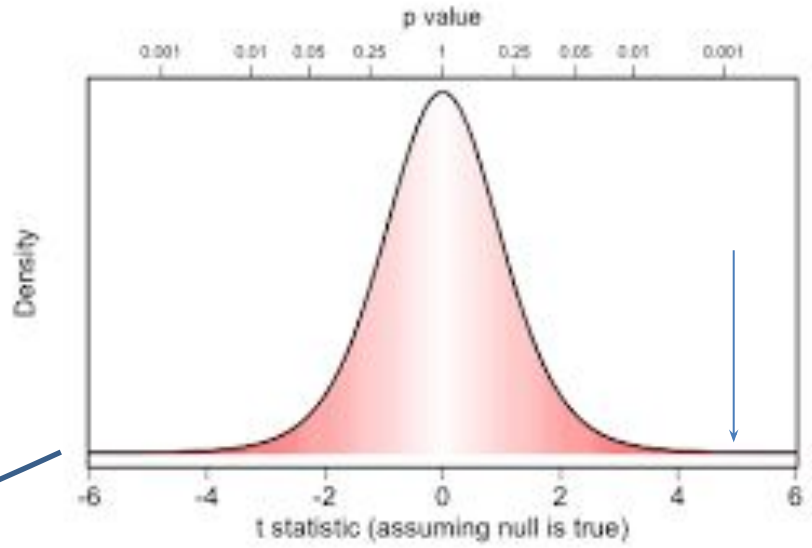
- Existe la posibilidad de **ver eventos extremos** en el **conjunto de todos los test individuales** realizados por azar, aún cuando no haya vínculo entre cluster y anotación

- Eventos con p_v=0.05 tienen un 5% de prob de ocurrir por azar Entonces espero ~5 Falsos Positivos si hago 100 tests y uso 0.05 como corte de significancia de test s individuales.

Ejemplo del problema de testeo múltiple de hipótesis

Quiero saber si dos condiciones son diferentes….Que test estadistico puedo usar?

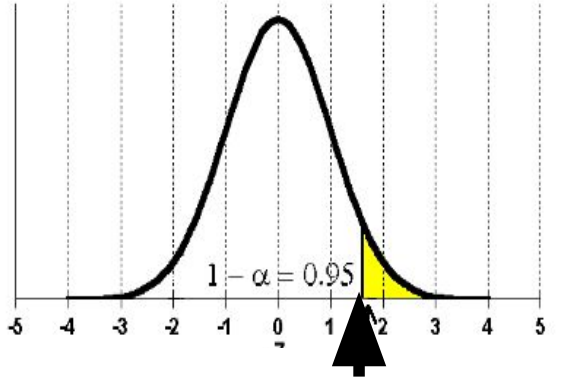




Student distribution

Ejemplo del problema de testeo múltiple de hipótesis

Distribution for “t” under Null Hypothesis



Observed t

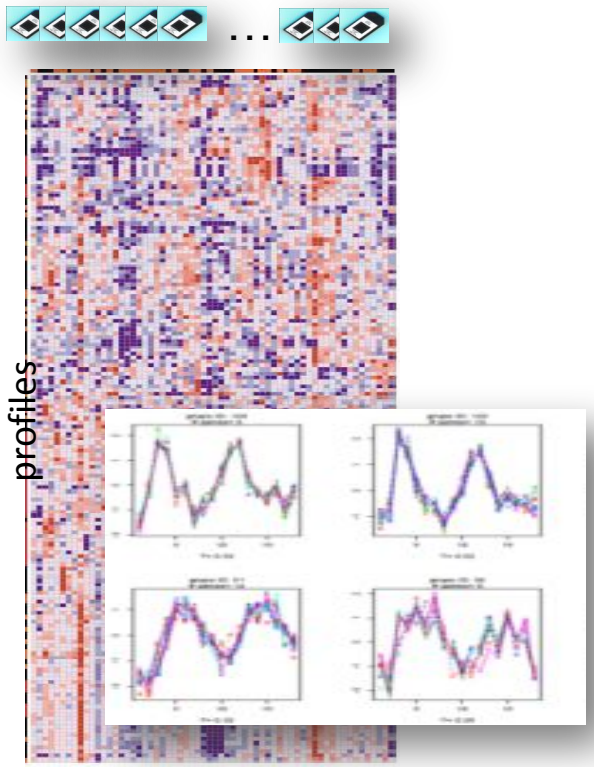
Random generated “expression matrix”

	A	A	A	A	B	B	B	B	Ttest p-value
1	0,15	0,51	0,08	0,84	0,17	0,91	0,69	0,37	0,58
2	0,47	0,06	0,28	0,12	0,83	0,95	0,42	0,29	0,08
3	0,07	0,94	0,17	0,59	0,66	0,1	0,21	0,2	0,55
4	0,17	0,49	0,88	0,05	0,33	0,11	0,7	0,7	0,8
5	0,86	0,4	0,6	0,26	0,57	0,48	0,7	0,28	0,87
6	0,22	0,04	0,41	0,7	0,29	0,4	0,5	0,44	0,68
7	0,05	0,45	0,06	0	0,96	0,3	0,53	0,45	0,06
8	0,49	0,45	0,55	0,25	0,59	0,83	0,28	0,89	0,22
9	0,83	0,51	0,26	0,64	0,83	0,46	0,63	0,79	0,46
10	0,03	0,33	0,6	0,86	0,09	0,78	0,02	0,96	0,99
11	0,72	0,81	0,54	0,14	0,64	1	0,22	0,82	0,61
12	0,44	0,16	0,21	0,58	0,6	0,89	0,85	0,64	0,02
13	0,5	0,56	0,11	0,75	0,55	0,91	0,7	0,88	0,13
14	0,48	0,17	0,07	0,3	0,83	0,16	0,11	0,87	0,33
15	0,03	0,84	0,35	0,26	0,45	0,71	0,03	0,96	0,55
16	0,38	0,11	0,02	0,58	0,15	0,93	0,23	0,51	0,43
17	0,39	0,56	0,12	0,48	0,02	0,93	0,95	0,89	0,25
18	0,58	0,57	0,22	0,4	0,03	0,76	0,81	0,57	0,63
19	0,82	0,68	0,04	0,63	0,91	0,83	0,88	0,75	0,14
20	0,84	0,77	0,64	0,89	0,72	0,23	0,11	0,91	0,19

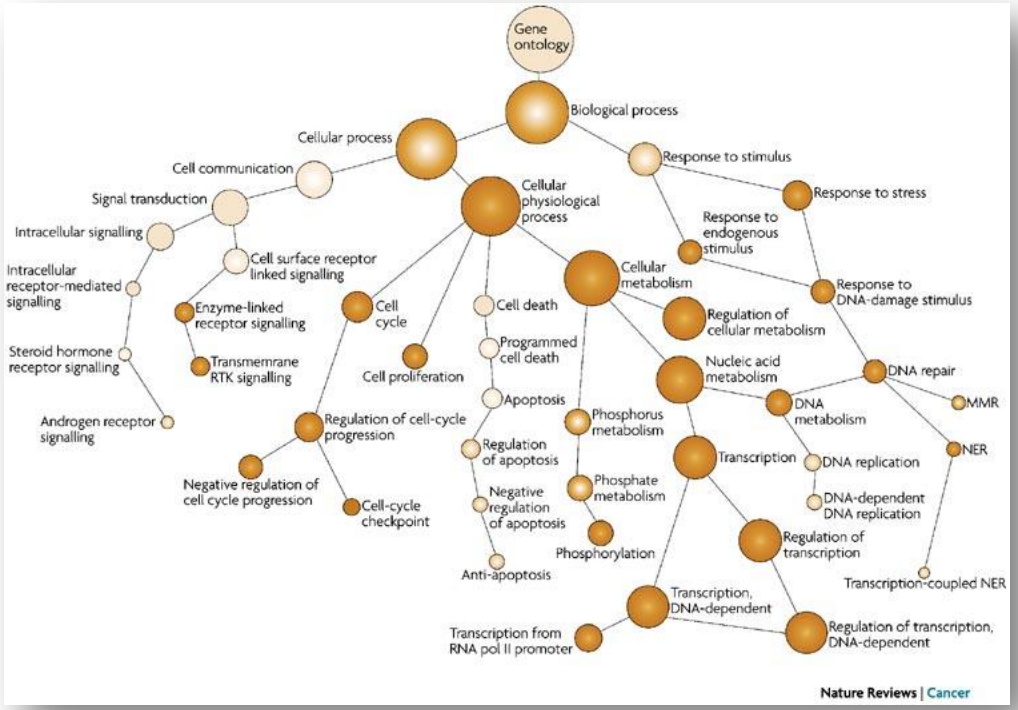
False Positive at p_cutoff=0.05 !!!

Need of **Multiple Hypothesis Testing Correction** when testing ~10000 genes

Integrando métricas



Gene Ontology



$d_X = d_{coexpression}$ — $d_{GO} = d_{semantic}$

Mixed Metrics

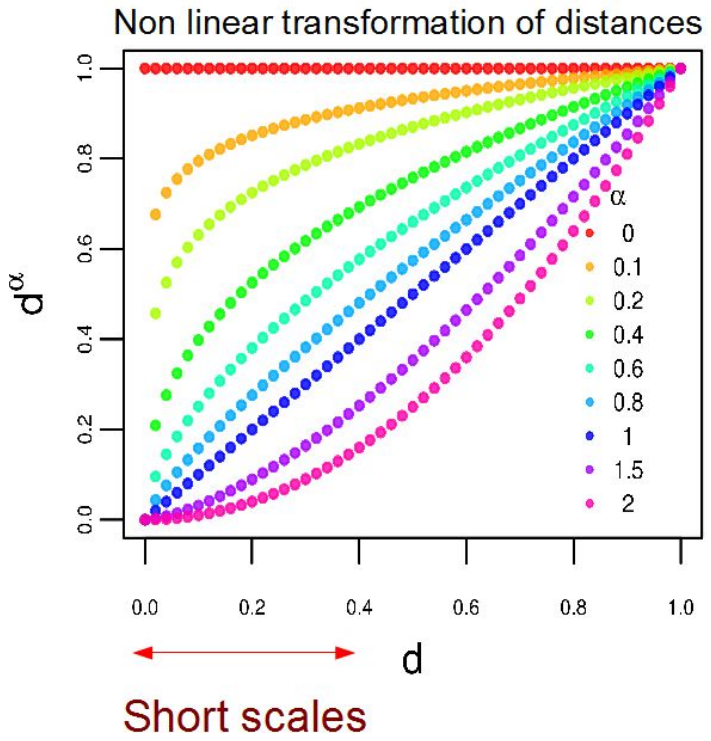
Buscamos estructuras **similares** en cuanto a la información transcripcional **Y** el espacio de conocimiento biológico

Agrupando genes con métricas mixtas

El uso de métricas mixtas en algoritmos de reconocimiento de clusters / comunas permite buscar estructura simultáneamente coherente en ambos espacios (por ejemplo: espacio transcripcional y GO)

Las métricas pueden combinarse de diferente forma

* $d^A_{mix} = \sqrt{(1-\alpha)d^2_x + \alpha d^2_{go}}$
* $d^B_{mix} = d_x . d^\alpha_{go}$
* $d^C_{mix} = d^{1-\alpha}_x . d^\alpha_{go}$



- El parámetro α controla la medida en la que el espacio GO es tenido en cuenta en la estimación de similaridades entre objetos.

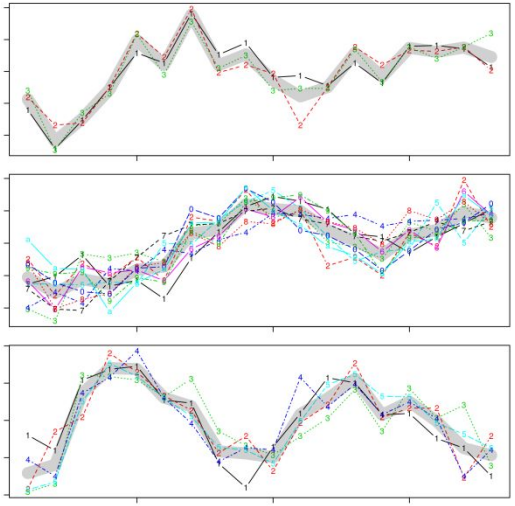
- d^B ***acerca*** genes **biologicamente relacionados** en el sentido de GO

La transformación no-lineal altera de manera diferente la resolución a diferentes escalas

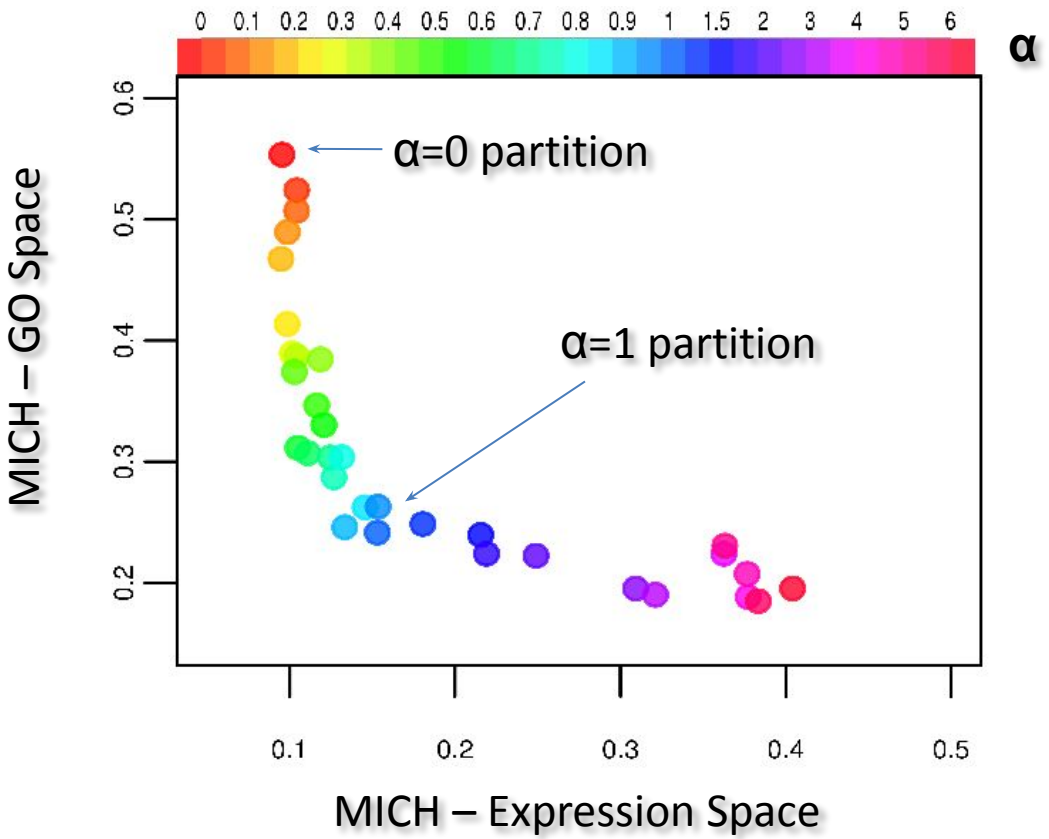
Agrupando genes con métricas mixtas

$$d_{mix} = d_x . d_{go}^{\alpha}$$

Cluster structure



Mean Intra Cluster Heterogeneity



$$MICH(Y) = \frac{1}{N_{cl}} \sum_{cl_k}^{N_{cl}} \frac{2}{n_k(n_k-1)} \sum_{i<j \in cl_k}^{n_k} d_{ij}^{Y}$$

Agrupando genes con métricas mixtas

$$d_{mix} = d_x . d_{go}^{\alpha}$$

Cluster structure