Análisis de Vías Biológicas

Luis Valenzuela Villa,
luis.valenz.v@gmail.com

Laboratorio de Genética, Facultad de Ciencias,
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,
Universidad de Chile.
4 de Octubre, 2017

Análisis de Vías Biológicas
Outline

1. Vías Biológicas.

2. Bases de datos.

3. Métodos estadísticos para el análisis de vías biológicas.

4. Práctico en R: Clusterprofiler, String.DB

Análisis de Vías Biológicas

1. Vías Biológicas.

Una serie de acciones entre moléculas que llevan a cierto producto o cambio en la célula.

Entonces, una vía puede gatillar:

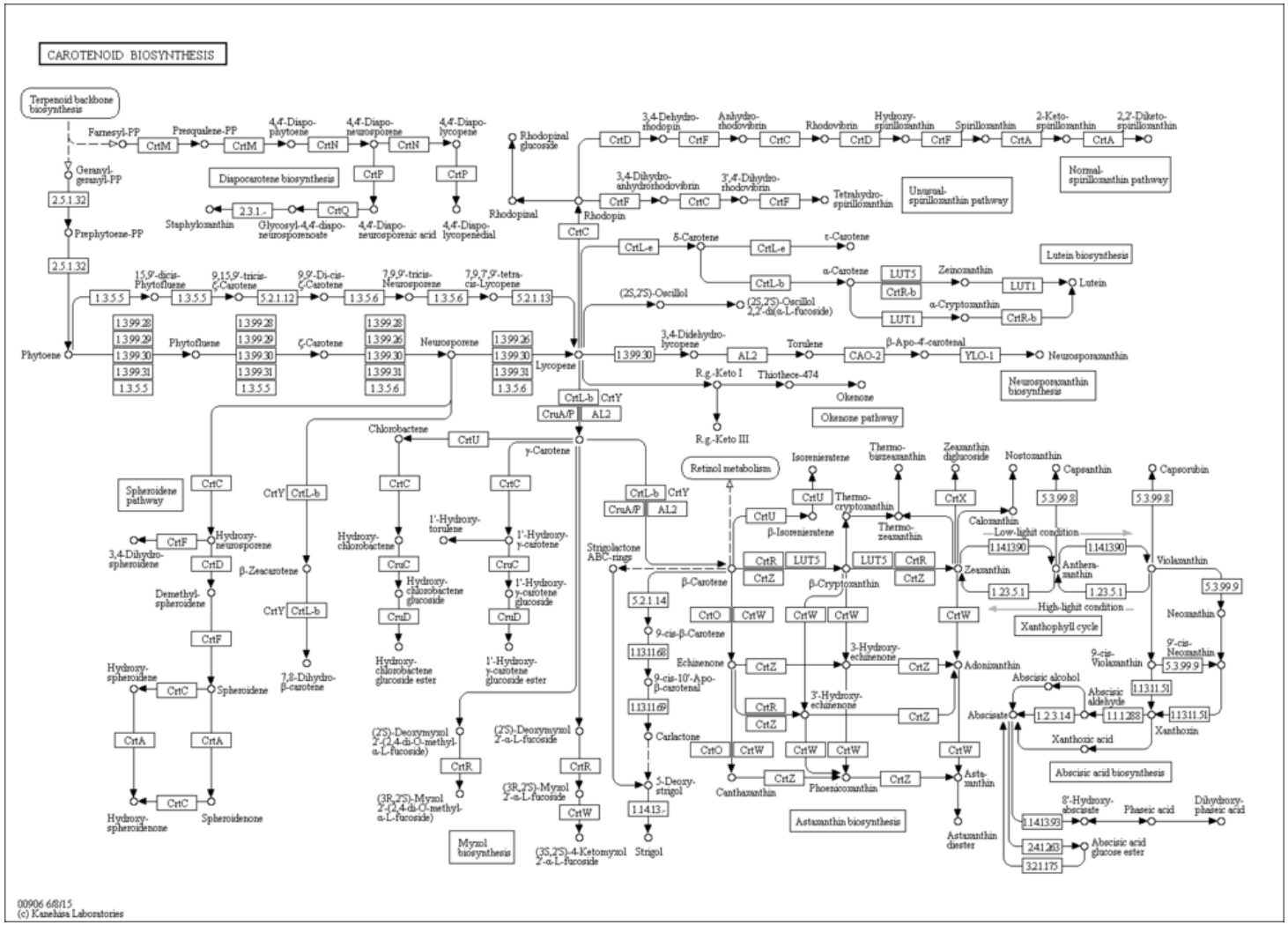
- La formación de moléculas, como proteínas o ácidos grasos.
- Promover o reprimir la transcripción de genes.

La mayoría de las vías biológicas están involucradas en:

- Metabolismo.
- Regulación de la expresión génica.
- Transducción de señales.

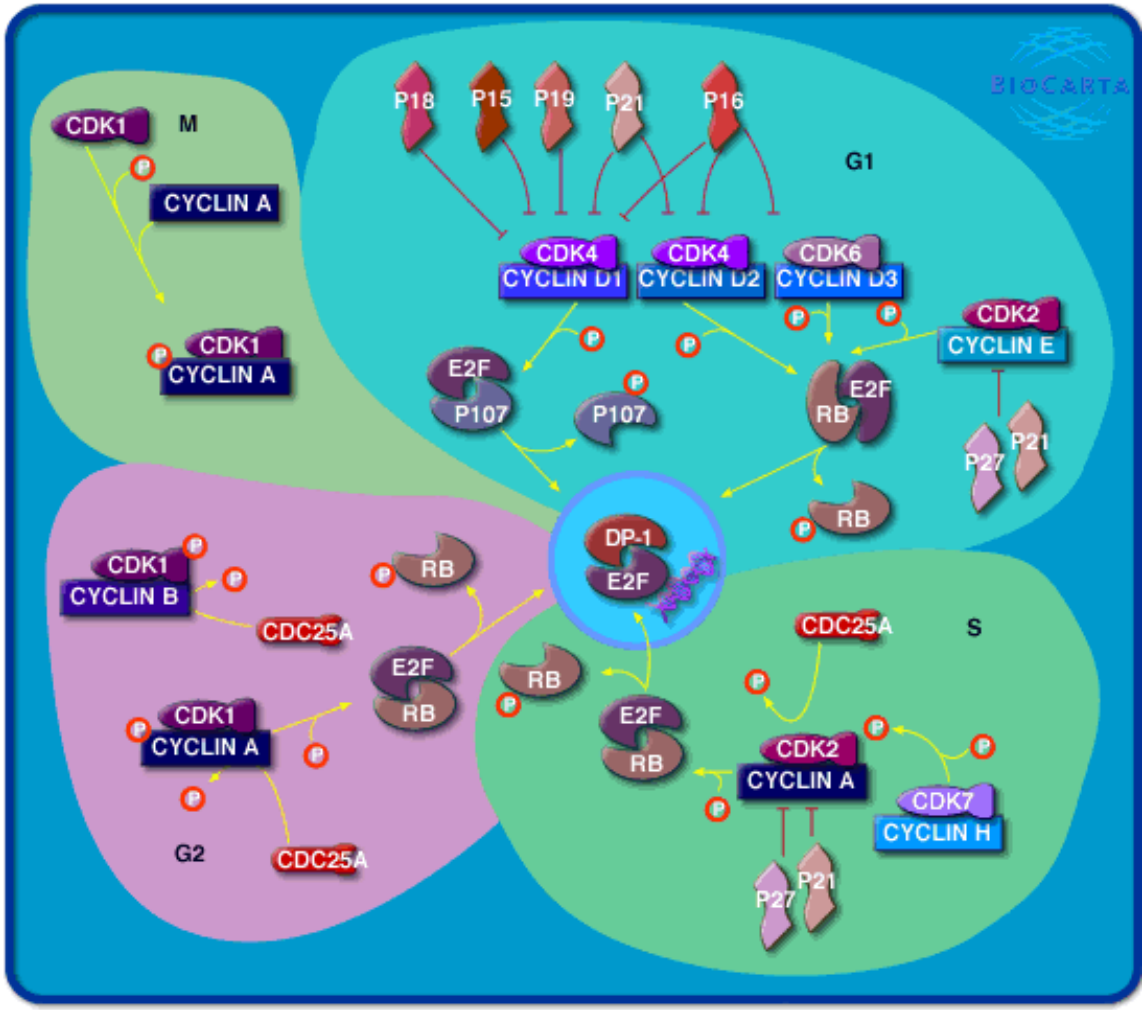
Análisis de Vías Biológicas

Vías Biológicas: Metabolismo



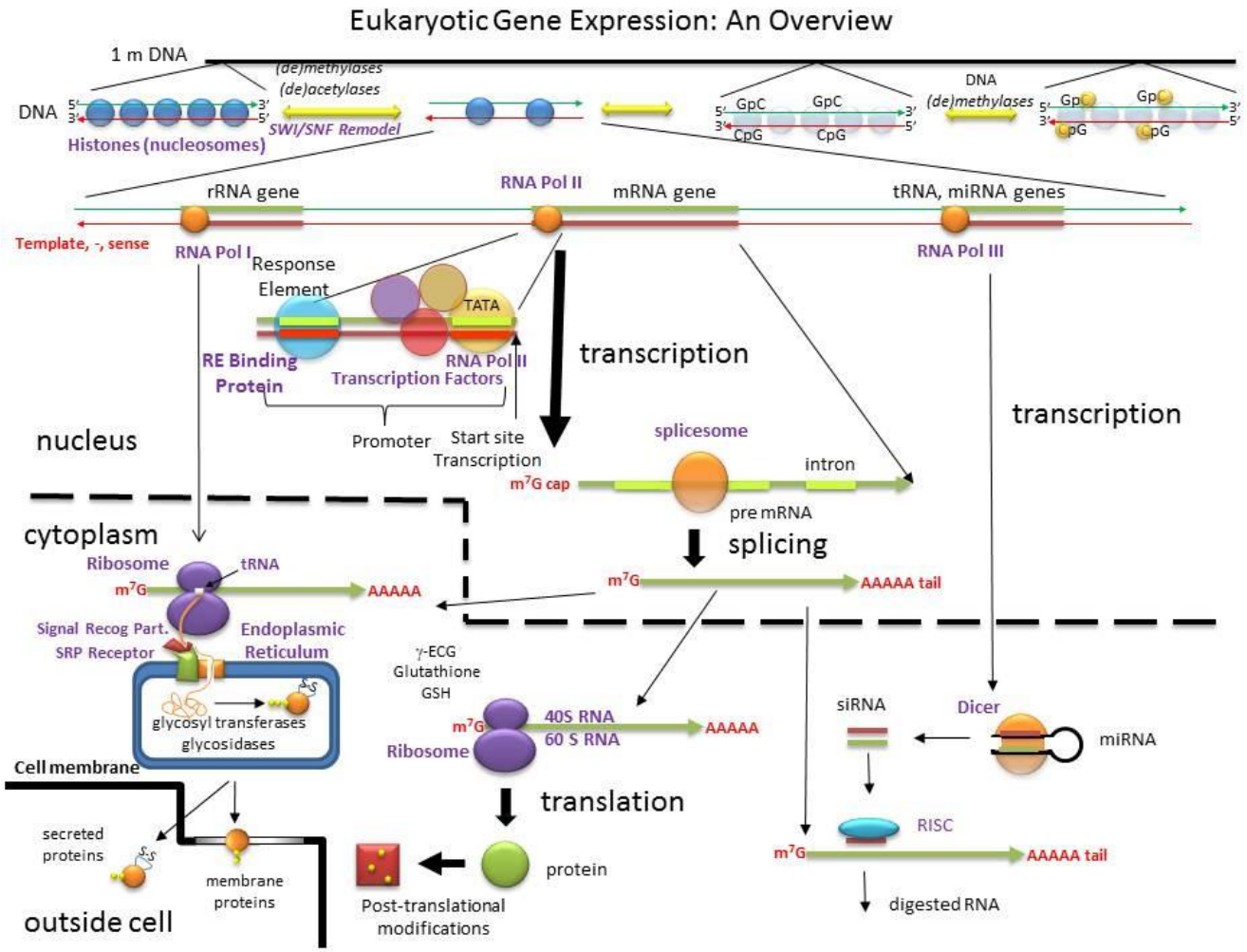
Análisis de Vías Biológicas

Vías Biológicas: Transducción de señales



Análisis de Vías Biológicas

Vías Biológicas:



Análisis de Vías Biológicas
Vías Biológicas.

Concepto amplio que puede ser utilizado en:

- Análisis de conjuntos de genes (ej. ORA, FCS, ToA)
- Redes de interacción física (ej. Interacciones proteicas)

- Simulación cinéticas de vías (ej. ODE análisis)
- Análisis de vías en estado estacionario (ej. FBA)

Análisis de Vías Biológicas
2. Bases de datos: GO

Gene Ontology Consortium | Home | Documentation | Downloads | Community | Tools | About | Contact us

Enrichment analysis

Your gene IDs here...

biological process

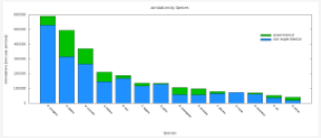
Homo sapiens

Submit

Help

Powered by PANTHER

Statistics



Other GOC tools

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Search

Ontology

Filter classes

Download ontology

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

molecular function
molecular activities of gene products
cellular component
where gene products are active
biological process
pathways and larger processes made up of the activities of multiple gene products.

Annotations

Download annotations (standard files)

Filter and download (customizable files <100k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. more

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. more

Search documentation

Search

User stories

Explore documentation related to your personal user story.

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology *Important*
- Enrichment analysis
- Downloads

http://geneontology.org/

Análisis de Vías Biológicas
Gene Ontology

Ontología:

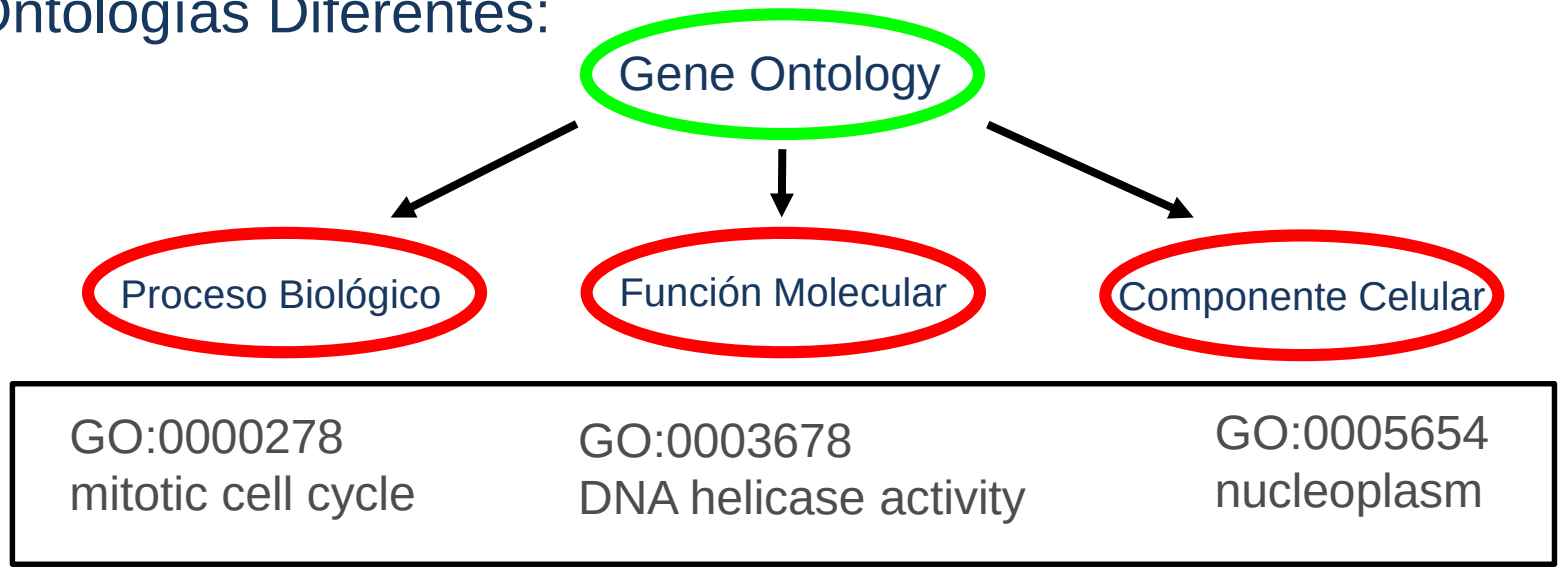
- Colección de vocabularios que describen la biología de un producto génico en algún organismo.

- Sumamente útil para interpretar la importancia biológica de los resultados de datos ómicos.

- GO se organiza en 3 conjuntos ontologías diferentes en un estructura tipo árbol.

Análisis de Vías Biológicas
Gene Ontology

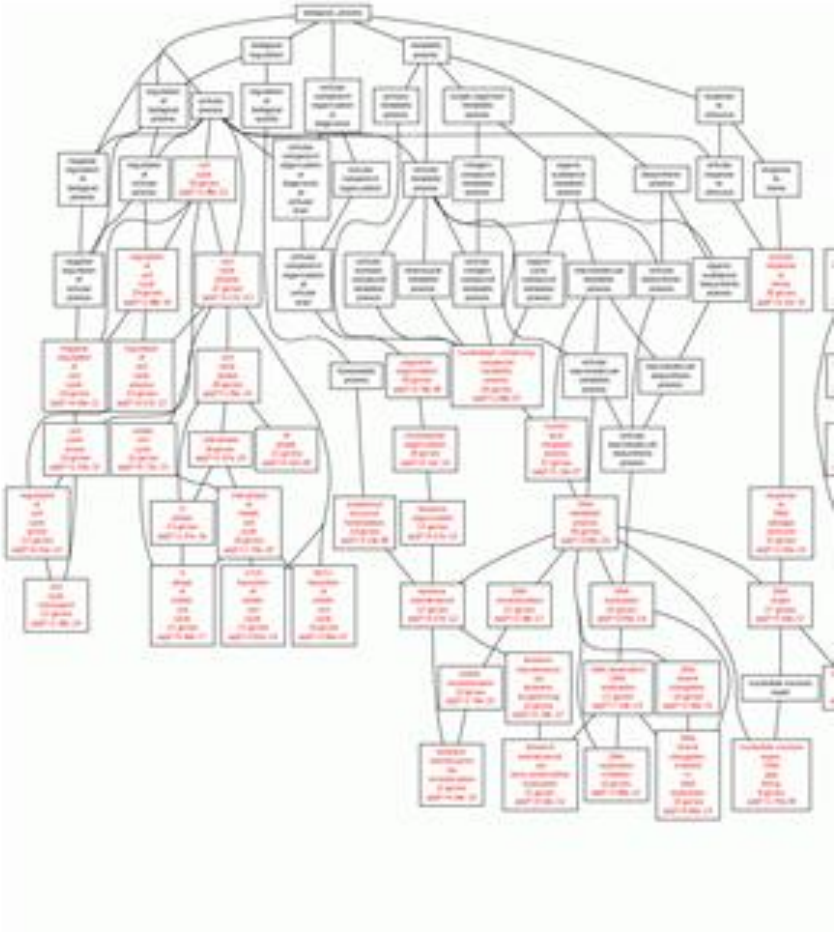
Tres Ontologías Diferentes:

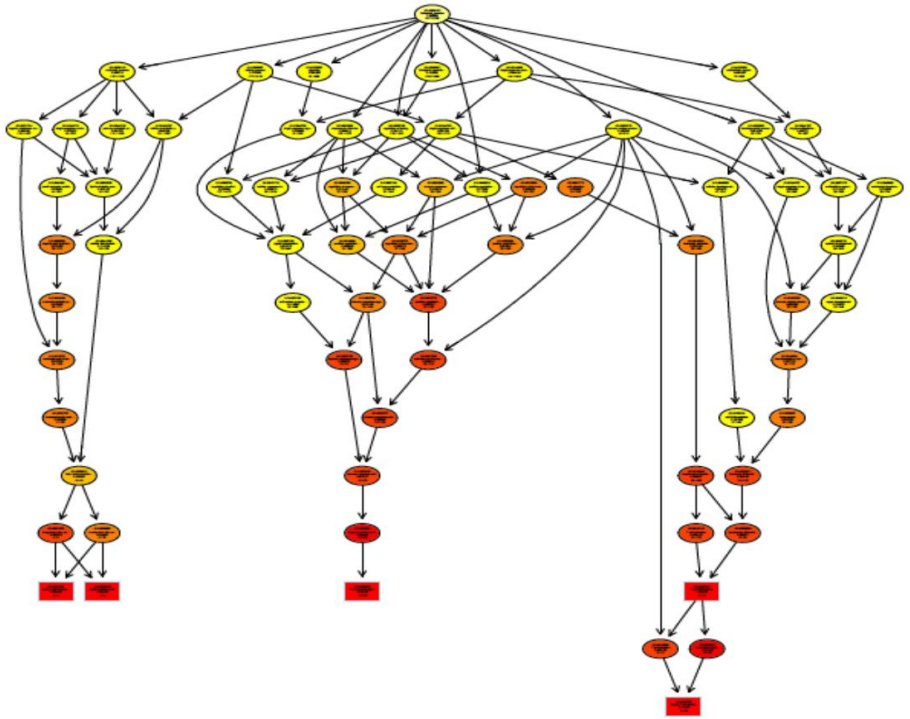


- Red con nodos parentales/hijos organizada mediante un grafo acíclico dirigido.
- A medida que bajamos en el árbol, más detallada es la descripción.
- Un gen puede estar en una o varias ontologías.
- Si está en una nodo hijo, también está anotado en el nodo parental.

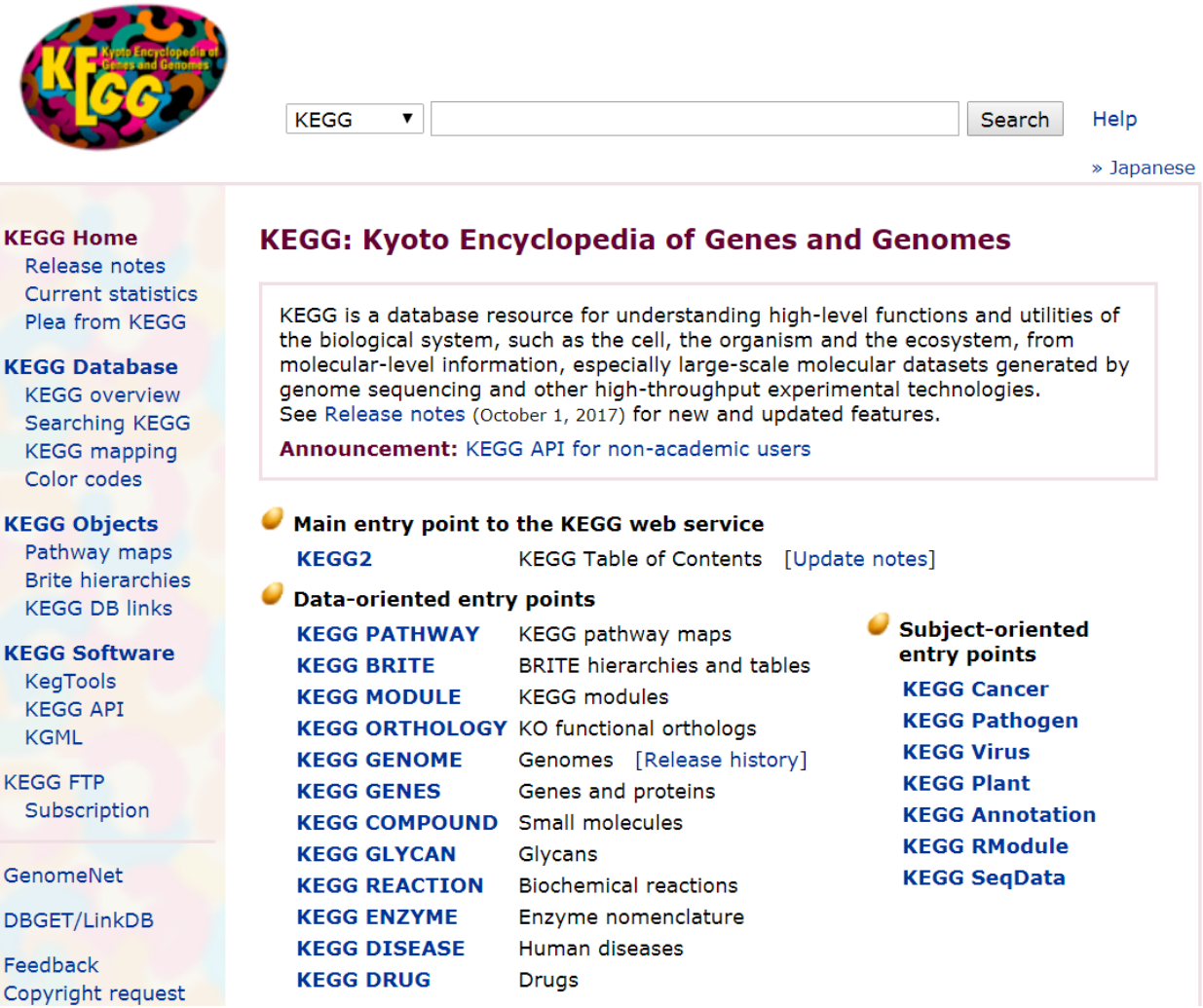
Análisis de Vías Biológicas
Gene Ontology

- Visualizaciones de enriquecimientos vía árboles





Análisis de Vías Biológicas
2. Bases de datos: KEGG



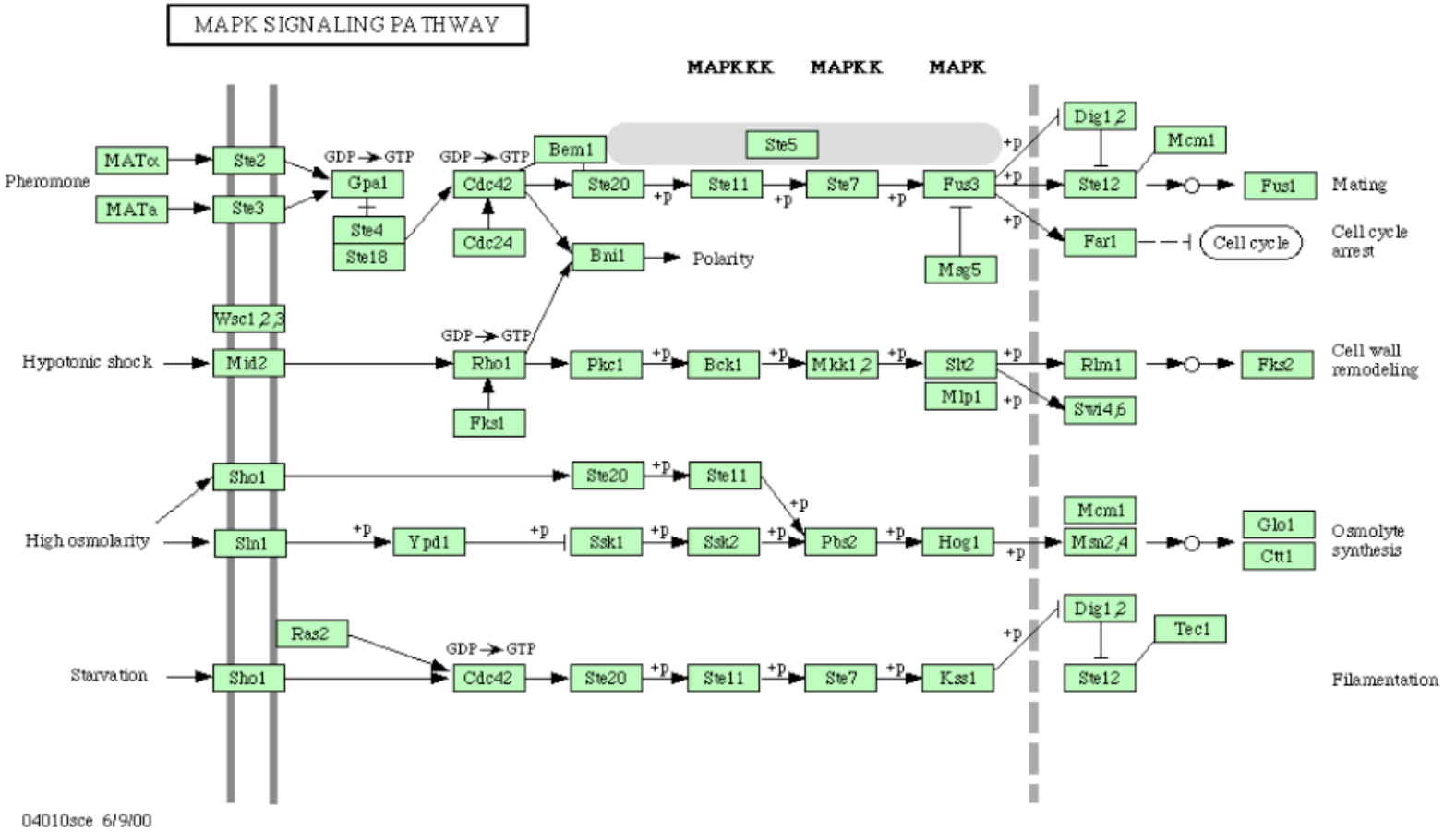
- Mapas de vías curadas manualmente representando nuestro conocimiento de la interacción molecular y redes de reacciones.

- Metabolismo.
- Procesamiento de Información genética.
- Procesamiento Información ambiental.
- Procesos Celulares

http://www.kegg.jp/

Análisis de Vías Biológicas
2. Bases de datos: KEGG

- Visualizaciones de vías KEGG



Análisis de Vías Biológicas
2. Bases de datos: String (práctico)

Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

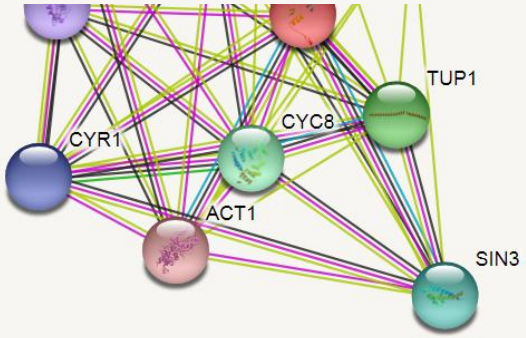
gene co-occurrence

Others

textmining

co-expression

protein homology

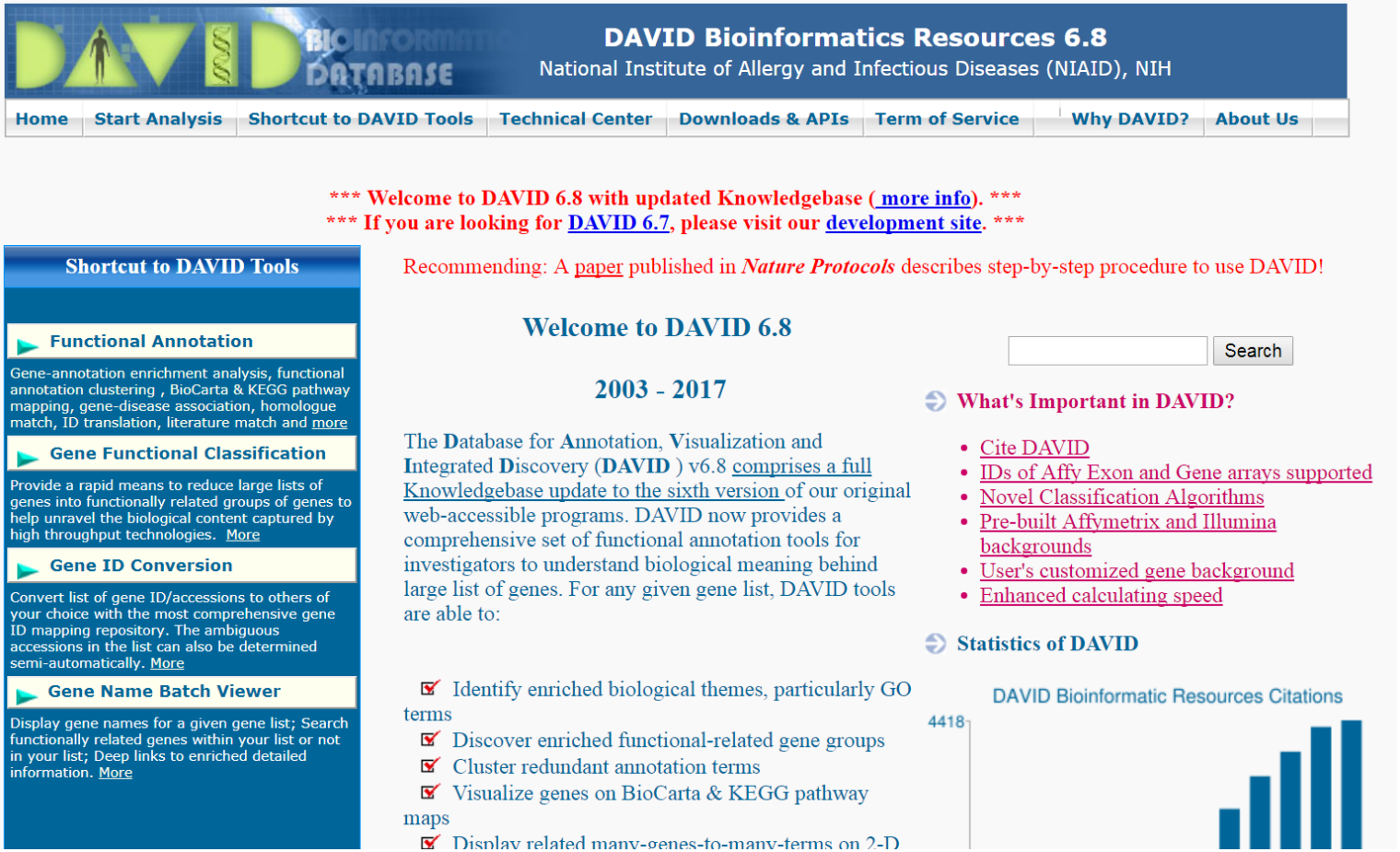


Viewers | Legend | Settings | Analysis | Exports | Clusters | More | Less

http://string-db.org/

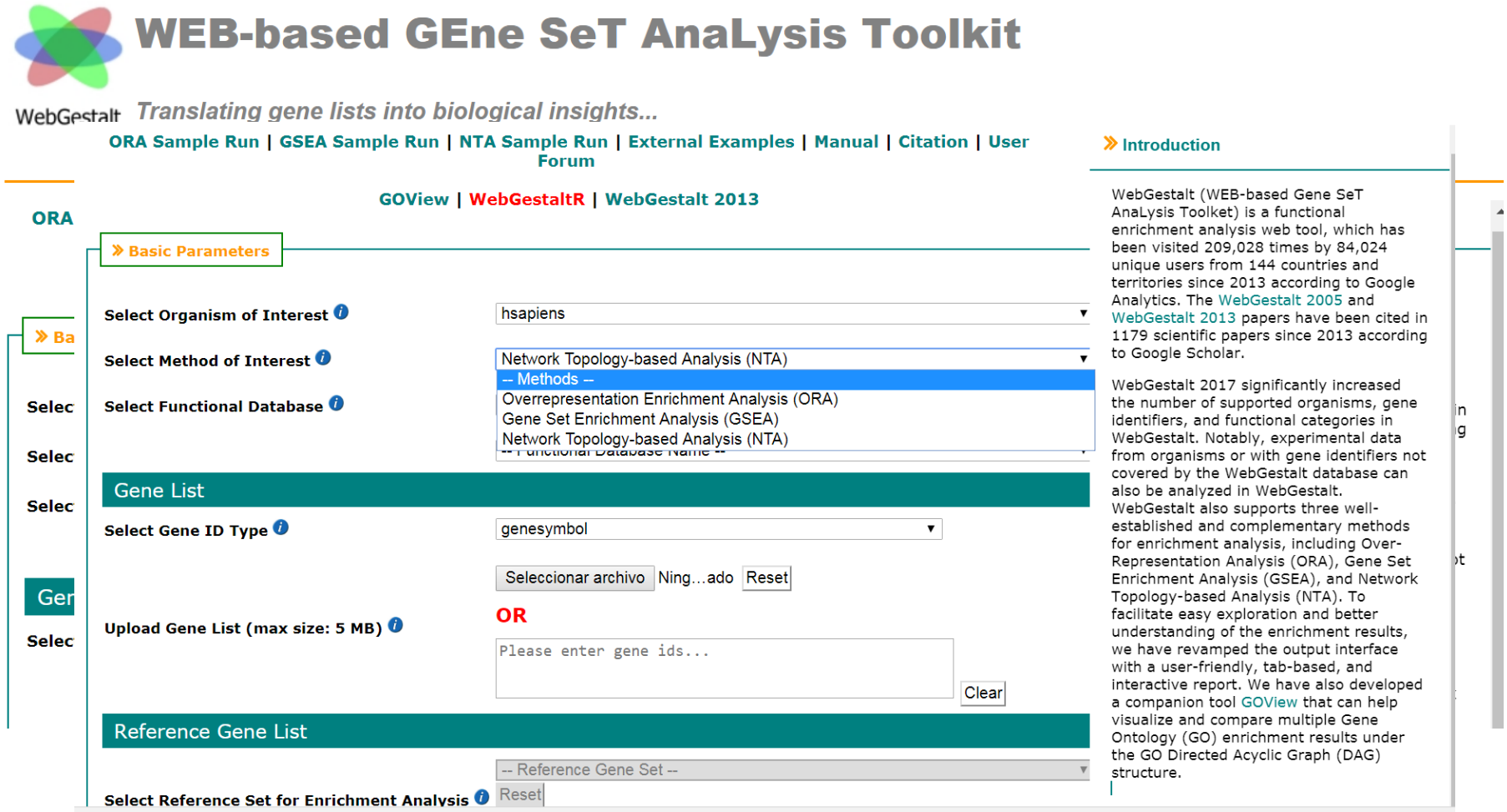
Análisis de Vías Biológicas
2. Bases de datos: DAVID

The **D**atabase for **A**nnotation, **V**isualization and**I**ntegrated **D**iscovery



https://david.ncifcrf.gov/

Análisis de Vías Biológicas
2. Bases de datos: webgestalt



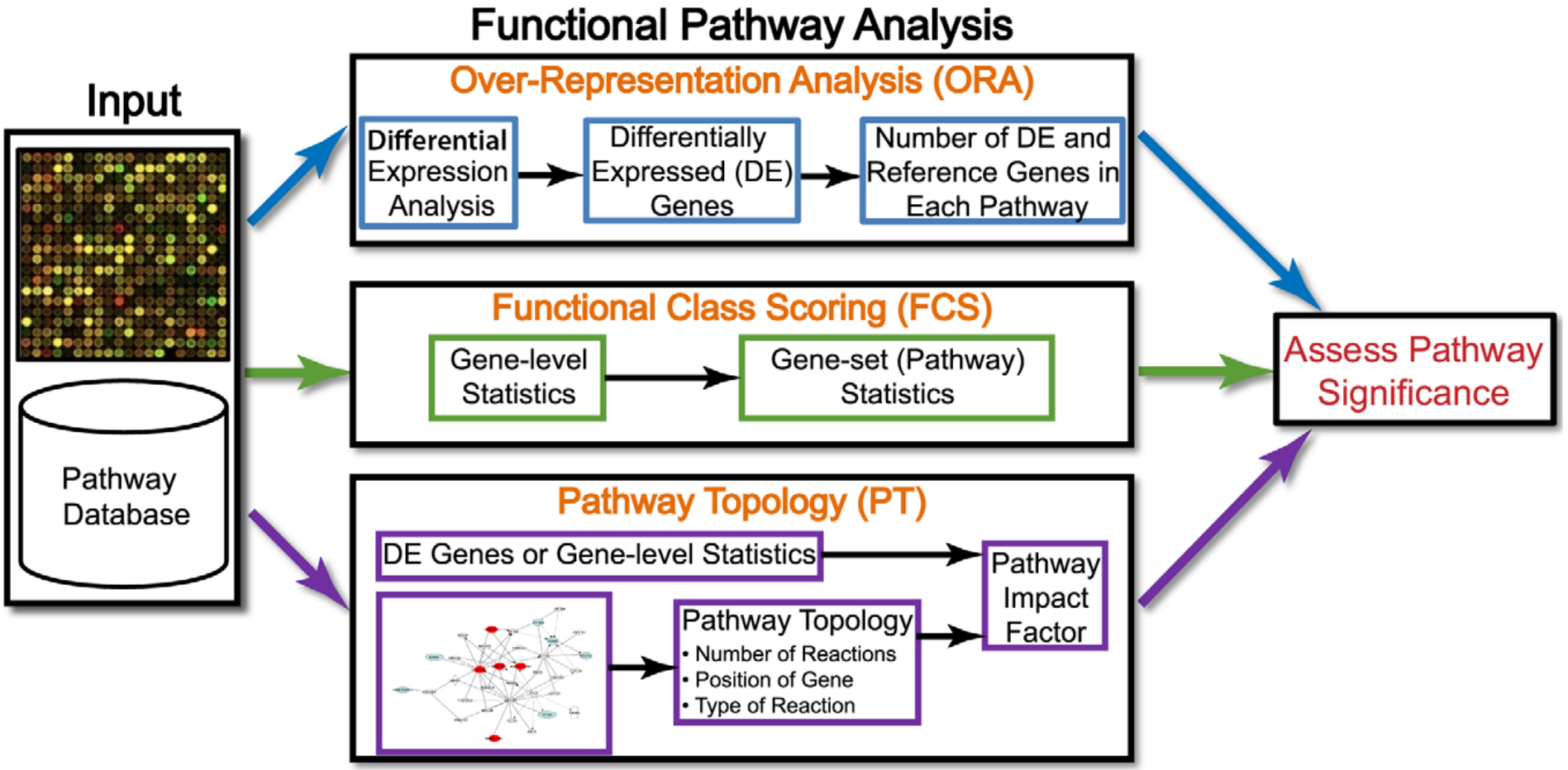
http://www.webgestalt.org/option.php

Análisis de Vías Biológicas
Otras Bases de Datos

- MsigDB
  http://software.broadinstitute.org/gsea/msigdb/
- Reactome
  http://reactome.org/
- Panther
  http://reactome.org/
- Biocyc
  http://biocyc.org/

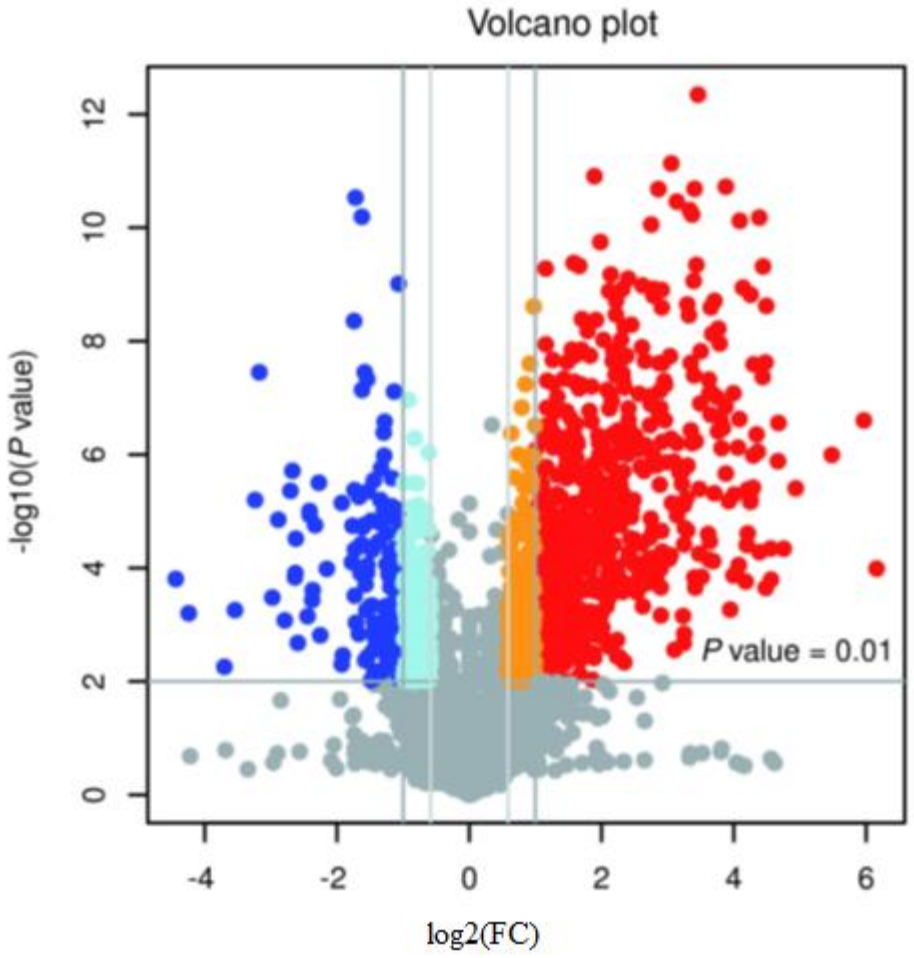
Análisis de Vías Biológicas

3. Métodos estadísticos para el análisis de vías biológicas.



Khatry et al. (2012) PloS ONE 8(2):e1002375

¿Cuántos genes deberíamos analizar?



- La matriz de datos completa.
- Sólo los N genes que más cambiaron.
- Los genes que pasaron una prueba estadística.
- Un conjunto predefinido de genes de interés.

Análisis de Vías Biológicas

3. ORA: Over Representation Analysis.

Este enfoque de análisis considera las vías como listados de genes, por lo tanto los genes por si mismos son los actores principales.

1. Supongamos que nos interesan los DEGs
2. Buscar vías enriquecidas

	IN	OUT	
DEG	a	b	n_1
UNIV	c	d	n_0
	m_1	m_0	N

En R:

- Prueba exacta de Fisher: fisher.test(matrix(c(a,b,c,d),2,2))
- P hipergeometrico:1-phyper(a-1,a+c,b+d,a+b) = 1-phyper(a-1,m_1,m_0,n_1)
- Chi cuadrado: chisq.test(matrix(c(a,b,c,d),2,2))
- Ease Score (DAVID): fisher.test(matrix(c(a-1,b,c,d),2,2))

Análisis de Vías Biológicas

3. Over Representation Analysis (ORA): Universo importa.

Elección del universo como fondo del estadígrafo:

Opciones:

- Usar todas las sondas del microarray.
- Usar todos los genes del organismo.
- Usar todos los genes anotados en la colección de genes a ser testeados.

Recordar:

- Si no se puede medir, no puede ser representado diferencialmente
- Si no puede ser anotado, no puede estar enriquecido en una anotación.

El uso de ORA en GO plantea algunos problemas debido a la estructura jerárquica inherente en las anotaciones.

¿Son los GO terms independientes entre sí?
¿Cómo construir los conjuntos de genes?
¿Debería ser el árbol cortado a cierto nivel?
¿Son los niveles consistentes con el respeto a las raíces y las hojas?

ORA es el modo más simple e intuitivo de analizar vías, pero:

- Usa los IDs de los genes, sin importar su expression.
- Típicamente usa solo DEGs
- Asume que cada gen es independiente del resto.
- Asume que las vías son independientes.

- Modelo usando distribución hipergeométrica.

- Calcula el valor p para determinar si los términos anotados es un listado de genes son mayores a los esperados por azar.

- Ajusta valor p por comparaciones múltiples.

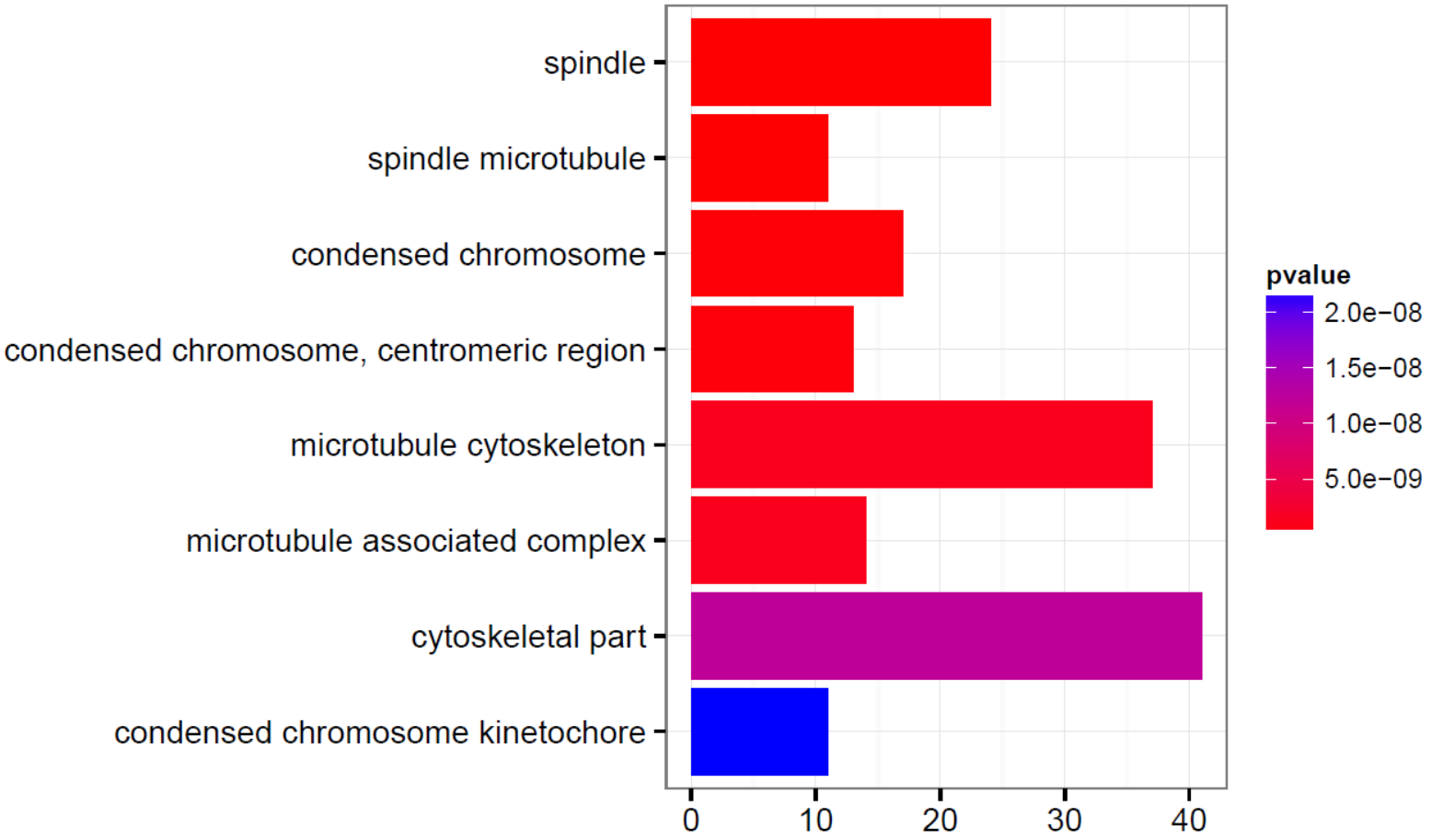
$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Análisis de Vías Biológicas
3. ORA en R: Clusterprofiler (práctico)

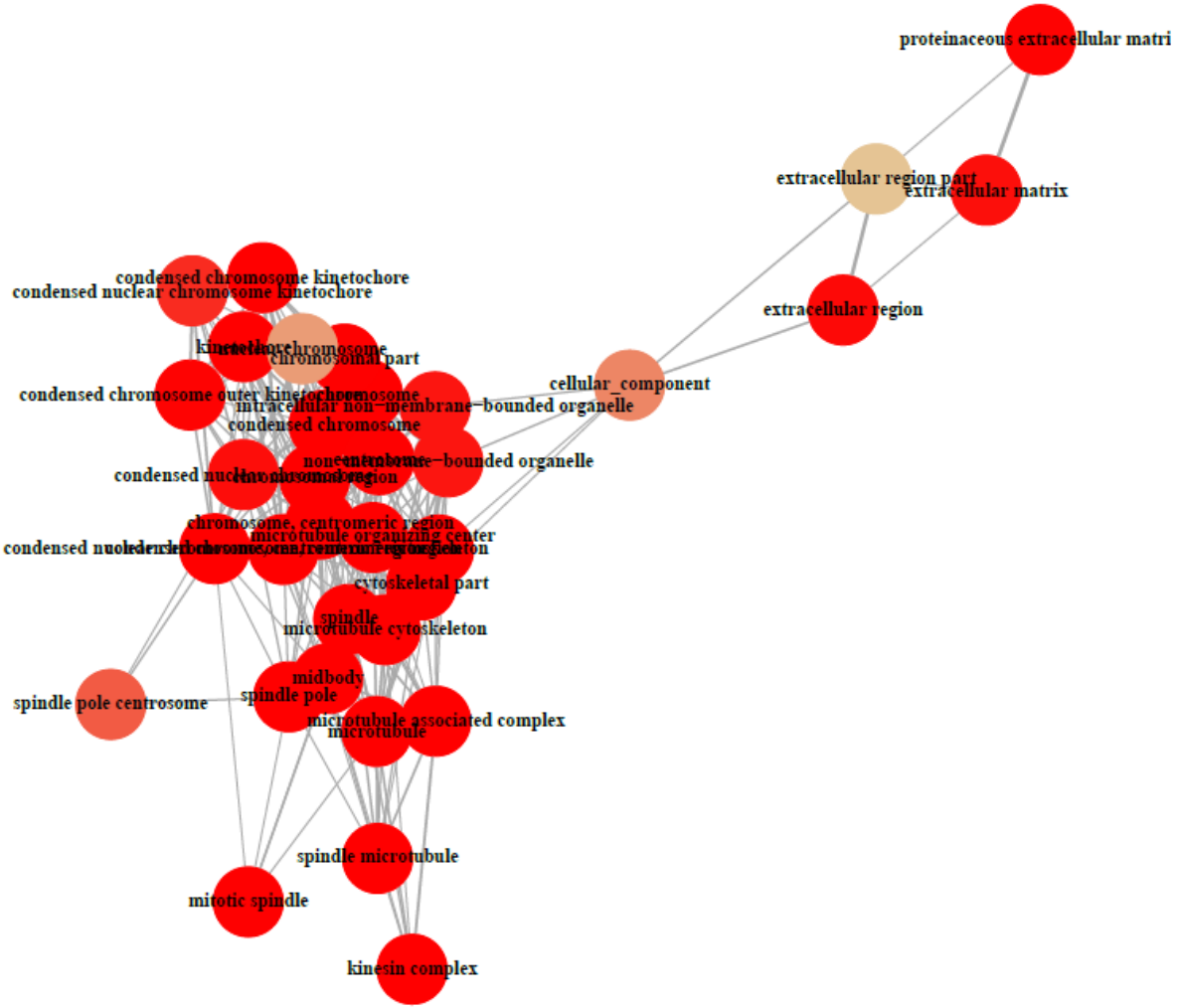
```
ego <- enrichGO(gene          = gene,
                universe      = names(geneList),
                organism      = "human",
                ont           = "CC",
                pAdjustMethod = "BH",
                pvalueCutoff  = 0.01,
                qvalueCutoff  = 0.05,
                readable      = TRUE)
```

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0019953	sexual reproduction	66/410	604/20103	1,77E-15	3,49E-12	3,49E-12	100504195/1	66
GO:0007283	spermatogenesis	52/410	367/20103	1,07E-14	8,06E-12	8,06E-12	100504195/1	52
GO:0048232	male gamete generation	52/410	368/20103	1,23E-15	8,06E-12	8,06E-12	100504195/1	52
GO:0044703	multi-organism reproductive process	66/410	701/20103	1,02E-11	5,02E-09	5,02E-09	100504195/1	66
GO:0007276	gamete generation	53/410	475/20103	3,74E-10	1,47E-07	1,47E-07	100504195/1	53
GO:0000003	reproduction	66/410	770/20103	1,99E-09	6,54E-07	6,54E-07	100504195/1	66

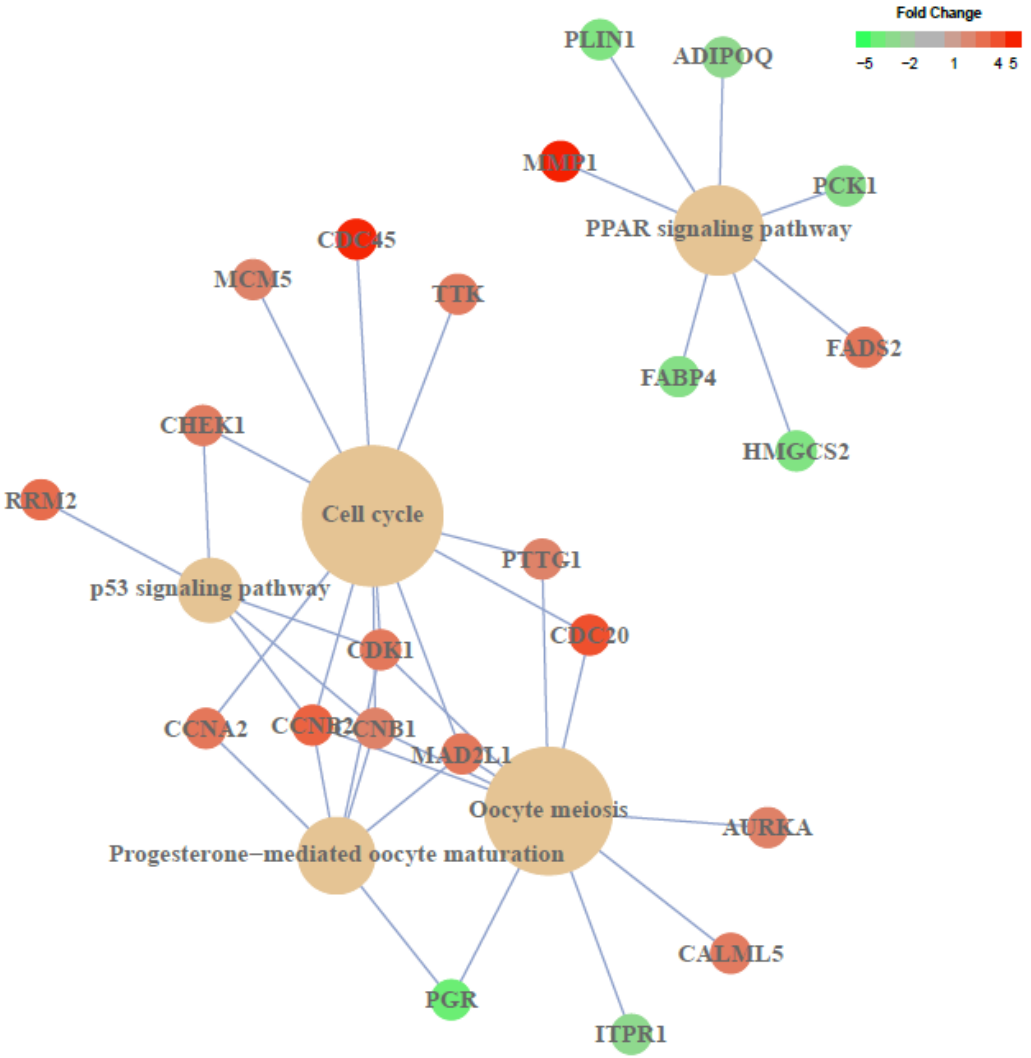
Análisis de Vías Biológicas
3. ORA en R: Clusterprofiler (práctico)



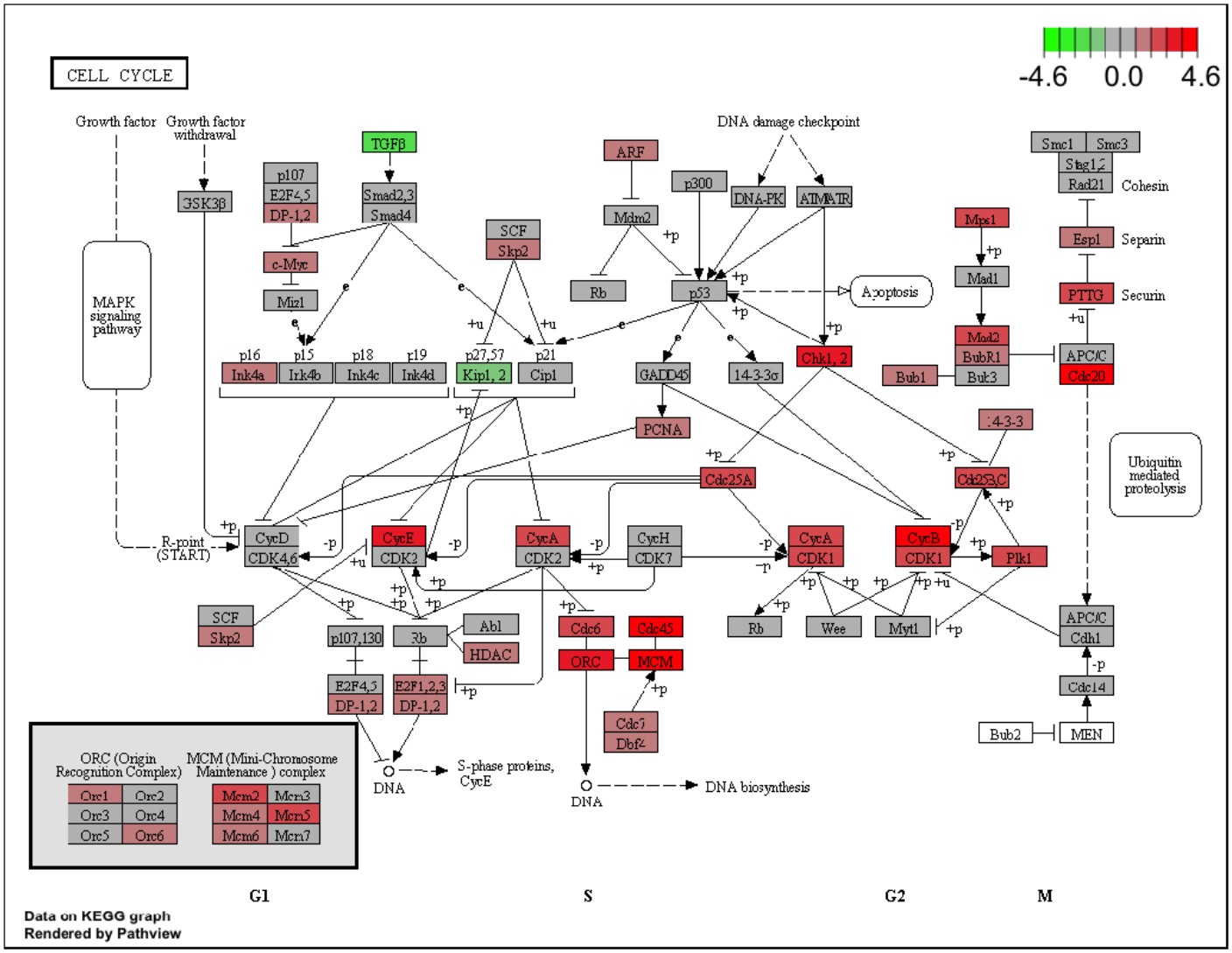
Análisis de Vías Biológicas
3. ORA en R: Clusterprofiler (práctico)



Análisis de Vías Biológicas
3. ORA en R: Clusterprofiler (práctico)



Análisis de Vías Biológicas
3. ORA en R: Clusterprofiler (práctico)



Análisis de Vías Biológicas
3. FCS: Functional Class Scoring

Aunque grandes cambios en genes individuales puede tener efectos significativos en las vías, pequeños cambios pueden coordinar cambios en los conjuntos de genes relacionados funcionalmente, teniendo así cambios significativos también.

Pasos principales:
1. Calcular individualmente en los genes un estadígrafo.

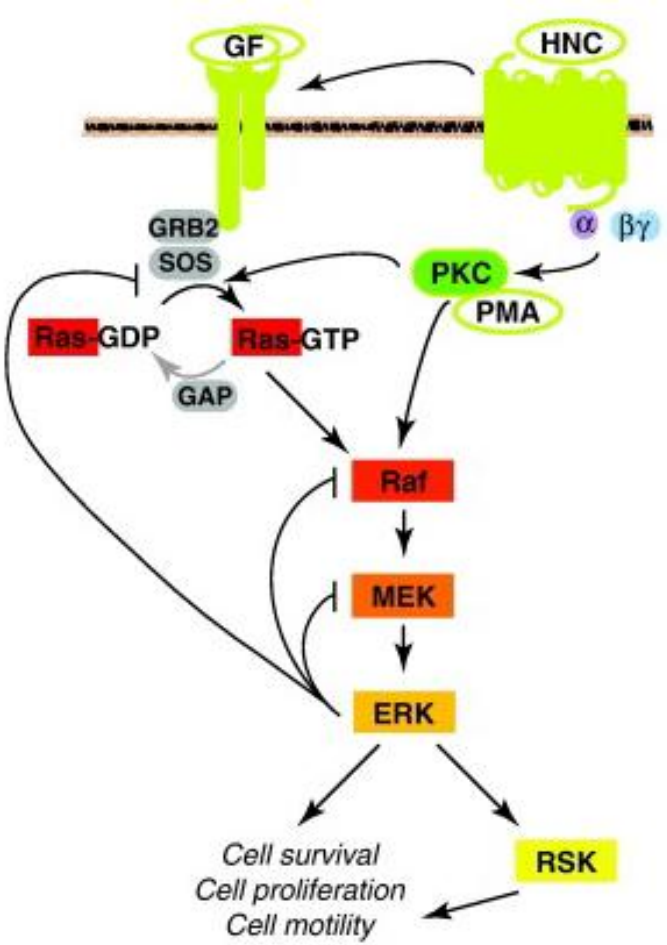
2. Los estadisgrafos de todos los genes en un set de genes son agregados en un solo estadísgrafo a nivel de vía.
(usando métodos uni o multivariados).

3. Llevar a cabo una nueva prueba de significancia a nivel de vía.

Análisis de Vías Biológicas
3. PT: Pathway Topology analysis

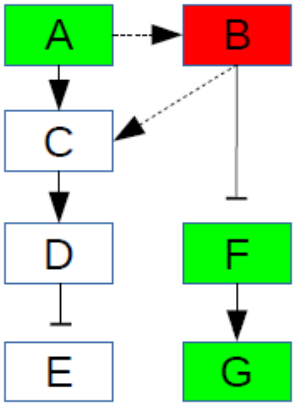
Método con los mismos pasos que FCS, pero toma en cuenta la topología de la vía.

Alternativas de paquetes en R:
SPIA,
CePa ORA
CePa GSA
Pathnet.



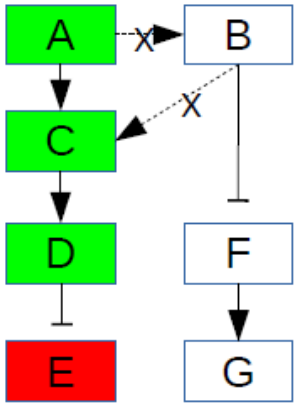
Análisis de Vías Biológicas
3. PT: Pathway Topology analysis

Método con los mismos pasos que FCS, pero toma en cuenta la topología de la vía.



(En ambos casos, 1 gen up y 3 down)
Genetic buffering entre A, B y C.
A controla a B y C, efecto indirecto sobre E y G
B controla a C y F, efecto indirecto sobre E y G

Si A es subexpresado, B aumenta para que CDE no sea alterado, mientras que FG disminuyen.

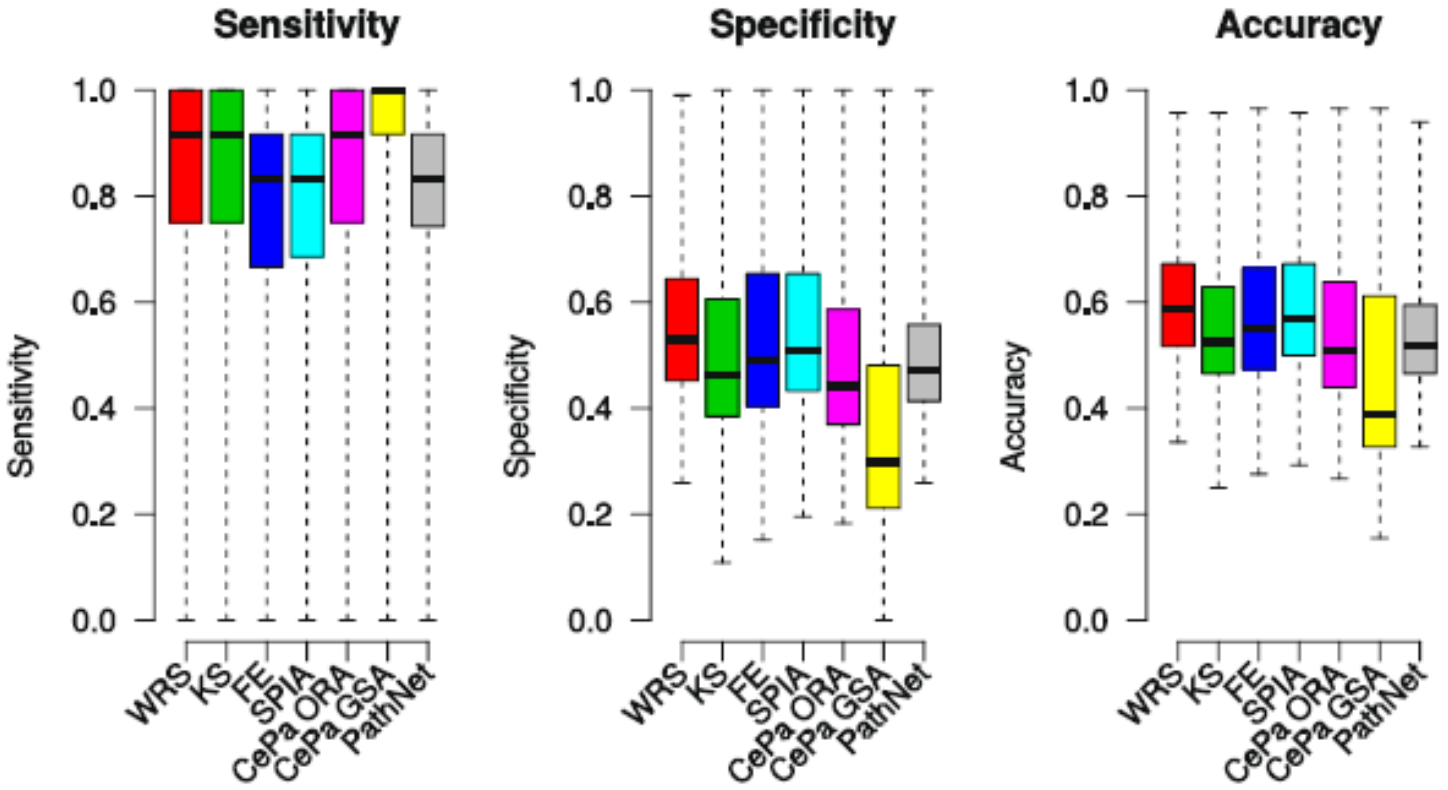


Sin *genetic buffering*
A solo controla a C
B solo controla a F

Si A es subexpresado, B sigue igual, y el efecto total en CDE y FG es diferente.

Análisis de Vías Biológicas
3. PT: Pathway Topology analysis

Bayerlova et al. BMC Bioinnformatics (2015)16:334



A pesar que los métodos basados en topología introducen más información, el aumento de su complejidad no refleja un aumento en el resultado biológico.

Análisis de Vías Biológicas

Break!