



UNIVERSIDAD  
DE CHILE

# 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

1. Vías Biológicas.
2. Bases de datos.
3. Métodos estadísticos para el análisis de vías biológicas.
4. Práctico en R: Clusterprofiler, String.DB



UNIVERSIDAD  
DE CHILE

# 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:

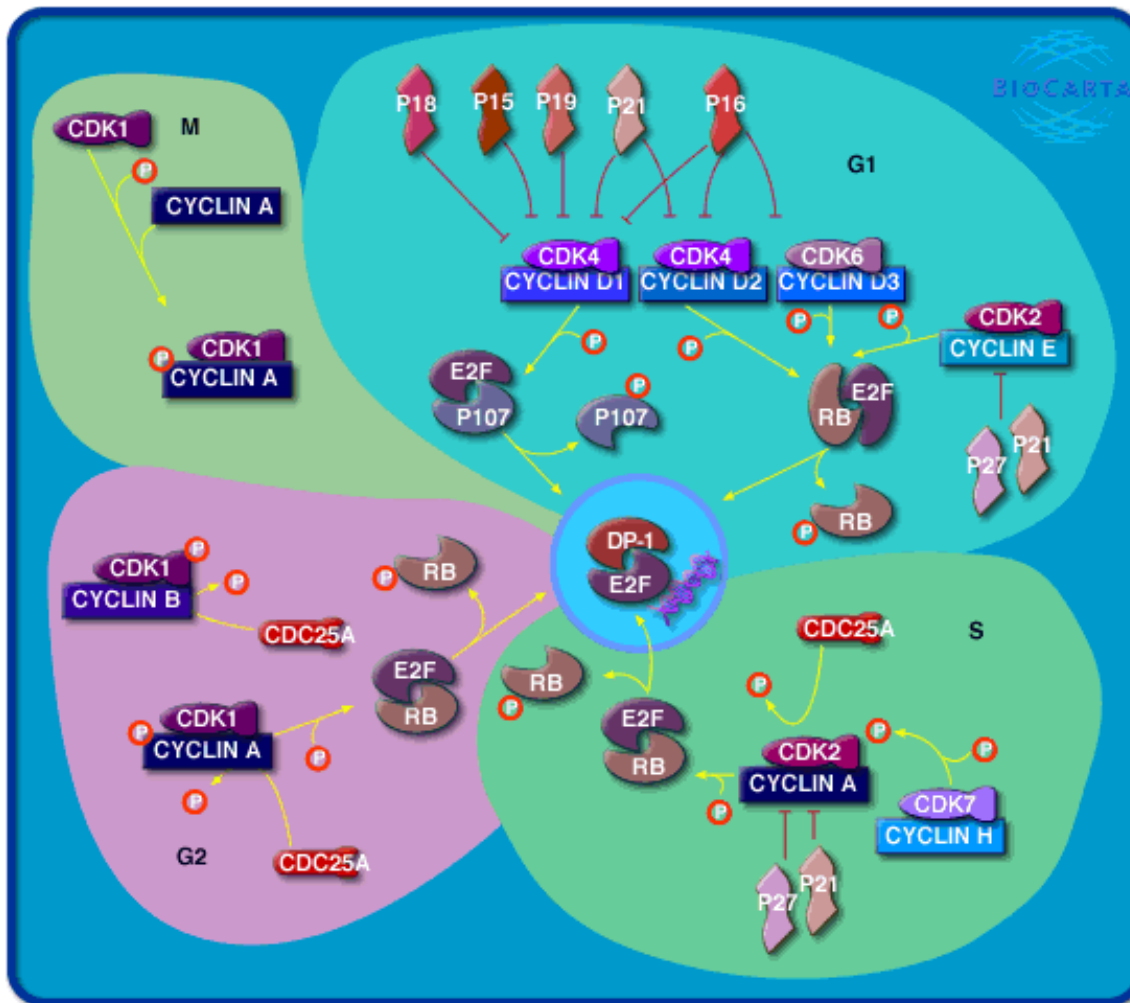
- Metabolismo.
- Regulación de la expresión génica.
- Transducción de señales.





# Análisis de Vías Biológicas

## Vías Biológicas: Transducción de señales



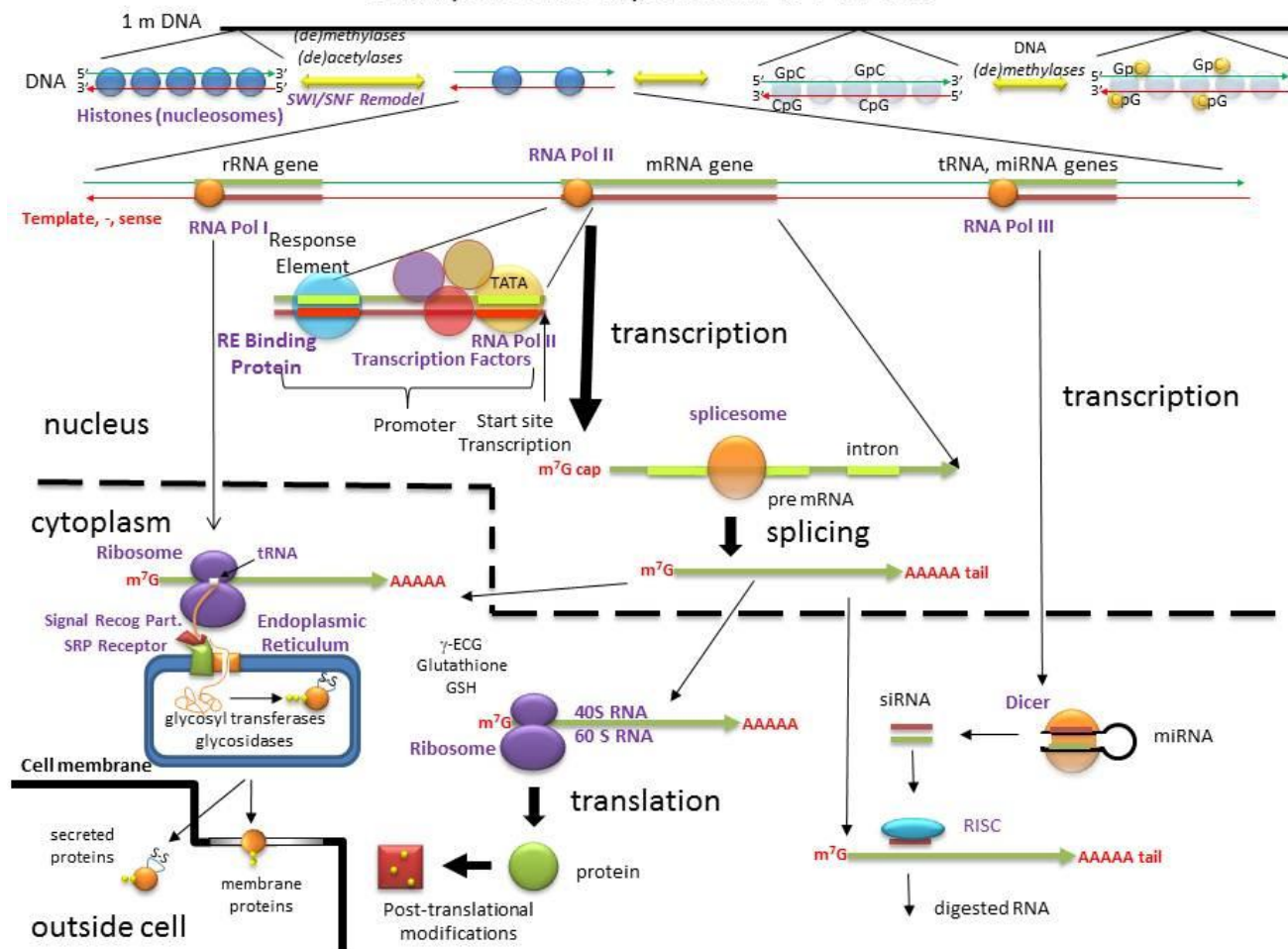


# Análisis de Vías Biológicas

## Vías Biológicas:



### Eukaryotic Gene Expression: An Overview





UNIVERSIDAD  
DE CHILE

# 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)



Fly  
EMBL-E  
GR  
Ast

### Enrichment analysis

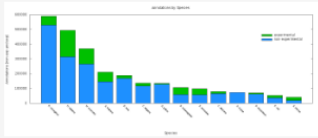
Your gene IDs here...

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/>



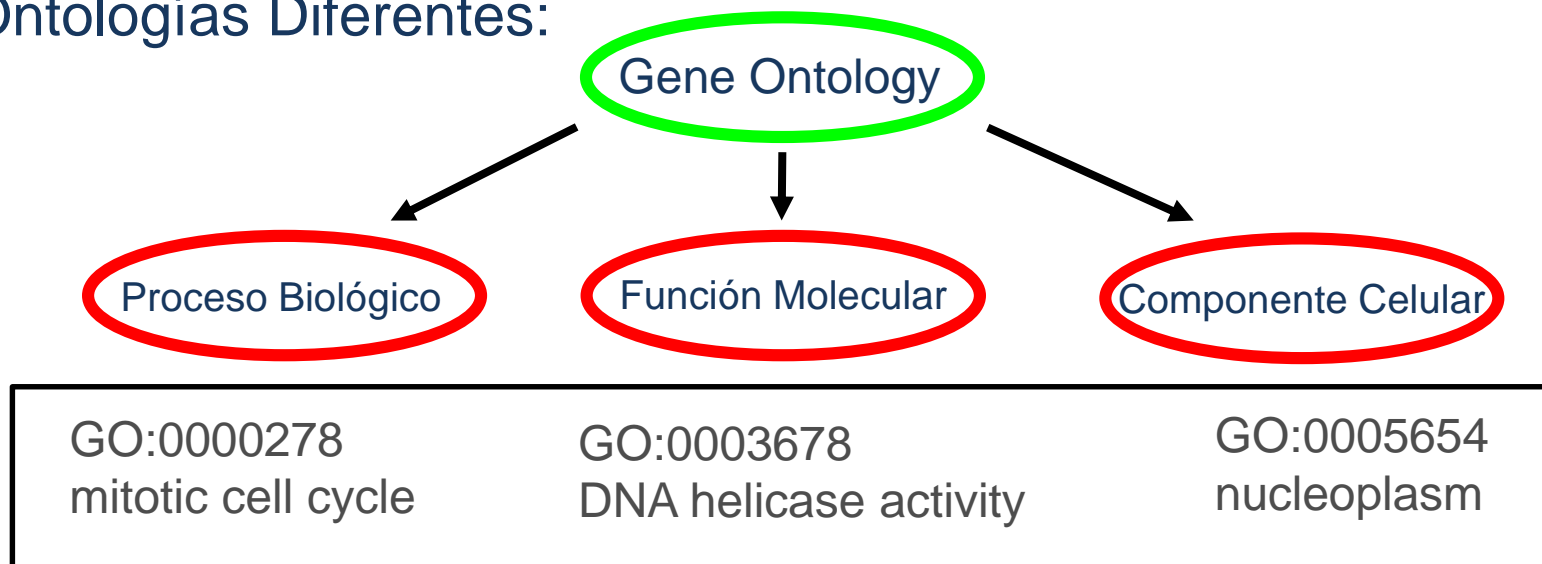


### Ontología:

- 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.



Tres Ontologías Diferentes:



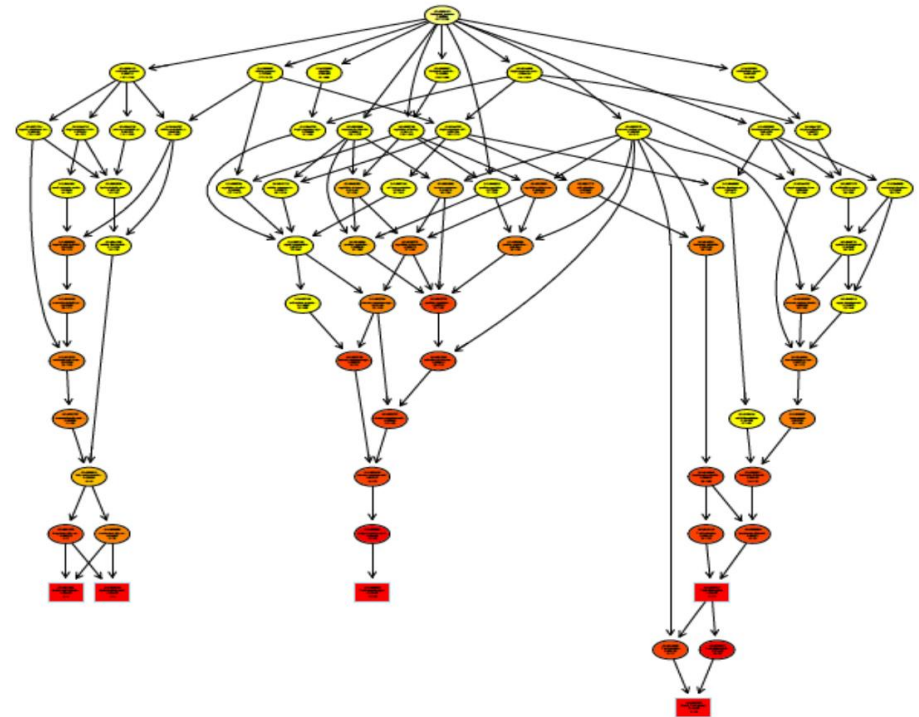
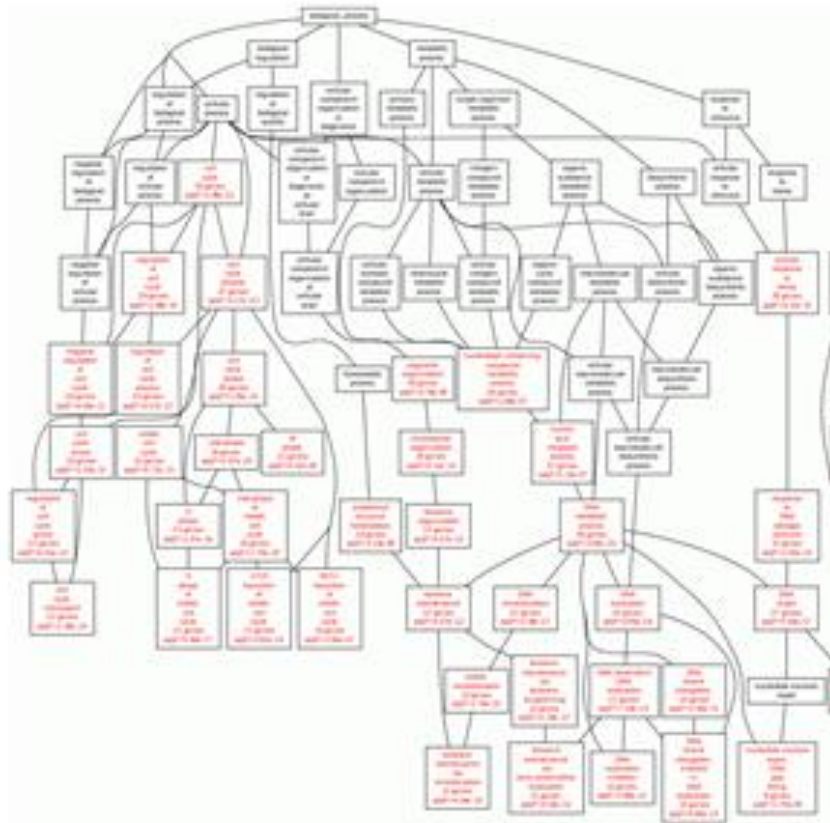
- Red con nodos parentales/hijos organizada mediante un grafo acíclico dirigido.
- A medida que bajamos en el árbol, más detallada es la descripción.
- Un gen puede estar en una o varias ontologías.
- Si está en una nodo hijo, también está anotado en el nodo parental.



# Análisis de Vías Biológicas Gene Ontology



- Visualizaciones de enriquecimientos vía árboles





[Help](#)
  
[» Japanese](#)

### KEGG Home

[Release notes](#)  
[Current statistics](#)  
[Plea from KEGG](#)

### KEGG Database

[KEGG overview](#)  
[Searching KEGG](#)  
[KEGG mapping](#)  
[Color codes](#)

### KEGG Objects

[Pathway maps](#)  
[Brite hierarchies](#)  
[KEGG DB links](#)

### KEGG Software

[KegTools](#)  
[KEGG API](#)  
[KGML](#)

### KEGG FTP

[Subscription](#)

[GenomeNet](#)

[DBGET/LinkDB](#)

[Feedback](#)  
[Copyright request](#)

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (October 1, 2017) for new and updated features.

**Announcement:** [KEGG API for non-academic users](#)

### ● Main entry point to the KEGG web service

**KEGG2** [KEGG Table of Contents](#) [[Update notes](#)]

### ● Data-oriented entry points

**KEGG PATHWAY** [KEGG pathway maps](#)  
**KEGG BRITE** [BRITE hierarchies and tables](#)  
**KEGG MODULE** [KEGG modules](#)  
**KEGG ORTHOLOGY** [KO functional orthologs](#)  
**KEGG GENOME** [Genomes](#) [[Release history](#)]  
**KEGG GENES** [Genes and proteins](#)  
**KEGG COMPOUND** [Small molecules](#)  
**KEGG GLYCAN** [Glycans](#)  
**KEGG REACTION** [Biochemical reactions](#)  
**KEGG ENZYME** [Enzyme nomenclature](#)  
**KEGG DISEASE** [Human diseases](#)  
**KEGG DRUG** [Drugs](#)

### ● Subject-oriented entry points

**KEGG Cancer**  
**KEGG Pathogen**  
**KEGG Virus**  
**KEGG Plant**  
**KEGG Annotation**  
**KEGG RModule**  
**KEGG SeqData**

- 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/>

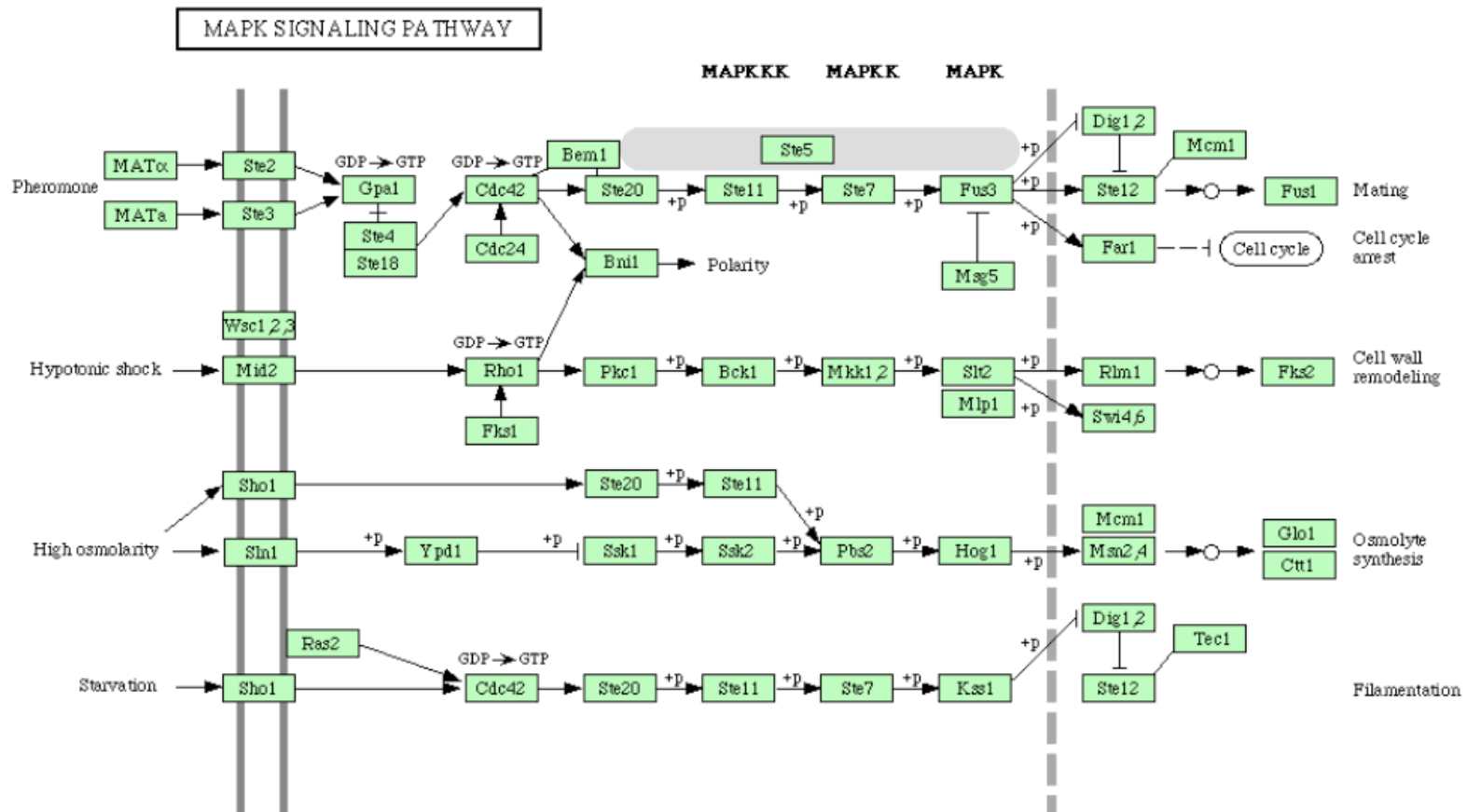


# Análisis de Vías Biológicas

## 2. Bases de datos: KEGG



### ● Visualizaciones de vías KEGG



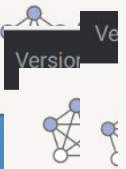
04010sce 6/9/00



UNIVERSIDAD DE CHILE

# 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



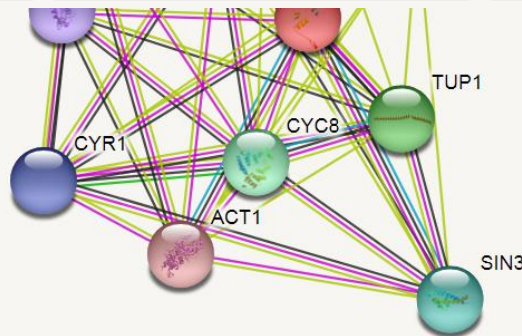
*textmining*



*co-expression*



*protein homology*



Viewers >

Legend v

Settings >

Analysis >

Exports >

Clusters >

+ More

- Less

<http://string-db.org/>



# The Database for Annotation, Visualization and Integrated Discovery



**DAVID Bioinformatics Resources 6.8**  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

\*\*\* Welcome to DAVID 6.8 with updated Knowledgebase ([more info](#)). \*\*\*  
\*\*\* If you are looking for [DAVID 6.7](#), please visit our [development site](#). \*\*\*

### Shortcut to DAVID Tools

- Functional Annotation**  
Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)
- Gene Functional Classification**  
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)
- Gene ID Conversion**  
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)
- Gene Name Batch Viewer**  
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. [More](#)

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

## Welcome to DAVID 6.8

2003 - 2017

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 [comprises a full Knowledgebase update to the sixth version](#) of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D

### What's Important in DAVID?

- [Cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

### Statistics of DAVID

DAVID Bioinformatic Resources Citations



<https://david.ncifcrf.gov/>



## WEB-based GENE SeT AnaLysis Toolkit

WebGestalt *Translating gene lists into biological insights...*

[ORA Sample Run](#) | [GSEA Sample Run](#) | [NTA Sample Run](#) | [External Examples](#) | [Manual](#) | [Citation](#) | [User Forum](#)

» Introduction

[GOView](#) | [WebGestaltR](#) | [WebGestalt 2013](#)

ORA

» Basic Parameters

» Ba

Sele

Sele

Sele

Ger

Sele

Select Organism of Interest <sup>i</sup>

hsapiens

Select Method of Interest <sup>i</sup>

Network Topology-based Analysis (NTA)

-- Methods --

Overrepresentation Enrichment Analysis (ORA)

Gene Set Enrichment Analysis (GSEA)

Network Topology-based Analysis (NTA)

-- Functional Database Name --

Select Functional Database <sup>i</sup>

Gene List

Select Gene ID Type <sup>i</sup>

genesymbol

Seleccionar archivo Ning...ado Reset

Upload Gene List (max size: 5 MB) <sup>i</sup>

Please enter gene ids...

Clear

Reference Gene List

-- Reference Gene Set --

Select Reference Set for Enrichment Analysis <sup>i</sup> Reset

WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) is a functional enrichment analysis web tool, which has been visited 209,028 times by 84,024 unique users from 144 countries and territories since 2013 according to Google Analytics. The [WebGestalt 2005](#) and [WebGestalt 2013](#) papers have been cited in 1179 scientific papers since 2013 according to Google Scholar.

WebGestalt 2017 significantly increased the number of supported organisms, gene identifiers, and functional categories in WebGestalt. Notably, experimental data from organisms or with gene identifiers not covered by the WebGestalt database can also be analyzed in WebGestalt. WebGestalt also supports three well-established and complementary methods for enrichment analysis, including Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA). To facilitate easy exploration and better understanding of the enrichment results, we have revamped the output interface with a user-friendly, tab-based, and interactive report. We have also developed a companion tool [GOView](#) that can help visualize and compare multiple Gene Ontology (GO) enrichment results under the GO Directed Acyclic Graph (DAG) structure.

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



- 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.

### Functional Pathway Analysis

#### Over-Representation Analysis (ORA)

Differential  
Expression  
Analysis

Differentially  
Expressed (DE)  
Genes

Number of DE and  
Reference Genes in  
Each Pathway

#### Functional Class Scoring (FCS)

Gene-level  
Statistics

Gene-set (Pathway)  
Statistics

#### Pathway Topology (PT)

DE Genes or Gene-level Statistics



Pathway Topology  
• Number of Reactions  
• Position of Gene  
• Type of Reaction

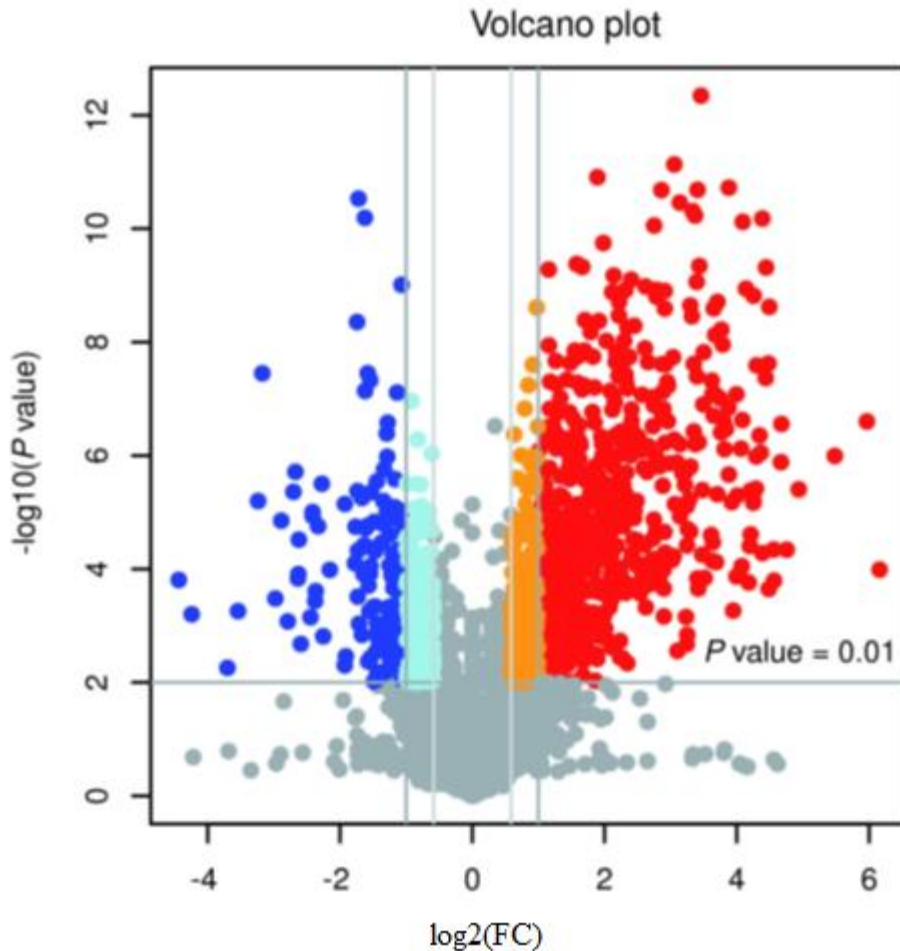
Pathway  
Impact  
Factor

Assess Pathway  
Significance

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.



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,m1,m0,n1)`
- 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))`



UNIVERSIDAD  
DE CHILE

# 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.



UNIVERSIDAD  
DE CHILE

# Análisis de Vías Biológicas

## 3. ORA: Temas pendientes

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?



UNIVERSIDAD  
DE CHILE

# Análisis de Vías Biológicas

## 3. ORA: Temas pendientes



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.



# Análisis de Vías Biológicas

## 3. ORA en R: Clusterprofiler (práctico)



- 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)
```

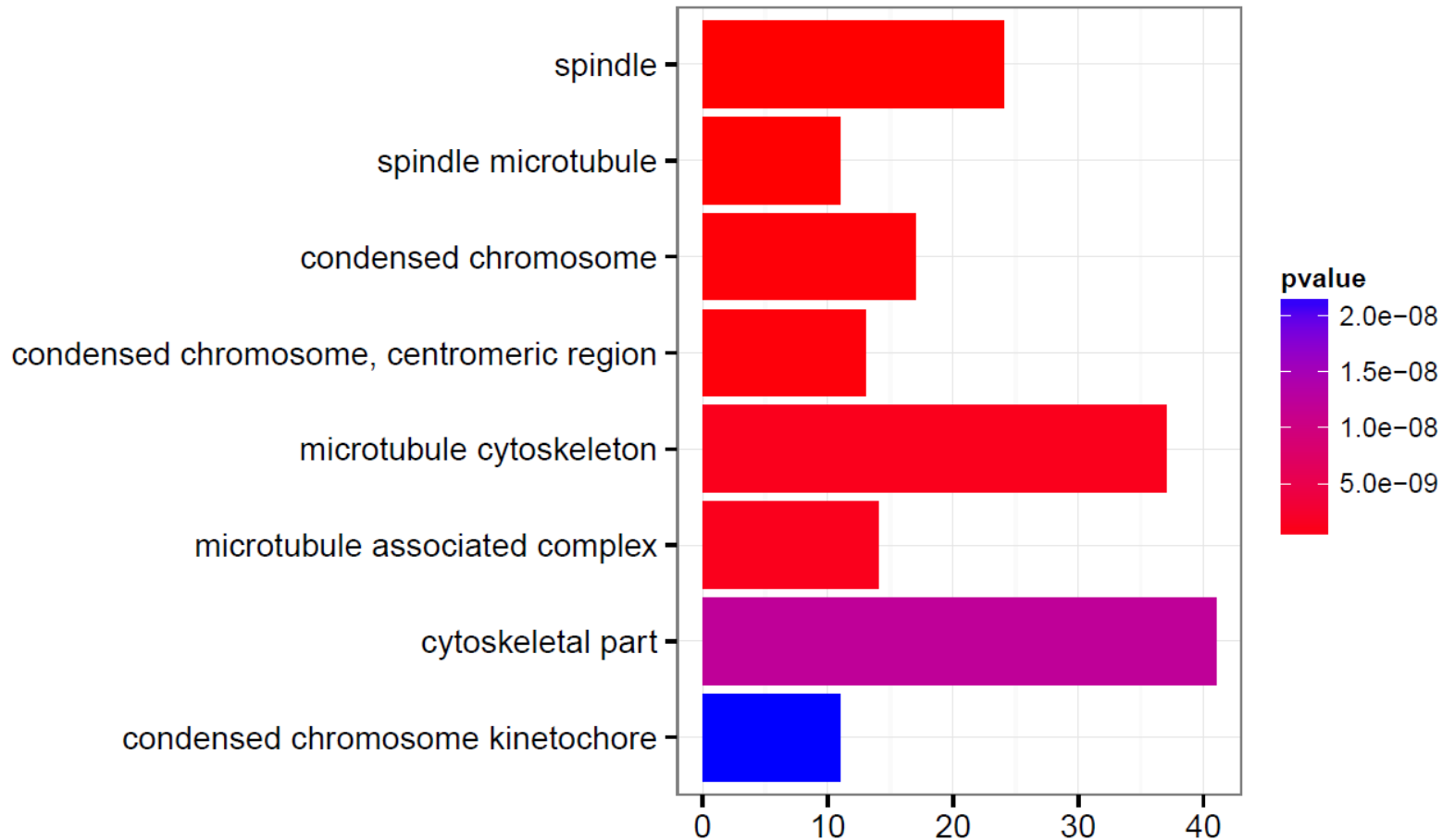
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0019953	sexual reproduction	66/410	604/20103	1,77E-15	3,49E-12	3,49E-12	100504195/1	66
GO:0007283	spermatogenesis	52/410	367/20103	1,07E-14	8,06E-12	8,06E-12	100504195/1	52
GO:0048232	male gamete generation	52/410	368/20103	1,23E-15	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



UNIVERSIDAD  
DE CHILE

# Análisis de Vías Biológicas

## 3. ORA en R: Clusterprofiler (práctico)

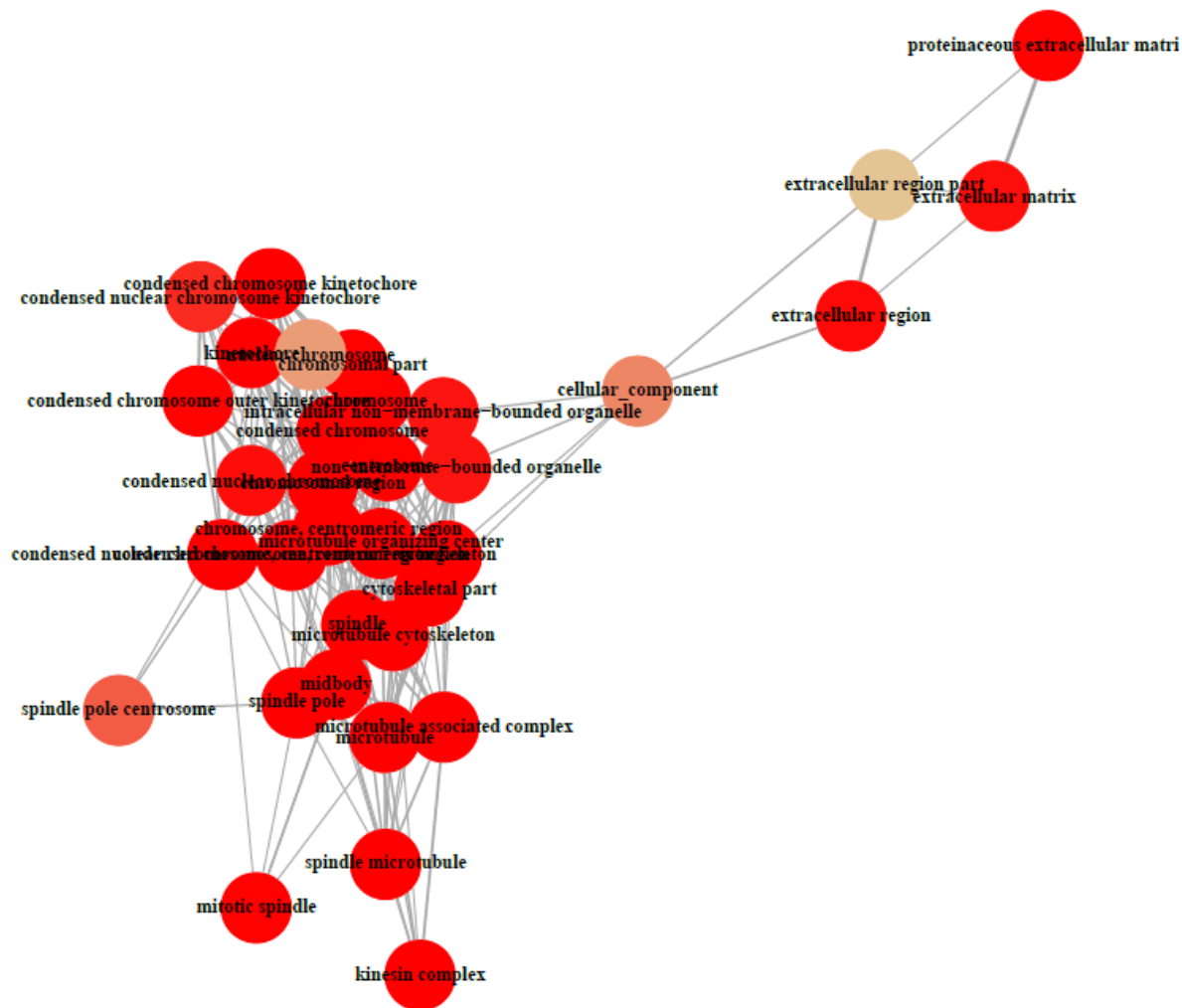




UNIVERSIDAD DE CHILE

# Análisis de Vías Biológicas

## 3. ORA en R: Clusterprofiler (práctico)

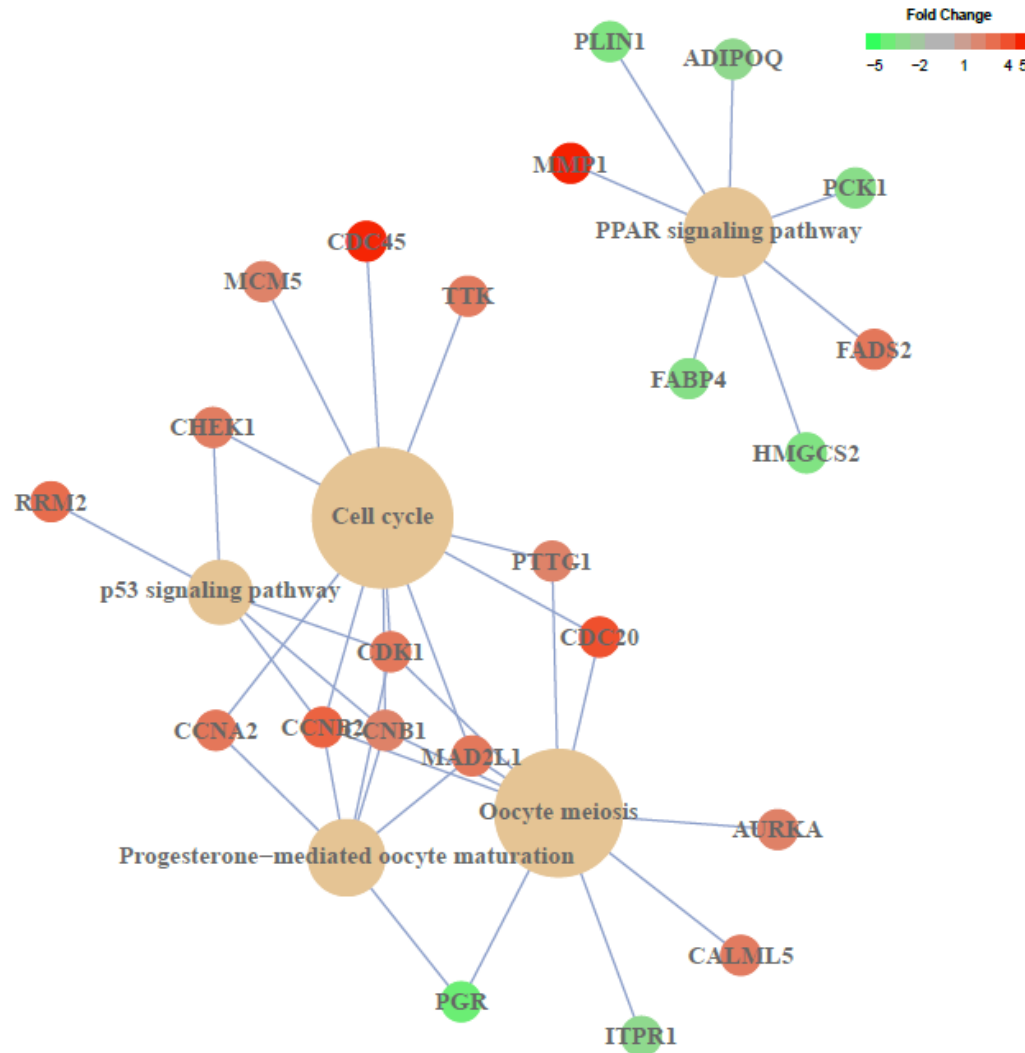




UNIVERSIDAD  
DE CHILE

# Análisis de Vías Biológicas

## 3. ORA en R: Clusterprofiler (práctico)





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.



UNIVERSIDAD  
DE CHILE

# 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.

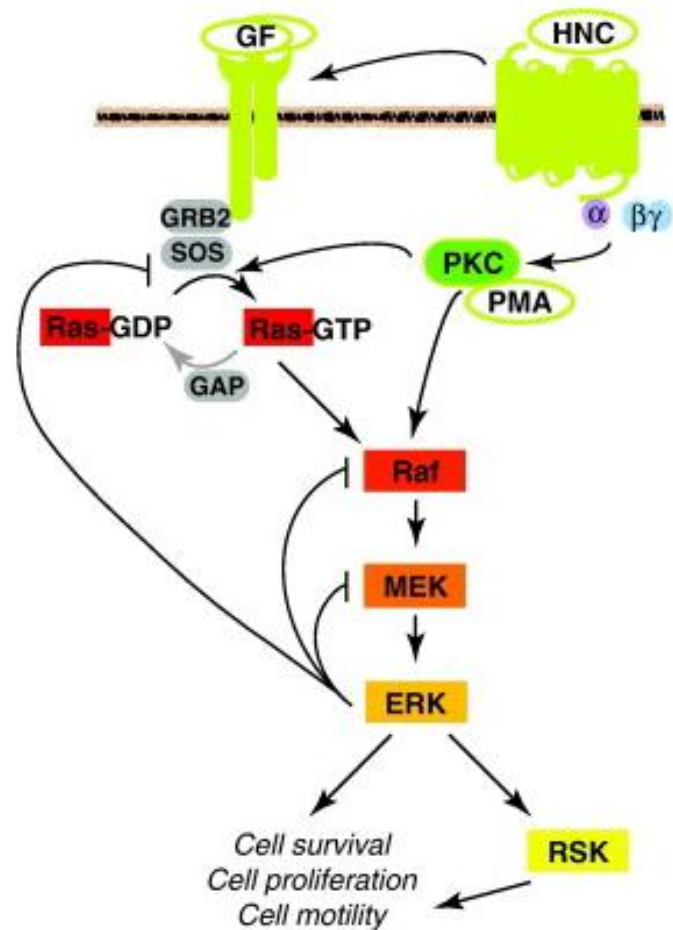
Alternativas de paquetes en R:

SPIA,

CePa ORA

CePa GSA

Pathnet.

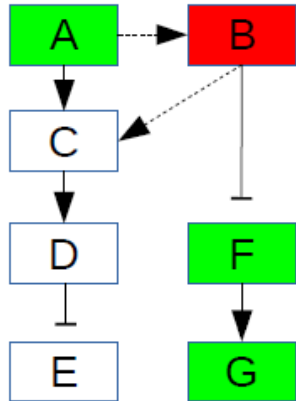




# Análisis de Vías Biológicas

## 3. PT: Pathway Topology analysis

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



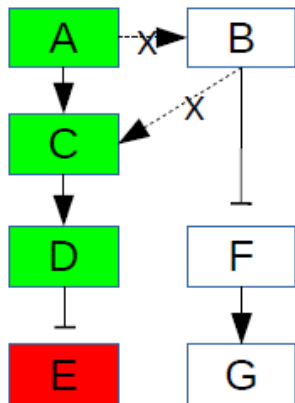
(En ambos casos, 1 gen up y 3 down)

*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.



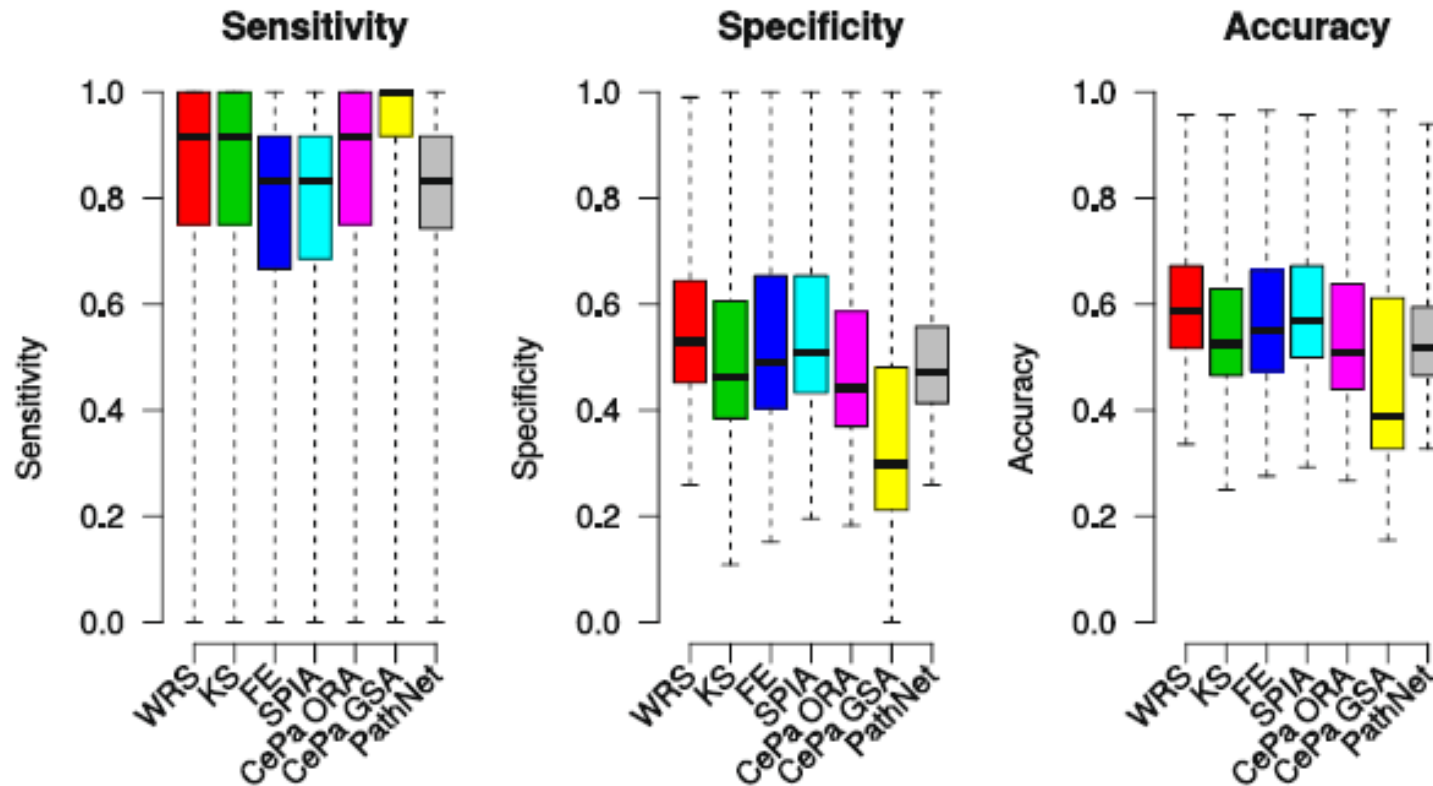


# Análisis de Vías Biológicas

## 3. PT: Pathway Topology analysis



Bayerlova et al. BMC Bioinformatics (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.



UNIVERSIDAD  
DE CHILE

# Análisis de Vías Biológicas



Break!